

Language-Informed Flow Matching for Trend-Guided Structure-Based 3D Molecular Generation

Tianyu Gao^{1,*} Zhikai Su^{1,*} Jiashu Li¹ Wenjun Gao¹
Zichuan Ying² Zhe Zhao³ Fei Zhang^{1,†} Ye Wei^{1,†}

¹City University of Hong Kong ²The University of Hong Kong ³Stanford University

*Equal contribution. †Corresponding authors.

feizhang010518@gmail.com ye.wei@cityu.edu.hk

Abstract

Structure-based drug design (SBDD) requires ligands that satisfy both 3D target affinity and 1D chemical validity. Existing controllable generation methods often rely on task-specific fine-tuning or externally imposed sampling-time guidance, adding cost and potentially conflicting with evolving 3D geometric constraints. We propose LiFT, a language-informed cross-modal framework built on Flow Matching for trend-guided 3D molecular generation across both de novo design and scaffold hopping. LiFT uses a “Sense-Evolve-Assemble” agent to generate target-aware SMILES as intermediate chemical conditions, from which a pre-trained chemical foundation model extracts continuous semantic priors. These priors are integrated into geometric generation through a lightweight semantic projector with zero-initialized adaptive normalization for stable cross-modal conditioning. We further introduce a Self-Conditioned Decoupled Router (SCDR), which modulates the velocity field according to intermediate structural states during ODE integration. Experiments on Cross-Docked2020 show that LiFT achieves competitive distribution matching while improving medicinal chemistry metrics and maintaining competitive structural validity under task-steering settings without additional generator fine-tuning. Our results suggest that language-derived chemical priors provide effective trend-level guidance for 3D molecular generation. Code and released artifacts are available at <https://github.com/kasurl/LiFT>.

1 Introduction

Structure-based drug design (SBDD) accelerates therapeutic discovery by designing small-molecule ligands that selectively bind disease-associated protein targets (Anderson, 2003; Zhang et al., 2024b). Computationally, this objective requires simultane-

ously satisfying two constraints: *3D physicochemical complementarity* within the binding pocket and *1D chemical validity* (Buttenschoen et al., 2024; Lin et al., 2025). Over-optimizing geometric affinity at the expense of chemical viability, or vice versa, often produces non-synthesizable or inactive molecules (Jiang et al., 2025). Consequently, a central challenge in SBDD is balancing 3D spatial interactions with valid chemical structures across the vast molecular space (Walters and Murcko, 2020).

Purely 3D SBDD models capture target-pocket geometry, but discrete chemical constraints such as valency, ring topology, and medicinal-chemistry preferences are often only implicit. Conversely, LLM-based molecular generators can exploit chemical knowledge from SMILES and natural-language descriptions, but standalone 1D proposals cannot ensure target-specific 3D compatibility. Recent 1D–3D molecular generators such as Liu et al. (2025) and Chen et al. (2025) demonstrate the promise of integrating chemical priors with geometric generation, yet they mainly transfer molecular-string representations into diffusion-style generators rather than addressing how human task preferences, pocket context, and property trade-offs can be translated into semantic conditions for pocket-aware 3D generation. They are also largely pocket-free, and their cross-modal conditioning is typically fixed rather than adapted to the evolving 3D state. In SBDD, such semantic–geometric mismatch can induce steric conflicts and reduce structural validity, motivating state-aware semantic guidance during generation (Kritikos et al., 2026; Singh, 2026; Lang et al., 2026).

To address these limitations, we propose LiFT, a language-informed cross-modal framework built on flow matching for trend-guided 3D target-aware molecular generation. LiFT asks whether molecular strings generated from human-readable de-

sign preferences can serve as reusable semantic priors for downstream geometric generation, rather than merely acting as final molecule proposals. A "Sense-Evolve-Assemble" LLM agent with Pocket-of-Thought reasoning converts pocket information, task preferences, and optional references into target-aware SMILES conditions. A frozen chemical foundation encoder then maps these conditions into dense semantic latents, which are integrated into an ODE-governed flow field through zero-initialized lightweight modulation. To coordinate semantic trends with evolving spatial configurations, we further introduce the Self-Conditioned Decoupled Router (SCDR), which adapts semantic influence according to invariant summaries of intermediate structural states while preserving the $SO(3)$ -equivariant geometry of the pocket-centered backbone (Satorras et al., 2021). This design enables property-aware steering across *de novo* design and reference-guided scaffold hopping without additional generator fine-tuning.

Extensive experiments on CrossDocked2020 show that LiFT provides effective trend-level guidance while maintaining competitive distribution matching and structural validity, offering a lightweight approach to language-informed SBDD generation. Our main contributions are summarized as follows:

- We introduce a language-informed 1D–3D conditioning pathway for target-aware SBDD, translating human-readable design preferences and pocket context into reusable SMILES-derived semantic priors. These priors are aligned with downstream geometric generation through state-aware cross-modal modulation, enabling property-aware trend guidance without treating LLM outputs as final molecules.
- We introduce the Self-Conditioned Decoupled Router (SCDR), coupled with zero-initialized lightweight modulation, to adapt semantic influence according to intermediate spatial states and coordinate language-derived chemical priors with geometric constraints along the integration trajectory.
- We design an evaluation protocol that jointly measures distribution matching, property-oriented generation, and industrial chemical filter compliance. Under this protocol, LiFT shows competitive distribution matching and

effective trend steering across both *de novo* design and scaffold hopping without additional generator fine-tuning.

2 Related Work

2.1 Generative Models in Structure-Based Drug Design

Structure-based drug design (SBDD) aims to generate ligands that are compatible with 3D protein pockets in both geometry and chemistry. Recent 3D generative models directly operate on atomic coordinates, including spatial autoregressive models (Luo et al., 2021; Peng et al., 2022), diffusion-based methods (Guan et al., 2023a; Huang et al., 2024; Guan et al., 2023b; Schneuing et al., 2024), and continuous-time Flow Matching frameworks (Zhang et al., 2024a; Dunn and Koes, 2026; Zhou et al., 2025; Schneuing et al., 2025). These approaches are well suited for modeling pocket-conditioned spatial configurations, but discrete chemical constraints such as valency, ring patterns, and medicinal-chemistry preferences are not always explicit in continuous coordinate generation.

A complementary line of work treats molecules as 1D sequences and applies language-modeling architectures to SMILES or related representations (Bagal et al., 2022). Such models naturally capture chemical syntax and topological regularities, but their connection to target-specific 3D pocket geometry is less direct. Target-aware sequence models have been explored for drug design (Wu et al., 2024), while recent methods such as ELILLM further investigate the use of language-model representations for structure-based molecular design (Hu et al., 2026). Our work follows this broader direction but treats language-derived molecular representations as intermediate semantic conditions for native pocket-conditioned 3D generation, with state-aware modulation along the flow trajectory.

2.2 Cross-Modal Conditioning and State-Aware Modulation

Cross-modal conditioning is commonly used to incorporate external information into generative models. Existing approaches include cross-attention or graph-language interaction modules (Liu et al., 2023), as well as FiLM- or AdaLN-style modulation in diffusion and transformer architectures (Perez et al., 2018; Peebles and Xie, 2023). Recent 1D–3D molecular generation methods further use learned projection modules to connect molecular

representations with 3D generation processes (Liu et al., 2025; Chen et al., 2025). Our work follows this direction, but uses zero-initialized lightweight modulation (Zhu et al., 2026) to integrate SMILES-derived priors into the flow model with limited architectural overhead.

Frozen chemical foundation models provide a useful interface for this conditioning by encoding SMILES into continuous representations that capture chemical syntax and property-related regularities from large molecular corpora (Ross et al., 2022; Soares et al., 2025a,b). We use these representations as soft chemical priors for 3D generation, rather than exact structural targets. Unlike inference-time guidance methods that modify sampling or optimization trajectories with external objectives (Choi et al., 2025; Jin et al., 2026; Guo et al., 2025), our conditioning signal is incorporated into the learned velocity field.

Because a fixed semantic signal may not be equally useful throughout ODE generation, our framework uses state-aware modulation to adapt the influence of SMILES-derived priors according to intermediate ligand states while preserving the geometric backbone of the 3D flow model.

3 Methodology

We propose LiFT, a language-conditioned framework for pocket-conditioned 3D molecular generation (Figure 1). LiFT builds on continuous-time Flow Matching (Schneuing et al., 2025), learning a vector field that transports prior noise toward the empirical ligand distribution conditioned on a 3D protein pocket. It consists of four stages: (i) target-aware SMILES generation with the *Sense-Evolve-Assemble LLM Agent* (§3.1); (ii) semantic latent extraction from generated SMILES (§3.2); (iii) scalar-domain conditioning through a lightweight semantic projector (§3.3); and (iv) state-aware velocity-field modulation with the *Self-Conditioned Decoupled Router* (§3.4). Detailed flow-matching derivations are provided in Appendix G.

3.1 Sense-Evolve-Assemble LLM Agent

To make language conditioning explicit and chemically grounded, we introduce a *Sense-Evolve-Assemble* agent based on a Pocket-of-Thought (PoT) prompting strategy. Instead of directly decoding a molecule from the protein pocket, the agent first constructs a structured pocket profile, then derives task-specific SMILES candidates, and finally

sanitizes the generated sequence before passing it to the 3D generator.

In the **Sense** phase, the agent summarizes the local microenvironment of the pocket $\mathcal{P}$. It defines the generative region with a spatial bounding box $\mathcal{B} \in \mathbb{R}^{3 \times 2}$ and applies Principal Component Analysis (PCA) to pocket coordinates to estimate the major geometric axes. We further restrict the interaction analysis to an 8.0 Å shell around the pocket centroid, within which the agent extracts key physicochemical cues, including hydrophobic residues, acidic and basic residues, and metal coordination centers such as ZN and MG.

During the **Evolve** phase, the agent performs task-specific ligand proposal under either reference-free (*de novo*) or reference-guided (*scaffold hopping*) settings. Guided by PoT, it first produces a structured reasoning path rather than a raw molecule, relating candidate substructures such as rings or heteroatoms to nearby pocket features and assessing whether the proposed molecular topology is compatible with the pocket geometry. The prompt also incorporates medicinal chemistry preferences, allowing different generation modes to emphasize distribution matching, binding-oriented design, drug-likeness, or balanced property profiles.

In the **Assemble** phase, the proposed components are serialized into a raw SMILES sequence S_{raw} . To reduce hallucinated or invalid outputs, we apply deterministic knowledge-guided decoding with cheminformatics checks for syntax errors, unmatched ring closures, and valency violations. The repaired and sanitized sequence S_{smiles} is then used as the 1D semantic condition for the downstream continuous 3D flow matching process.

3.2 Semantic Latent Extraction for Cross-Modal Priming

To connect discrete SMILES conditions with continuous geometric generation, we encode the agent-generated S_{smiles} using SMI-TED, a chemical foundation model pre-trained on 91M molecules. The encoder maps the 1D molecular syntax and associated chemical context into a global semantic vector:

$$\mathbf{z}_{\text{sem}} = \text{Encoder}_{\text{SMI-TED}}(S_{\text{smiles}}) \quad (1)$$

We use $\mathbf{z}_{\text{sem}}$ as the cross-modal prior for the downstream flow model. Compared with discrete token injection, this latent representation provides

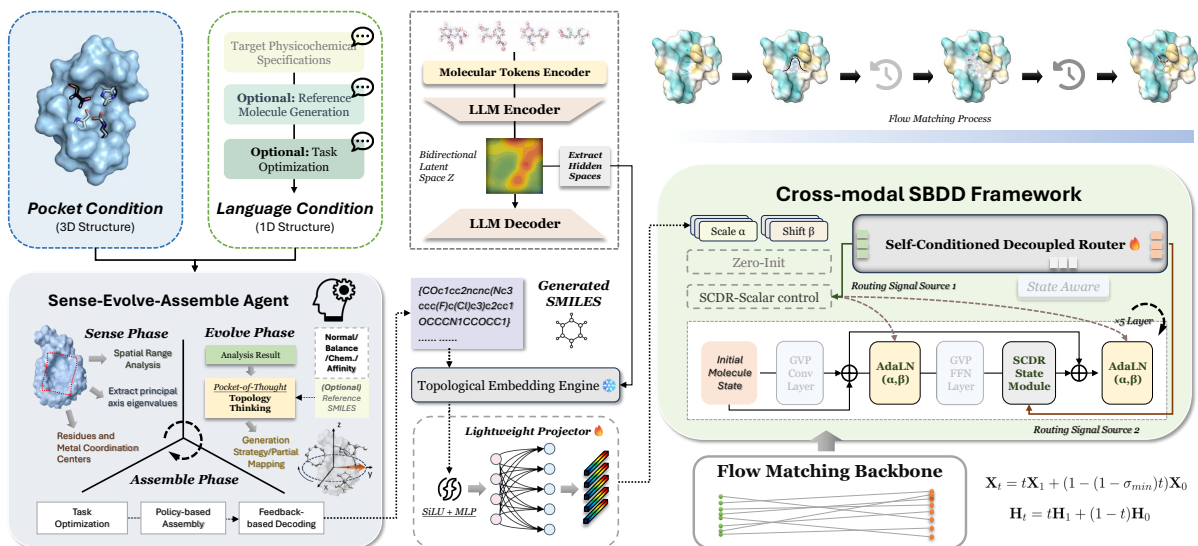


Figure 1: **Overview of the LiFT Framework.** The left panel details the language-conditioned semantic pathway, utilizing the PoT agent, SMILES sanitization, and a foundation encoder to project sequences into dense priors. The right panel illustrates the cross-modal generation framework, injecting these semantic priors into an ODE-governed vector field with SCDR providing state-aware dynamic velocity modulation during flow matching.

a compact continuous interface for conditioning the 3D velocity field with SMILES-derived chemical information. Because SMI-TED is pre-trained on large molecular corpora, $\mathbf{z}_{sem}$ carries pharmacological priors that can bias geometric generation toward property-oriented trends without fine-tuning the generator.

3.3 Scalar Priming via Lightweight Semantic Projector

SBDD requires combining 1D symbolic descriptors ($\mathbf{z}_{sem}$) with 3D geometric configurations. We represent each ligand state as a decoupled pair of invariant scalar features and proper-rotation-equivariant vector features:

$$\mathbf{x} = (\mathbf{s}, \mathbf{V}) \in \mathbb{R}^{N_L \times d_s} \times \mathbb{R}^{N_L \times d_v \times 3} \quad (2)$$

Within this centered frame, $SO(3)$ rotations are the geometric transformations preserved by the reported reflection-sensitive backbone. We therefore apply cross-modal priming only to the scalar domain $\mathbf{s}$. This design keeps the equivariant vectors $\mathbf{V}$ unchanged by the non-equivariant 1D signal, allowing semantic conditioning to enter the model without disrupting the geometric update rules.

A lightweight projector $\mathcal{P}_\phi$ maps the semantic latent into affine modulation parameters:

$$\gamma, \beta = \mathcal{P}_\phi(\mathbf{z}_{sem}) = \mathbf{W}_p \cdot \sigma(\mathbf{z}_{sem}) + \mathbf{b}_p \quad (3)$$

where $\gamma, \beta \in \mathbb{R}^{d_s}$. We employ a **zero-initialization** strategy ($\mathbf{W}_p, \mathbf{b}_p = \mathbf{0}$), so the projec-

tor initially recovers the original scalar normalization path and introduces semantic conditioning only through learned deviations. Let $\text{AdaLN}_{\gamma, \beta}(\mathbf{s}) = \text{LN}(\mathbf{s}) \odot (1 + \tanh(\gamma)) + \beta$. The primed node state is then defined as:

$$\mathbf{s}_{base} = \text{AdaLN}_{\gamma, \beta}(\mathbf{s}), \mathbf{V}_{base} = \mathbf{V}. \quad (4)$$

Thus, $\mathbf{s}_{base}$ carries SMILES-derived priors before state-aware refinement by SCDR.

3.4 Self-Conditioned Decoupled Router (SCDR)

To coordinate semantic conditioning with geometric generation, we propose the *Self-Conditioned Decoupled Router* (SCDR), a cross-modal routing module that adjusts semantic influence according to intermediate 3D structures, thereby reducing conflicts between 1D chemical priors and pocket geometry.

During ODE integration, SCDR summarizes the evolving ligand state through invariant structural statistics. A DeepSets encoder projects node features into $SO(3)$ -invariant latents $\mathbf{z}_{node} = \phi([\mathbf{s} \parallel \|\mathbf{V}\|])$, which are aggregated by a statistical read-out operator Ψ_{stat} :

$$\mathbf{z}_{stat} = [\mathbb{E}(\mathbf{z}_{node}) \parallel \max(\mathbf{z}_{node}) \parallel \sigma(\mathbf{z}_{node}) \parallel \log(N)] \quad (5)$$

where N is the atom count. Combining mean, maximum, dispersion, and scale statistics, $\mathbf{z}_{stat}$ pro-

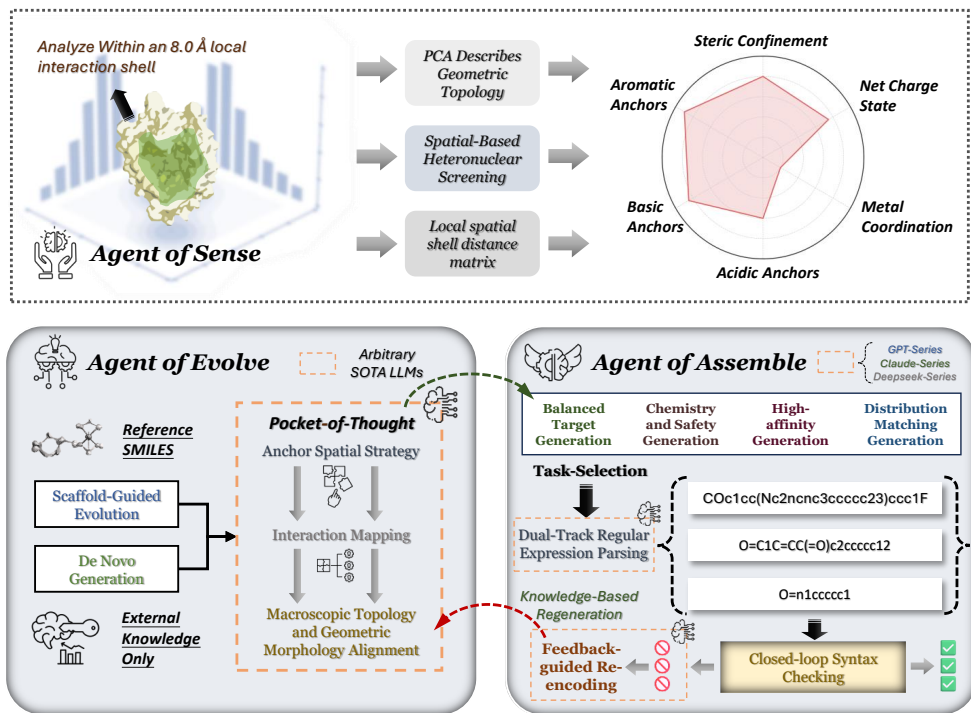


Figure 2: Workflow of the “Sense-Evolve-Assemble” LLM agent. **(Top)** The *Sense* phase extracts spatial geometry and physicochemical anchors within an 8.0 Å pocket interaction shell to build the target profile. **(Bottom Left)** The *Evolve* phase executes Pocket-of-Thought (PoT) reasoning, leveraging text specifications or reference structures for both *de novo* design and scaffold hopping. **(Bottom Right)** The *Assemble* phase performs rule-based sequence serialization, employing deterministic cheminformatics verification to enforce 1D topological and chemical validity.

vides a compact permutation-invariant summary of the current ligand configuration.

To integrate structural context with semantic conditioning, both sources are projected into a shared latent space, yielding $\hat{\mathbf{h}}_{struct}$ and $\hat{\mathbf{h}}_{cond}$, where $\hat{\mathbf{h}}_{cond}$ encodes the semantic prior $\mathbf{z}_{sem}$ and time embedding $\mathbf{t}_{emb}$. Instead of directly adding the two representations, SCDR uses a dual-gated fusion mechanism:

$$g_t = \text{Sigmoid}(\text{MLP}_t(\mathbf{t}_{emb})) \quad (6)$$

$$g_c = \text{Sigmoid}(\text{MLP}_c([\hat{\mathbf{h}}_{cond} \parallel \hat{\mathbf{h}}_{struct}])) \quad (7)$$

The temporal gate g_t controls how much structural feedback is used at different ODE stages, while the context gate g_c modulates the interaction between semantic and structural representations. The unified router latent $\mathbf{h} \in \mathbb{R}^{d_h}$ is obtained by gated residual fusion:

$$\mathbf{h} = \text{LayerNorm}(\text{MLP}_{fuse}(\hat{\mathbf{h}}_{cond} + g_t \cdot g_c \cdot \hat{\mathbf{h}}_{struct})) \quad (8)$$

Conditioned on $\mathbf{h}$, the router branches into two pathways. In the **State Path**, the graph-level latent $\mathbf{h}_{b(i)}$ is broadcast to atom i to infer a bounded scalar

gate $\alpha_{state,i}^s \in \mathbb{R}^{d_s}$:

$$\alpha_{state,i}^s = 1 + w(p)b_{a_s}(\mathbf{W}_{state}\mathbf{h}_{b(i)}) \quad (9)$$

$$\mathbf{s}_{out,i} = \mathbf{s}_{base,i} + \text{LN}(\mathbf{s}_i) \odot (\alpha_{state,i}^s - 1) \quad (10)$$

where $\mathbf{W}_{state} \in \mathbb{R}^{d_s \times d_h}$ is a learnable projection and $b_a(\mathbf{z}) = a\mathbf{z}/(1 + |\mathbf{z}|)$. Setting the state scale $a_s = 0.5$ bounds the gate within $[0.5, 1.5]$ for stable calibration. The vector state bypasses this modulation ($\mathbf{V}_{out,i} = \mathbf{V}_i$), preserving the pocket-centered $SO(3)$ equivariance of the geometric backbone.

Simultaneously, the **Update Path** modulates the geometric feed-forward network $\mathcal{F} = (\mathcal{F}_s, \mathcal{F}_v)$ to regulate residual updates during generation. For each channel $k \in \{s, v\}$, it infers decoupled scaling factors $\mathbf{g}_{up,i}^k$:

$$\mathbf{g}_{up,i}^k = 1 + w(p)b_{a_u}(\mathbf{W}_{up}^k\mathbf{h}_{b(i)}) \quad (11)$$

$$\mathbf{x}_{next,i} = \mathbf{x}_{out,i} + \mathbf{g}_{up,i} \odot \mathcal{F}(\mathbf{x}_{out,i}) \quad (12)$$

where $\mathbf{W}_{up}^k$ are the corresponding learnable weights, and $\mathbf{g}_{up,i} = (\mathbf{g}_{up,i}^s, \mathbf{g}_{up,i}^v)$. A broader update scale $a_u = 0.9$ permits residual scaling within

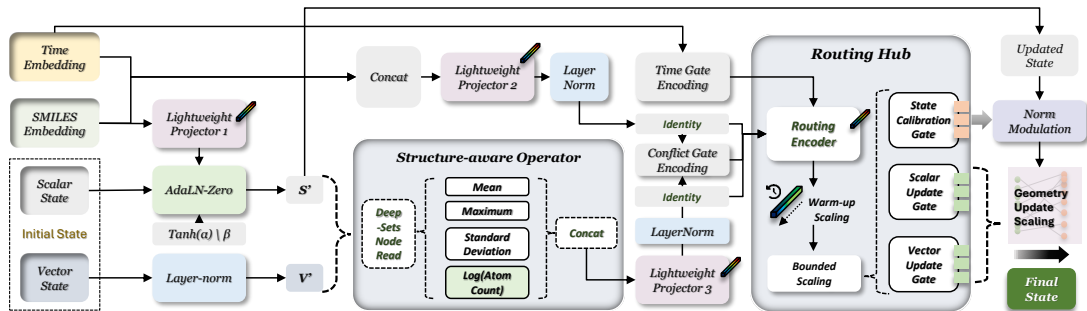


Figure 3: **Detailed architecture of the Self-Conditioned Decoupled Router (SCDR).** Temporal/SMILES priors and intermediate geometric states (S' , V') are contextualized and aggregated by the **Structure-aware Operator** through a multi-scale DeepSets node readout. The **Routing Module** then decodes the fused representation into three decoupled pathways: the *State Calibration Gate* (α_{state}^s), the *Scalar Update Gate* (g_{up}^s), and the *Vector Update Gate* (g_{up}^v) for downstream geometric update scaling.

[0.1, 1.9]. By separating scalar state calibration from residual update modulation, SCDR enables semantic priors to interact with intermediate 3D structural states while maintaining equivariant geometric updates.

4 Experiments

In this section, we evaluate LiFT from two complementary perspectives: distribution matching fidelity and property-oriented generation. We report distribution matching in Sec. 4.2, property-oriented results in Sec. 4.3, and ablation studies in Sec. 4.4. Extended analyses of language-derived conditions are provided in Appendix A and B, while implementation details, extended metrics, and qualitative cases are provided in Appendix C, Appendix E, and Appendix F, respectively.

4.1 Experimental Setup

Datasets and Baselines. We train on the refined CrossDocked2020 dataset (Francoeur et al., 2020) (100K complexes) and compare with autoregressive models (Peng et al., 2022; Luo et al., 2021), 3D diffusion methods (Guan et al., 2023a,b; Huang et al., 2024), flow-matching frameworks (Schneuing et al., 2025; Zhou et al., 2025), and LLM-based SBDD pipelines (Wu et al., 2024; Hu et al., 2026).

Implementation Details. Our framework uses a 5-layer heterogeneous GVP-GNN (Jing et al., 2021) with a frozen SMI-TED encoder (Soares et al., 2025b), trained by continuous flow matching with $T = 500$ ODE steps. Hardware, inference speed, and hyperparameters are provided in Appendix C.4.

Evaluation Metrics. We evaluate 1D pharma-

cological/topological properties (QED, SA, LogP, and ChEMBL ring frequencies), 3D binding and validity (Vina/Gnina and PoseBusters), and filter compliance (REOS and RDKit alerts). For distribution matching, Wasserstein Distance (WD) is computed against the empirical ligand distribution from the 100K training complexes. Metric details are provided in Appendix C.3.

Information Access. Reference-free variants use only pocket-derived textual specifications and task preferences, without access to ground-truth ligand SMILES, graphs, coordinates, docking poses, or ligand-derived labels. Reference-guided variants additionally use an available ligand SMILES only as a scaffold-hopping anchor, not as an evaluation target. Property-oriented variants change only inference-time task instructions after training the base 3D generator, without additional generator fine-tuning or external guidance models.

4.2 Comprehensive Distribution Matching and Generalization

To evaluate distribution matching, we compare generated molecules against the empirical ligand distribution estimated from the 100K CrossDocked2020 training complexes. We report reference-free and reference-guided variants: *Ligand-Ref Embedding* directly encodes the available reference SMILES, whereas *Norm No-Ref* and *Norm Ref* use distribution-oriented LLM prompting without and with a ligand reference. We also include two representative property-oriented variants, *Balanced Ref* and *Vina No-Ref*, to examine reference-guided scaffold hopping and *de novo* design without additional 3D generator fine-tuning.

Although LiFT is not specifically optimized for

Table 1: Comprehensive distribution matching benchmark on CrossDocked2020. All metrics represent the Wasserstein Distance (WD $\downarrow$) between generated molecular-property distributions and the empirical ligand distribution estimated from the 100K training complexes. The evaluation is rigorously divided into Binding Efficiency (Vina, Gnina), Medicinal Chemistry limits (QED, SA, LogP, Rotatable Bonds), and Topological Ring Systems (minimum ring frequencies > 0 , > 10 , > 100). The best results are highlighted in **bold**, the second best are underlined, and the third best are marked with a dagger ($\dagger$).

Method	Binding Eff. (WD $\downarrow$)		MedChem Properties (WD $\downarrow$)				Topological Rings (WD $\downarrow$)		
	Vina	Gnina	QED	SA	LogP	Rot. Bonds	> 0	> 10	> 100
<i>Autoregressive Models</i>									
AR (Luo et al., 2021)	0.032 $\dagger$	<u>0.020</u>	0.036	1.066	1.389	1.370	0.429	0.420	0.402
Pocket2Mol (Peng et al., 2022)	0.047	0.023	0.057	0.127	0.737	3.087	0.320	0.316	0.300
<i>3D Diffusion Architectures</i>									
TargetDiff (Guan et al., 2023a)	0.035	0.031	0.054	1.513	0.497 $\dagger$	0.355 $\dagger$	0.488	0.504	0.498
DecompDiff (Guan et al., 2023b)	0.112	0.070	0.080	1.307	0.643	2.421	0.343	0.355	0.336
BindDM (Huang et al., 2024)	0.018	0.026	0.025 $\dagger$	1.559	0.648	0.723	0.522	0.532	0.533
<i>Flow-Matching Frameworks</i>									
PAFlow (Zhou et al., 2025)	0.053	0.022 $\dagger$	0.050	1.664	2.462	1.392	0.608	0.623	0.602
DrugFlow (Schneuing et al., 2025)	0.042	0.024	<u>0.023</u>	0.243 $\dagger$	<u>0.475</u>	<u>0.233</u>	0.139	0.127	0.098
<i>LLM-based SBDD Pipelines</i>									
TamGen (Wu et al., 2024)	0.376	0.186	0.076	4.303	0.945	5.149	0.028 $\dagger$	<u>0.008</u>	0.047 $\dagger$
ELILLM-Diff (Hu et al., 2026)	0.377	0.190	0.048	5.341	3.424	1.307	0.521	0.540	0.527
ELILLM-Rand (Hu et al., 2026)	0.376	0.184	0.074	4.888	2.152	2.153	0.382	0.414	0.406
<i>LiFT Variants</i>									
LiFT (Ligand-Ref Embedding)	0.034	<u>0.020</u>	0.054	0.437	0.954	0.409	0.260	0.244	0.207
LiFT (Norm No-Ref)	0.056	0.040	0.203	0.538	1.221	1.012	<u>0.027</u>	0.020 $\dagger$	0.010
LiFT (Norm Ref)	0.041	0.029	0.083	<u>0.232</u>	0.374	0.624	0.162	0.163	0.145
LiFT (Balanced Ref)	0.044	0.027	0.020	0.416	0.762	0.229	0.192	0.185	0.165
LiFT (Vina No-Ref)	<u>0.031</u>	0.019	0.197	0.568	1.104	1.400	0.008	0.004	<u>0.028</u>

distribution matching, it achieves competitive WD performance across key metrics. Notably, No-Reference variants perform strongly in binding efficiency and topological ring distributions, surpassing the Ligand-Ref variant on several metrics. Their higher WD on QED and SA reflects a distributional shift rather than uniform degradation: generated molecules show improved pharmacological profiles relative to training-set ligands, which naturally moves them away from the empirical reference distribution, as indicated in Table 2.

4.3 Property-Oriented Generation without Additional Generator Fine-Tuning

Moving beyond distribution matching, we evaluate property-oriented generation in Table 2. *No-Reference* variants generally outperform their *Reference* counterparts on core medicinal-chemistry and topological metrics, suggesting that strict anchoring to empirical ligands can constrain language-derived priors, whereas reference-free prompting allows exploration of alternative valid chemical regions. Accordingly, *No-Reference* configurations

improve QED up to 0.757 and SA down to 2.659, while maintaining strong ring-frequency alignment (78.86% for rings > 10), filter compliance (RD-Kit/REOS $> 71\%$), and competitive PoseBusters validity (up to 73.56%). Detailed subindicators are provided in Appendix E.

Task-specific prompts further induce directional property shifts: Vina-oriented instructions bias generation toward higher binding affinity, whereas QED-centric instructions improve drug-likeness. Rather than deterministic optimization, this semantic conditioning acts as a trend-guided prior, with text specifications corresponding to downstream empirical scores as analyzed in Appendix B. Overall, LiFT improves medicinal-chemistry profiles and filter compliance while preserving competitive binding and structural-validity indicators, suggesting a practical trade-off between language-derived property steering and 3D pocket compatibility.

4.4 Ablation Studies and Diagnostic Analyses

Structural Module Ablation. Table 3 evaluates the main conditioning modules. The Ligand-

Table 2: Property-oriented generation without additional generator fine-tuning on CrossDocked2020. Binding efficiency metrics (Gnina $\uparrow$, Vina $\downarrow$) are reported as absolute means. MedChem metrics (QED $\uparrow$, SA $\downarrow$) reflect the mean molecular properties, and topological/safety filters (Rings $\uparrow$, PoseB./RDKit/REOS $\uparrow$) are reported as pass rates (%). Best results are highlighted in **bold**, second best underlined, third best $\dagger$.

Category	Method	Binding Aff.		MedChem Means		Rings Freq (%) $\uparrow$			Filters (%) $\uparrow$		
		Gnina $\uparrow$	Vina $\downarrow$	QED $\uparrow$	SA $\downarrow$	>0	>10	>100	RDKit	REOS	PoseB.
3D-based Models	AR (Luo et al., 2021)	<u>0.295</u>	-0.414 $\dagger$	0.509	4.289	40.99	36.48	31.09	52.69	43.17	59.86
	PAFlow (Zhou et al., 2025)	0.262	-0.429	0.491	4.889	23.02	16.20	10.99	75.55	58.28	14.68
	BindDM (Huang et al., 2024)	0.249	-0.362	0.508	4.784	31.68	25.31	17.98	69.81	56.83	31.93
	DecompDiff (Guan et al., 2023b)	0.205	-0.267	0.454	4.532	49.58	43.01	37.60	60.27	44.62	54.92
	TargetDiff (Guan et al., 2023a)	0.244	-0.344	0.480	4.738	35.08	28.07	21.41	65.19	53.17	51.28
	DrugFlow (Schneuing et al., 2025)	0.252	-0.338	0.553	3.43	69.97	65.80	61.41	75.86	64.84	78.45
	Pocket2Mol (Peng et al., 2022)	0.297	<u>-0.425</u>	0.573	3.198	51.83	46.90	41.22	83.46	65.95	72.02
LLM-based Models	TamGen (Wu et al., 2024)	0.089	-0.003	0.460	7.528	81.02 $\dagger$	37.08	66.50	37.08	57.75	5.64
	ELILLM-Diff (Hu et al., 2026)	0.085	-0.002	0.487	8.567	31.78	24.50	18.53	22.62	71.31 $\dagger$	3.60
	ELILLM-Rand (Hu et al., 2026)	0.091	-0.003	0.460	8.113	45.66	37.06	30.67	27.97	65.39	4.14
LiFT Variants	Ligand-Ref Embedding	0.263 $\dagger$	-0.352	0.480	3.660	57.88	54.07	50.54	59.69	47.14	71.74
	LiFT (QED-Reference)	0.242	-0.329	0.521	3.694	63.90	59.06	54.20	60.20	48.62	67.17
	LiFT (Vina-Reference)	0.243	-0.337	0.515	3.757	62.55	57.73	53.28	61.37	49.90	68.34
	LiFT (Bal-Reference)	0.247	-0.335	0.533	3.640	64.65	59.96	54.72	63.52	51.25	68.31
	LiFT (QED-No-Reference)	0.243	-0.334	<u>0.744</u>	2.724 $\dagger$	80.43	74.88 $\dagger$	69.09 $\dagger$	83.19 $\dagger$	71.10	72.21 $\dagger$
	LiFT (Vina-No-Reference)	0.255	-0.373	0.732 $\dagger$	<u>2.662</u>	83.08	78.86	74.07	<u>83.36</u>	73.81	<u>73.56</u>
	LiFT (Bal-No-Reference)	0.241	-0.341	0.757	2.659	<u>81.92</u>	<u>77.76</u>	<u>73.84</u>	81.27	<u>71.78</u>	70.73

Ref Embedding setting serves as a reference-conditioned diagnostic baseline for the downstream generator. Removing zero-initialization slightly weakens binding, QED, and filter compliance, suggesting that zero-initialized modulation helps inject SMILES-derived priors without destabilizing the flow backbone. Removing SCDR causes larger drops in QED, RDKit, and REOS, indicating that state-aware routing mainly benefits the drug-likeness and filter compliance of semantically steered generation.

Cross-LLM Robustness and Selection. Benchmarking GPT-4o, Claude-4-Sonnet, and DeepSeek-V3 shows relatively small performance variation, suggesting that LiFT is not tied to a single language engine. DeepSeek-V3 achieves the strongest filter compliance, while GPT-4o provides the best QED and a stable property-oriented profile; we therefore use GPT-4o as the default backbone.

Table 3: Ablation analysis of core components and cross-LLM inference performance.

Model Variant	Gnina $\uparrow$	Vina $\downarrow$	QED $\uparrow$	SA $\downarrow$	RDKit% $\uparrow$	REOS% $\uparrow$
<i>Structural/Component Ablation</i>						
LiFT (Ligand-Ref Embedding)	0.263	-0.352	0.480	3.660	59.69	47.14
w/o Zero-Init	0.248	-0.339	0.470	3.660	58.00	47.00
w/o SCDR (naive semantic injection)	0.245	-0.335	0.464	3.591	57.28	44.62
<i>Generalization Across LLM Backbones</i>						
LiFT (GPT-4o)	0.241	-0.341	0.757	2.659	<u>81.27</u>	71.78
LiFT (Claude-4-Sonnet)	0.249	-0.330	0.700	2.777	78.20	<u>73.98</u>
LiFT (DeepSeek-V3)	<u>0.258</u>	<u>-0.346</u>	<u>0.707</u>	<u>2.660</u>	83.44	75.03

Language-Condition Diagnostics. Beyond architectural ablations, Appendix A and Appendix B examine LLM-derived molecular conditions from

complementary perspectives. Appendix A compares random, retrieved, direct-prompted, ablated, and full Sense-Evolve-Assemble SMILES conditions under the same 3D generator, testing whether structured language conditions provide more balanced guidance than arbitrary or retrieval-based priors. Appendix B further analyzes whether these conditions retain trend-level semantic effects after cross-modal projection: Appendix B.1 measures property-level correspondence between SMILES conditions and final molecules, Appendix B.2 compares generated and empirical ligand SMILES in ECFP4 space, and Appendix B.3 profiles SCDR gate behavior across training, ODE trajectories, and latent channels.

5 Conclusion

In this study, we introduce **LiFT**, a language-informed cross-modal framework built on flow matching for trend-guided 3D molecular generation. Our evaluations show (i) strict ligand-reference anchoring can constrain exploration, whereas reference-free *de novo* prompting enables exploration of valid chemical regions; and (ii) property-oriented prompting can improve medicinal chemistry profiles while maintaining competitive binding and structural validity. These results demonstrate the potential of language-derived chemical priors as semantic conditions for pocket-conditioned 3D generation, enabling trend-level guidance without geometric-backbone fine-tuning.

Limitations

While LiFT provides trend-guided language conditioning for SBDD, challenges in cross-modal generation remain. Natural language captures useful chemical and pharmacological preferences but lacks the geometric granularity for atomic-level control. Accordingly, LiFT provides trend-level guidance rather than exact property or structural control. Since LiFT relies on LLM-derived symbolic priors, its performance may vary with the chemical competence and reliability of the underlying LLM, especially where text-based chemical knowledge is sparse or biased. Future work may explore intermediate representations that better bridge language-level chemical intent and evolving 3D states, as well as richer modalities such as cryo-EM density maps or quantum interaction graphs.

Current SBDD evaluation lacks consensus on balancing distributional fidelity, such as Wasserstein Distance, with absolute property-oriented improvement. Validation on the static Cross-Docked2020 benchmark enables controlled comparison but does not establish generalization across broader pocket distributions or capture real biological dynamics. Extending LiFT to additional targets, datasets, and wet-lab feedback remains important.

References

- Amy C. Anderson. 2003. [The process of structure-based drug design](#). *Chemistry & Biology*, 10(9):787–797.
- Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2022. [MolGPT: Molecular generation using a transformer-decoder model](#). *Journal of Chemical Information and Modeling*, 62(9):2064–2076.
- Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. 2024. [Posebusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences](#). *Chemical Science*, 15(9):3130–3139.
- Zhanpeng Chen, Weihao Gao, Shunyu Wang, Yanan Zhu, Hong Meng, and Yuexian Zou. 2025. [MolSculpt: Sculpting 3D molecular geometries from chemical syntax](#). *Preprint*, arXiv:2512.10991.
- Seungyeon Choi, Hwanhee Kim, Chihyun Park, Dahyeon Lee, Seungyong Lee, Yoonju Kim, Hyoungjoon Park, Sein Kwon, Youngwan Jo, and Sanghyun Park. 2025. [Controllable 3D molecular generation for structure-based drug design through bayesian flow networks and gradient integration](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 125612–125647. Curran Associates, Inc.
- Ian Dunn and David R. Koes. 2026. [FlowMol3: Flow matching for 3D de novo small-molecule generation](#). *Digital Discovery*, 5(5):2052–2066.
- Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David Ryan Koes. 2020. [Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design](#). *Journal of Chemical Information and Modeling*, 60(9):4200–4215.
- Bowen Gao, Yanwen Huang, Yiqiao Liu, Wenxuan Xie, Bowei He, Haichuan Tan, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. 2025. [Cidd: Collaborative intelligence for structure-based drug design empowered by llms](#). In *Advances in Neural Information Processing Systems*.
- Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 2023a. [3D equivariant diffusion for target-aware molecule generation and affinity prediction](#). In *International Conference on Learning Representations*.
- Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. 2023b. [Decompdiff: Diffusion models with decomposed priors for structure-based drug design](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11827–11846. PMLR.
- Yingqing Guo, Yukang Yang, Hui Yuan, and Mengdi Wang. 2025. [Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Xuanning Hu, Anchen Li, Qianli Xing, Jinglong Ji, Hao Tuo, and Bo Yang. 2026. [Empowering llms for structure-based drug design via exploration-augmented latent inference](#). *Preprint*, arXiv:2601.15333.
- Zhilin Huang, Ling Yang, Zaixi Zhang, Xiangxin Zhou, Yu Bao, Xiawu Zheng, Yuwei Yang, Yu Wang, and Wenming Yang. 2024. [Binding-adaptive diffusion models for structure-based drug design](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12671–12679.
- Lei Jiang, Shuzhou Sun, Biqing Qi, Yuchen Fu, Xiaohua Xu, Yuqiang Li, Dongzhan Zhou, and Tianfan Fu. 2025. [Chem3DLLM: 3D multimodal large language models for chemistry](#). *Preprint*, arXiv:2508.10696.
- Jirui Jin, Cheng Zeng, Pawan Prakash, Ellad B. Tadmor, Adrian Roitberg, Richard G. Hennig, Stefano Martiniani, and Mingjie Liu. 2026. [Molguidance: Advanced guidance strategies for conditional molecular generation with flow matching](#). *Journal of Chemical Information and Modeling*, 66(15):8860–8874.

- Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. 2021. [Learning from protein structure with geometric vector perceptrons](#). In *International Conference on Learning Representations*.
- Antonios Kritikos, Nikolaos Spanos, and Athanasios Voulodimos. 2026. [Crossflowdg: Bridging the modality gap with cross-modal flow matching for domain generalization](#). *Preprint*, arXiv:2604.16892.
- Xiaolei Lang, Yang Wang, Yukun Zhou, Chaojun Ni, Kerui Li, Jiagang Zhu, Tianze Liu, Jiajun Lv, Xingxing Zuo, Yun Ye, Guan Huang, Xiaofeng Wang, and Zheng Zhu. 2026. [VAG: Dual-stream video-action generation for embodied data synthesis](#). *Preprint*, arXiv:2604.09330.
- Haitao Lin, Guojiang Zhao, Odin Zhang, Yufei Huang, Lirong Wu, Cheng Tan, Zicheng Liu, Zhifeng Gao, and Stan Z. Li. 2025. [CBGBench: Fill in the blank of protein-molecule complex binding graph](#). In *The Thirteenth International Conference on Learning Representations*. Spotlight.
- Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023. [MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638.
- Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2025. [NExT-Mol: 3D diffusion meets 1D language modeling for 3D molecule generation](#). In *The Thirteenth International Conference on Learning Representations*.
- Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. 2021. [A 3D generative model for structure-based drug design](#). In *Advances in Neural Information Processing Systems*, pages 6229–6239.
- William Peebles and Saining Xie. 2023. [Scalable diffusion models with transformers](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.
- Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. 2022. [Pocket2Mol: Efficient molecular sampling based on 3D protein pockets](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17644–17655. PMLR.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. [FiLM: Visual reasoning with a general conditioning layer](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Jerret Ross, Brian M. Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. [Large-scale chemical language representations capture molecular structure and properties](#). *Nature Machine Intelligence*, 4(12):1256–1264.
- Victor Garcia Satorras, Emiel Hoogetboom, and Max Welling. 2021. [E\(n\) equivariant graph neural networks](#). In *Proceedings of the 38th International Conference on Machine Learning*.
- Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom Blundell, Pietro Lio, Max Welling, Michael Bronstein, and Bruno Correia. 2024. [Structure-based drug design with equivariant diffusion models](#). *Nature Computational Science*.
- Arne Schneuing, Ilia Igashov, Adrian W. Dobbelsstein, Thomas Castiglione, Michael M. Bronstein, and Bruno Correia. 2025. [Multi-domain distribution learning for de novo drug design](#). In *The Thirteenth International Conference on Learning Representations*.
- Guransh Singh. 2026. [AEGIS: Anchor-enforced gradient isolation for knowledge-preserving vision-language-action fine-tuning](#). *Preprint*, arXiv:2604.16067.
- Eduardo Soares, Emilio Vital Brazil, Victor Shira-suna, Dmitry Zubarev, Renato Cerqueira, and Kristin Schmidt. 2025a. [A mamba-based foundation model for materials](#). *npj Artificial Intelligence*, 1:8.
- Eduardo Soares, Emilio Vital Brazil, Victor Shira-suna, Dmitry Zubarev, Renato Cerqueira, and Kristin Schmidt. 2025b. [An open-source family of large encoder-decoder foundation models for chemistry](#). *Communications Chemistry*, 8:193.
- W. Patrick Walters and Mark A. Murcko. 2020. [Assessing the impact of generative AI on medicinal chemistry](#). *Nature Biotechnology*, 38:143–145.
- Kehan Wu, Yingce Xia, Pan Deng, Renhe Liu, Yuan Zhang, Han Guo, Yumeng Cui, Qizhi Pei, Lijun Wu, Shufang Xie, Si Chen, Xi Lu, Song Hu, Jinzhi Wu, Chi Kin Chan, Shawn Chen, Liangliang Zhou, Neng-hai Yu, Enhong Chen, and 4 others. 2024. [Tamgen: drug design with target-aware molecule generation through a chemical language model](#). *Nature Communications*, 15(1):9360.
- Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. 2017. [Deep sets](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Daiheng Zhang, Chengyue Gong, and Qiang Liu. 2024a. [Rectified flow for structure based drug design](#). *Preprint*, arXiv:2412.01174.
- Wei Zhang, Zekun Guo, Yingce Xia, Peiran Jin, Shufang Xie, Tao Qin, and Xiang-Yang Li. 2025. [Molchord: Structure-sequence alignment for protein-guided drug design](#). *Preprint*, arXiv:2510.27671.

Zaixi Zhang, Jiaxian Yan, Yining Huang, Qi Liu, Enhong Chen, Mengdi Wang, and Marinka Zitnik. 2024b. [Geometric deep learning for structure-based drug design: A survey](#). *Preprint*, arXiv:2306.11768.

Jingyuan Zhou, Hao Qian, Shikui Tu, and Lei Xu. 2025. [Prior-guided flow matching for target-aware molecule design with learnable atom number](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jie Zhu, Mingyu Ding, Boqiang Duan, Leye Wang, and Jingdong Wang. 2026. [Unveiling the secret of adaln-zero in diffusion transformer](#). *Preprint*, arXiv:2608.09438.

Appendix Overview

The supplementary material is organized into eight complementary parts that further analyze the role of language-derived molecular conditions, provide reproducibility details, extend the experimental diagnostics beyond the main paper, and clarify the methodological positioning of LiFT.

- **Appendix A: Extended Analysis of Language-Derived Molecular Conditions.** This appendix analyzes the intermediate SMILES conditions generated by the language agent before geometric generation. It includes condition-source ablations, 1D semantic evaluations, downstream 3D prior effects, and analyses of how structured language-derived conditions influence molecular generation under a fixed 3D backbone.
- **Appendix B: Semantic Preservation and Diagnostic Analyses.** This appendix examines whether language-derived molecular conditions retain measurable trend-level effects after cross-modal projection into 3D generation. It includes property-level correspondence analysis, ECFP4 manifold comparisons, and SCDR routing diagnostics across training and ODE trajectories.
- **Appendix C: Experimental Details and Reproducibility.** This appendix reports datasets, implementation settings, hyperparameters, evaluation protocols, hardware configurations, runtime and efficiency, pocket-level uncertainty estimates, absolute docking scores, and additional reproducibility details.
- **Appendix D: Controlled SMILES-Condition Protocols and Prompting**

Templates. This appendix describes the controlled prompting interfaces, information-access boundaries, task-specific steering directives, and structured generation templates used by the language agent.

- **Appendix E: Rule-Level Structural, Med-Chem, and Pose Validity Diagnostics.** This appendix decomposes aggregate RDKit, REOS, and PoseBusters metrics into fine-grained rule-level analyses and geometric validity sub-checks.
- **Appendix F: Qualitative Case Studies on Representative Targets.** This appendix provides qualitative analyses of generated molecules under different conditioning settings, including comparisons of structural validity, medicinal chemistry properties, and pocket compatibility.
- **Appendix G: Detailed Mathematical Formulations.** This appendix contains extended derivations and theoretical details for the proposed flow-matching and conditioning framework.
- **Appendix H: Extended Comparison with Related Work.** This appendix provides a systematic methodological comparison between LiFT and representative language-based, 1D-3D, and native pocket-conditioned SBDD methods, together with a focused comparison with ELILLM.

A Extended Analysis of Language-Derived Molecular Conditions

This appendix provides an extended analysis of the intermediate 1D SMILES conditions used by LiFT. The main experiments in the paper evaluate the final 3D generated molecules, while this appendix focuses on the symbolic molecular conditions before they are encoded by SMI-TED and injected into the 3D flow model. The purpose of this analysis is to isolate whether the proposed language-derived conditions are chemically meaningful and structurally diverse, rather than merely acting as arbitrary SMILES priors.

We evaluate six condition sources under the same reference-free balanced generation setting. All variants are evaluated on the same target-pocket

set and produce SMILES conditions that are subsequently used by the same downstream 3D generation pipeline. The only difference among these variants is how the intermediate SMILES condition is obtained.

- **C0-random**: valid SMILES are randomly sampled from the training ligand pool. This setting controls for whether injecting any valid molecular string is sufficient to provide useful chemical priors.
- **C1-retrieved**: SMILES are selected from the training ligand pool through a non-LLM retrieval procedure. This setting represents a stronger non-generative molecular-prior baseline than random sampling, while avoiding direct access to test ligand SMILES.
- **C2-direct**: the LLM directly generates SMILES from the pocket description and task preference, without the structured Sense-Evolve-Assemble procedure. This setting controls for the effect of ordinary direct prompting.
- **C3-w/o-Sense**: the language agent generates SMILES without the explicit pocket-sensing stage. This removes the structured extraction of local pocket features, such as geometric shape, charge state, and residue anchors.
- **C4-w/o-PoT**: the agent receives the pocket information but does not use Pocket-of-Thought reasoning before SMILES generation. This setting isolates the role of structured intermediate reasoning.
- **C5-full**: the full Sense-Evolve-Assemble agent used in LiFT. This setting combines pocket sensing, Pocket-of-Thought reasoning, and task-aware molecular assembly to produce the final SMILES condition.

This comparison separates three sources of signal: generic valid molecular priors (**C0-random**), non-LLM retrieved molecular priors (**C1-retrieved**), and LLM-derived molecular conditions with different levels of structured reasoning (**C2-C5**). Therefore, the analysis directly tests whether the full language agent provides more chemically structured 1D conditions than simpler alternatives.

A.1 1D Semantic Condition Analysis

We first evaluate the intermediate SMILES conditions before they are injected into the 3D generator. This analysis isolates whether the language-derived molecular conditions are chemically meaningful in their own right, rather than merely serving as arbitrary textual inputs to the downstream flow model. All variants are evaluated on the same target set and differ only in how the 1D SMILES condition is obtained.

Using the six condition sources defined above, we first evaluate the resulting SMILES before they are injected into the 3D generator. Table 4 reports representative 1D condition-quality metrics. We focus on validity, uniqueness, scaffold uniqueness, QED, SA, LogP, ring count, and scaffold novelty with respect to the training set. These metrics cover basic chemical validity, molecular diversity, medicinal-chemistry plausibility, and whether the generated conditions collapse to training-set scaffolds.

Several observations emerge. First, validity is saturated across all variants, indicating that the comparison is not driven by invalid SMILES filtering. Second, the retrieved baseline has zero scaffold novelty by construction and substantially lower scaffold diversity than generated conditions, showing that retrieval provides chemically valid but non-novel molecular priors. Third, direct prompting achieves strong QED and SA, but it produces lower scaffold uniqueness and simpler ring profiles, suggesting that direct LLM generation tends to optimize local drug-likeness without maintaining the same level of topological diversity. Finally, the full agent achieves the highest uniqueness, high scaffold uniqueness, strong scaffold novelty, and a balanced property profile. This suggests that the Sense-Evolve-Assemble design provides useful 1D molecular conditions beyond random, retrieved, or directly prompted SMILES.

A.2 Downstream Effect of 1D Condition Sources

We next examine whether the differences observed at the 1D condition level translate into downstream 3D molecular generation behavior. For this analysis, all condition sources are passed through the same frozen SMI-TED encoder and the same trained 3D flow generator. The sampling configuration and evaluation pipeline are also fixed, so the only changing factor is the source of the intermedi-

Table 4: 1D molecular condition quality under different SMILES condition sources. All variants are evaluated before downstream 3D generation. C0 samples random valid SMILES from the training ligand pool; C1 retrieves training-set SMILES; C2 uses direct LLM prompting; C3 removes explicit pocket-sensing information; C4 removes Pocket-of-Thought reasoning; and C5 is the full Sense-Evolve-Assemble agent. The full agent preserves perfect validity, achieves the highest uniqueness, maintains high scaffold diversity, and produces chemically plausible property profiles without collapsing to retrieved training-set scaffolds. Boldface is omitted because this table is intended to compare overall trade-offs among validity, diversity, novelty, and medicinal-chemistry plausibility rather than single-metric dominance.

Condition Source	Validity	Unique	Scaf. Unique	QED	SA	LogP	Rings	Novel Scaf.
C0 Random	1.000	0.987	0.777	0.543	3.241	2.101	2.877	0.000
C1 Retrieved	1.000	0.767	0.000	0.673	2.512	3.647	1.977	0.000
C2 Direct	1.000	0.947	0.624	0.801	2.048	2.583	1.466	0.641
C3 w/o Sense	1.000	0.987	0.832	0.777	2.151	3.059	2.096	0.857
C4 w/o PoT	1.000	0.963	0.786	0.754	2.129	2.938	2.390	0.814
C5 Full	1.000	0.993	0.824	0.778	2.239	2.529	2.672	0.750

Table 5: Downstream 3D generation results under different 1D SMILES condition sources. All variants use the same SMI-TED encoder, 3D flow generator, sampling configuration, and evaluation pipeline; only the source of the intermediate SMILES condition changes. C0 samples random valid SMILES from the training ligand pool; C1 uses retrieved training-set SMILES; C2 directly prompts the LLM to generate SMILES; C3 removes explicit pocket-sensing information; C4 removes Pocket-of-Thought reasoning; and C5 is the full Sense-Evolve-Assemble agent. Boldface is omitted because this table evaluates trade-offs across medicinal chemistry, binding-efficiency, topological, and structural-validity indicators rather than optimizing a single metric.

Condition	QED $\uparrow$	SA $\downarrow$	VinaEff $\downarrow$	GninaEff $\uparrow$	LogP	Rot. $\uparrow$	Ring $>0\uparrow$	Ring $>10\uparrow$	Ring $>100\uparrow$	PB-Int. $\uparrow$
C0 Random	0.579	3.699	-0.302	0.233	1.766	4.755	0.629	0.583	0.512	0.794
C1 Retrieved	0.693	2.855	-0.318	0.225	3.159	3.931	0.837	0.824	0.797	0.856
C2 Direct	0.760	2.683	-0.345	0.236	2.547	2.925	0.777	0.713	0.666	0.900
C3 w/o Sense	0.749	2.531	-0.343	0.238	3.117	2.584	0.766	0.706	0.656	0.881
C4 w/o PoT	0.736	2.732	-0.329	0.234	2.851	3.110	0.748	0.680	0.615	0.900
C5 Full	0.757	2.659	-0.331	0.241	2.347	3.716	0.819	0.778	0.738	0.902

ate SMILES condition.

Table 5 reports downstream 3D generation results under the six condition sources defined in Appendix A.1. Random SMILES conditions lead to the weakest overall performance, with lower QED, worse SA, weaker binding-efficiency scores, and lower PoseBusters interaction-energy validity. This confirms that injecting an arbitrary valid molecular prior is insufficient for effective 3D generation. Retrieved SMILES provide stronger ring-frequency statistics, which is expected because retrieved molecules are drawn from the empirical training ligand pool. However, this retrieval-based prior does not yield the best binding-related scores and also shows a relatively high LogP, suggesting that simply reusing training-set molecular priors does not provide the most effective target-aware guidance.

Direct LLM prompting achieves the highest QED and the best Vina efficiency, indicating that an LLM can generate useful drug-like conditions even

without structured reasoning. However, its ring-frequency metrics are consistently lower than those of the full agent, suggesting that direct prompting tends to favor local medicinal-chemistry optimization while providing less balanced topological guidance. Removing the Sense phase or Pocket-of-Thought reasoning also leads to weaker overall trade-offs. The w/o Sense variant obtains the lowest SA score, but it has reduced ring-frequency coverage and lower PoseBusters interaction-energy validity than the full model. Similarly, w/o PoT weakens ring statistics and does not improve binding-efficiency metrics.

The full Sense-Evolve-Assemble agent provides the most balanced downstream behavior. Although it is not the best on every single metric, it achieves the best Gnina efficiency and the highest PoseBusters interaction-energy validity, while maintaining strong QED, SA, and ring-frequency statistics. Compared with Direct LLM and w/o PoT, the full agent better preserves topological richness after 3D

generation. Compared with Retrieved SMILES, it avoids collapsing to empirical training-set priors while producing stronger binding-related and structural-validity indicators. These results support the view that structured language-derived SMILES conditions act as informative soft priors for 3D generation, rather than arbitrary molecular strings or simple retrieval anchors.

Together with the 1D condition-quality analysis in Appendix A.1, these results indicate that the proposed language agent improves both the intermediate symbolic conditions and their downstream effect after cross-modal projection into 3D molecular generation.

B Semantic Preservation and Diagnostic Analyses

This appendix provides additional diagnostic analyses for the language-conditioned generation pipeline. Rather than treating the LLM-generated SMILES as a black-box textual input, we examine whether these 1D conditions remain chemically meaningful at different stages of the framework. The goal is not to prove exact preservation or one-to-one recovery of ground-truth ligands. Instead, we ask whether the agent-generated SMILES and their downstream 3D generations preserve measurable trend-level and distribution-level relationships with empirical molecular data.

We study this question from three complementary perspectives. Appendix B.1 evaluates whether property tendencies encoded in the agent-generated SMILES are reflected in the calculated properties of the final generated 3D molecules. Appendix B.2 analyzes the topo-chemical distribution of the agent-generated SMILES themselves by comparing their ECFP4 fingerprint embeddings with the 100 held-out ligand SMILES from the evaluated pockets. Appendix B.3 further profiles the learned behavior of the State Calibration Gate, Scalar Update Gate, and Vector Update Gate across training, ODE inference, and latent-channel distributions. Together, these analyses provide diagnostic evidence that the language-derived conditions are not arbitrary text strings, but chemically structured inputs that remain informative before and during 3D generation.

Table 6: Cross-modal trend correspondence between text-derived 1D conditioning targets and the final generated 3D molecular ensembles under the *LiFT (Balanced No-Ref)* setting. Pearson correlations are computed at the pocket-ensemble level over 100 target pockets.

Property	Pearson r	p -value
QED	0.7344	3.457×10^{-18}
LogP	0.8379	1.646×10^{-27}
Heavy atoms	0.7392	1.613×10^{-18}
Rotatable bonds	0.7619	3.427×10^{-20}
6-ring count	0.7578	7.066×10^{-20}

B.1 Property-Level Trend Correspondence under Cross-Modal Projection

To examine whether text-derived conditions remain reflected in the generated molecules, we measure the correlation between agent-generated SMILES-derived conditions and calculated molecular properties after 3D generation. This analysis is conducted at the ensemble level rather than the single-molecule level. For each target pocket, we average the properties of its generated ligands, which reduces sampling noise and focuses on macroscopic generation trends.

Under the unconstrained *LiFT (Balanced No-Ref)* setting, we generated more than 6,000 ligands for 100 target pockets from CrossDocked2020. For each pocket, we computed the average property value over its generated ligand ensemble and compared it with the corresponding text-derived conditioning target. As shown in Figure 4 and Table 6, the generated ensembles show consistently positive correlations across multiple physicochemical and topological dimensions. All reported correlations are statistically significant, suggesting that language-derived conditions preserve measurable trend-level signals after cross-modal projection into 3D molecular generation.

These results should be interpreted as evidence of trend-level correspondence rather than exact property control. The generated molecules are not expected to match the conditioning values point by point, since the final structures are also constrained by pocket geometry, flow dynamics, and chemical validity. Nevertheless, the observed correlations suggest that the language-derived conditions remain informative after projection into the 3D generation process. This supports the use of text-derived priors as soft guidance signals for property-oriented generation, while leaving the precise causal role of each architectural component to the ablation and

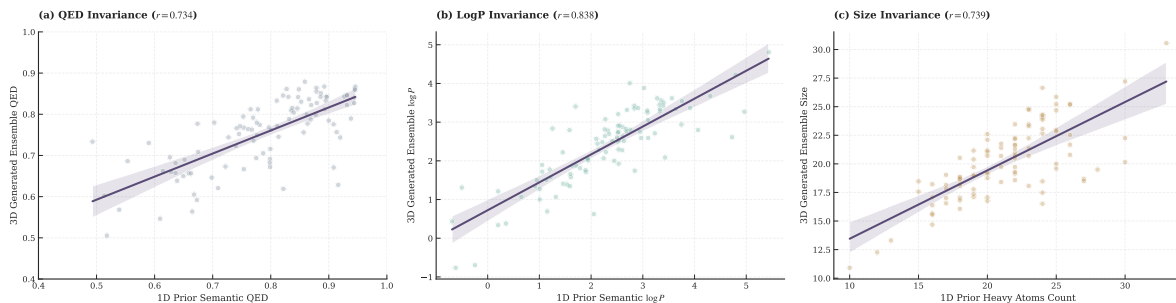


Figure 4: **Property-level trend correspondence under the *LiFT (Balanced No-Ref)* setting.** The panels show ensemble-level correlations between properties derived from the agent-generated SMILES conditions and calculated properties of the final generated molecules across 100 test pockets: (a) QED ($r = 0.734$), (b) $\log P$ ($r = 0.838$), and (c) molecular size measured by heavy atom count ($r = 0.739$). All reported Pearson correlations are statistically significant ($p < 10^{-17}$). These results indicate that the SMILES-derived conditions preserve measurable property-level trends after cross-modal projection and 3D generation, without implying exact molecule-level regression.

diagnostic analyses.

B.2 Topo-Chemical Distribution of Agent-Generated SMILES

While Appendix B.1 examines whether SMILES-derived property tendencies are reflected after 3D generation, we further evaluate the agent-generated SMILES before they are injected into the geometric generator. This analysis asks whether the 1D molecules proposed by the language agent occupy topo-chemical regions comparable to the empirical ligand SMILES from the test set.

For each molecule, we compute 2,048-bit Morgan fingerprints (ECFP4, radius = 2) to represent local substructural patterns. We then project these high-dimensional fingerprints into two dimensions using *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE). The comparison is performed between the agent-generated SMILES and the 100 ground-truth ligand SMILES corresponding to the test pockets. Since *t*-SNE is stochastic and depends on the full input matrix of each run, absolute coordinates should not be compared across panels. We therefore use the visualization only as a qualitative diagnostic of relative overlap, separation, and neighborhood structure within each panel.

As shown in Figure 5a, the reference-guided *LiFT (Balanced Ref)* setting produces SMILES whose ECFP4 embeddings broadly overlap with the ground-truth ligand SMILES. This behavior is expected because the reference-guided setting provides an explicit molecular anchor to the agent. The observed overlap suggests that, when reference information is available, the agent tends to generate 1D molecular structures with local substructural patterns similar to those of the empirical ligands.

In the reference-free *LiFT (Balanced No-Ref)* setting, shown in Figure 5b, the agent-generated SMILES display a clearer shift relative to the ground-truth distribution. This is also expected, since the agent does not receive a reference molecule and therefore has more freedom to propose alternative scaffolds. Importantly, the generated SMILES remain close to the empirical distribution in the projected fingerprint space, rather than forming a visually detached cluster. This suggests that the reference-free agent explores nearby topo-chemical regions while still preserving recognizable substructural similarity under the ECFP4 representation.

Overall, this analysis complements the property-level correspondence in Appendix B.1. Appendix B.1 shows that the agent-generated SMILES provide property-level signals that remain visible after 3D generation. In contrast, the present analysis examines the SMILES conditions themselves and shows that their fingerprint distributions remain close to the ground-truth ligand SMILES, especially under reference-guided generation. These results do not prove exact recovery of empirical ligands, nor do they imply that *t*-SNE distances are chemically quantitative. They provide a qualitative check that the language agent produces chemically structured 1D conditions before the downstream 3D flow model uses them.

B.3 Holistic Spatiotemporal Diagnostics of SCDR

Figure 6 provides a diagnostic view of the learned behavior of the Self-Conditioned Decoupled Router (SCDR). The goal of this analysis is not to repeat the architectural design of SCDR,

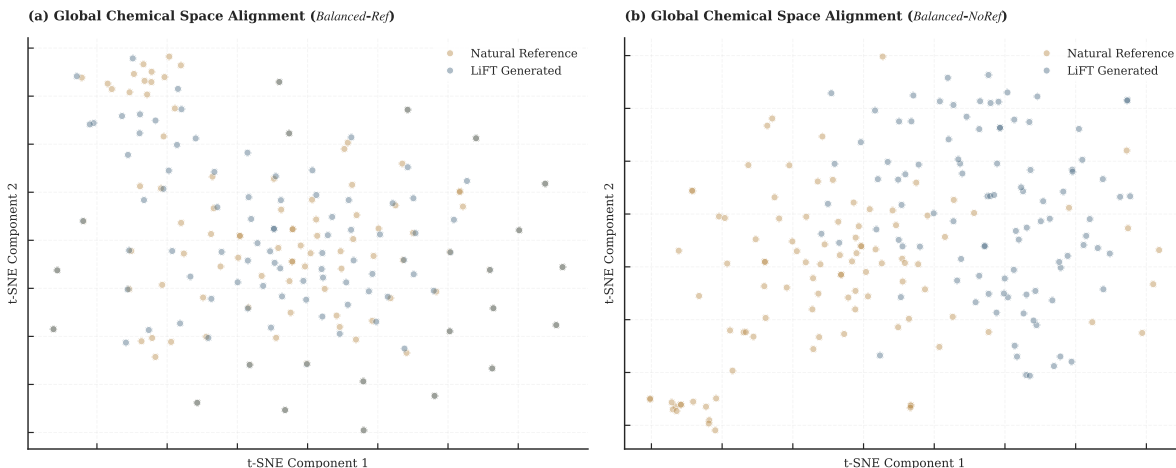


Figure 5: **Qualitative t -SNE visualization of topo-chemical distributions for agent-generated SMILES.** The plots compare ECFP4 fingerprint embeddings of SMILES produced by the language agent with the 100 ground-truth ligand SMILES from the test pockets. (a) Under the reference-guided *LiFT (Balanced Ref)* setting, the agent-generated SMILES show broad overlap with the ground-truth ligand distribution in the projected fingerprint space. (b) Under the reference-free *LiFT (Balanced No-Ref)* setting, the agent-generated SMILES exhibit a more visible distributional shift while remaining near the empirical SMILES distribution. Because t -SNE is a qualitative embedding method, this figure is intended to show distributional tendencies rather than precise chemical-space distances.

which has been described in Figure 3, Section 3.4, and Appendix F. Instead, we use this figure to examine how the learned gates vary across training, ODE inference, and latent-channel distributions. We focus on three outputs: the State Calibration Gate α_{state}^s , the Scalar Update Gate g_{up}^s , and the Vector Update Gate g_{up}^v .

Macro-level training evolution. The left column of Figure 6 shows the aggregated gate values across training epochs. The State Calibration Gate α_{state}^s stays within a relatively high range during training, with moderate fluctuations across pockets. This suggests that the state-calibration branch remains active rather than being ignored by the model. The Scalar Update Gate g_{up}^s rises during the early stage and then stays in a more stable range, indicating that scalar-state updates are learned early and remain consistently used afterward. In contrast, the Vector Update Gate g_{up}^v shows a gradual downward trend after the early phase. This pattern is consistent with a more conservative use of vector-state updates during later training, although we do not interpret it as direct proof of a specific physical mechanism.

Meso-level ODE trajectory dynamics. The middle column profiles the same gates along the ODE integration trajectory. The State Calibration Gate α_{state}^s increases during the middle stage of inte-

gration and then decreases toward the final steps. This indicates that state calibration is most active in the intermediate part of generation, when the ligand and state has already moved away from the initial noise but has not yet reached the final configuration. The Scalar Update Gate g_{up}^s generally decreases along the trajectory, suggesting that scalar updates are stronger at earlier stages and gradually relaxed as integration proceeds. By comparison, the Vector Update Gate g_{up}^v increases in the middle stage before declining near the end. This shows that scalar and vector updates follow different temporal patterns during inference. Across the three target pockets, the curves differ most visibly in the middle stage but become closer near the terminal steps, suggesting pocket-dependent intermediate behavior with relatively similar final regimes.

Micro-level latent-channel spectra. The right column shows the distribution of gate weights across latent channels for different target pockets. The distributions are not uniform or single-peaked. For α_{state}^s , the violin plots show separated high- and mid-value regions, indicating that state calibration is unevenly distributed across channels. The Scalar Update Gate g_{up}^s and Vector Update Gate g_{up}^v also show broad and partly multi-modal distributions, with visible differences among pockets. These observations suggest that SCDR does not apply the same update strength to all latent channels.

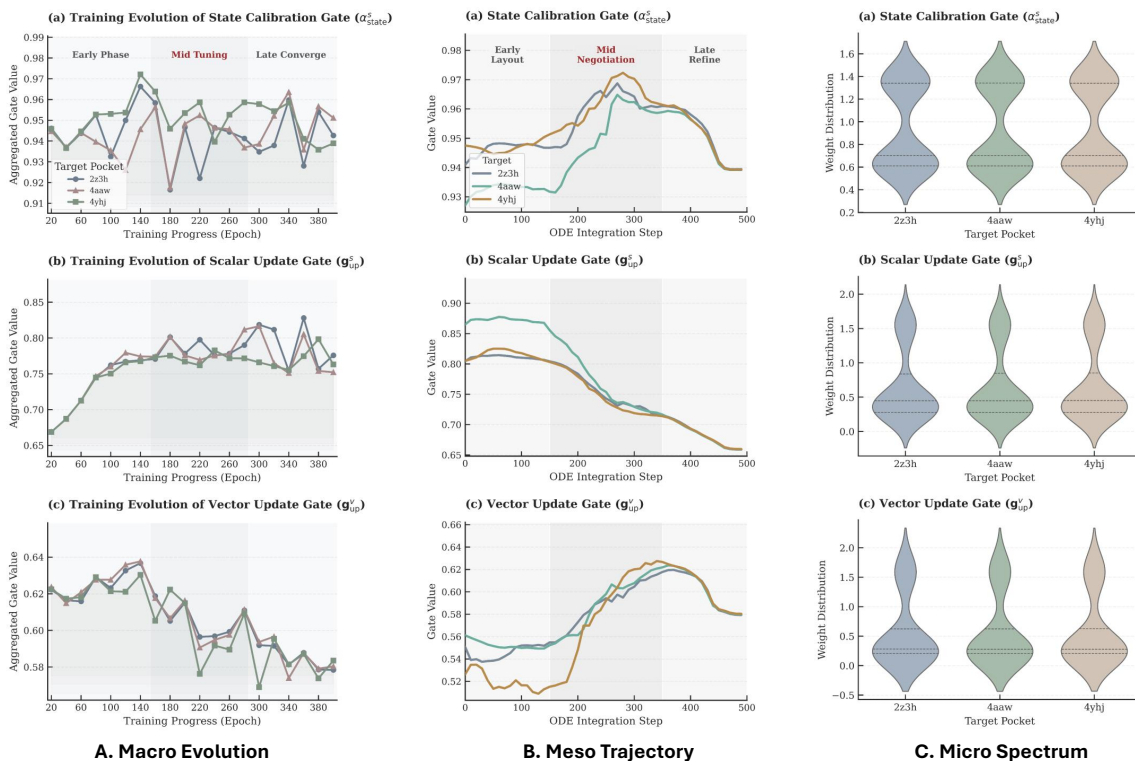


Figure 6: **Holistic spatiotemporal diagnostics of SCDR.** This figure analyzes the learned gate behavior of SCDR rather than its architectural design. **A. Macro evolution:** aggregated gate values across training epochs for the State Calibration Gate α_{state}^s , Scalar Update Gate g_{up}^s , and Vector Update Gate g_{up}^v . **B. Meso trajectory:** gate values along ODE integration steps, showing different temporal patterns for scalar and vector updates and visible pocket-level variation in the middle stage. **C. Micro spectrum:** latent-channel distributions of gate weights across target pockets, showing heterogeneous and pocket-dependent modulation patterns. These diagnostics provide evidence that SCDR learns structured gate dynamics across training, inference, and latent-channel scales.

Instead, different channels receive different levels of modulation, and the resulting distributions vary with the target pocket.

Summary. Overall, Figure 6 shows that SCDR has non-trivial learned gate behavior across three scales. During training, the three gates exhibit different evolution patterns. During ODE inference, scalar and vector updates follow different temporal trends, with stronger pocket-level variation in the middle of the trajectory. At the channel level, the gate distributions show heterogeneous and pocket-dependent patterns rather than uniform collapse. These results should be viewed as diagnostic evidence rather than causal proof. They support a modest interpretation: SCDR learns structured, stage-dependent modulation signals that are compatible with the ablation results in Table 3. Together with the property-invariance analysis in Appendix B.1 and the topo-chemical manifold analysis in Appendix B.2, this provides additional evidence that the language-derived conditions behave as useful

trend-level signals in the 3D generation process.

C Experiment Details

This section describes the dataset construction, conditioning settings, baseline replication protocol, and implementation details used in our experiments. The goal is to make the evaluation protocol explicit and auditable across distribution matching, reference-guided generation, de novo generation, and preference-alignment settings.

C.1 Dataset Preparation and Preprocessing

We use the refined *CrossDocked2020* benchmark for structure-based molecular generation. Each example is derived from a protein–ligand complex and provides two complementary views: a local three-dimensional pocket–ligand geometry for training the generative dynamics, and a one-dimensional molecular representation for constructing semantic conditions. This setting matches the target-aware generation problem considered in LiFT, where ligand generation is conditioned on ex-

plicit pocket geometry while semantic information is provided through a controlled molecular prior.

Complex curation. We follow the standard refined CrossDocked preprocessing protocol and construct examples from the provided pocket-level structures. Complexes with unreliable binding poses, invalid molecular structures, or failed feature construction are excluded. Ligands are restricted to the small-molecule regime relevant to SBDD, ensuring that the benchmark focuses on realistic pocket-conditioned ligand generation rather than out-of-distribution molecular sizes.

Structural condition. For each retained complex, the receptor is represented by the local binding pocket defined by the benchmark preprocessing protocol. This pocket definition is used only to construct the geometric conditioning graph and should not be confused with reference-ligand information exposed to the LLM. In the no-reference setting, the LLM does not receive the ground-truth ligand identity, SMILES, docking scores, or molecular properties. The protein pocket is encoded as a compact residue-level graph with scalar residue features and geometry-aware vector features, while the ligand is represented by coordinates, atom types, and bond types for flow-matching training and sampling.

Semantic condition. For semantic embedding construction, ligand strings are canonicalized with RDKit and encoded by the frozen SMI-TED encoder into a fixed-dimensional molecular semantic vector. This vector provides the one-dimensional chemical prior used by the cross-modal conditioning interface. For controlled comparisons, we construct Ligand-Ref and pure LLM embedding variants under the same processed protein–ligand split, so that the compared settings share identical pocket geometries and differ only in the origin of the semantic condition.

Evaluation. After preprocessing, the training split contains approximately 10^5 refined complexes, and a held-out subset of 100 protein pockets is used for in-distribution evaluation. For each held-out pocket, LiFT generates 100 samples using the same inference configuration across all reported settings. The resulting molecules are evaluated with a unified post-processing and metric-computation pipeline, including cheminformatics validity checks, structural filters, docking-based scores, and distributional property analyses.

C.2 Baseline Specification and Replication Protocol

We compare LiFT with representative structure-based molecular generation methods spanning four major modeling families: autoregressive construction, 3D diffusion, flow matching, and LLM-based SBDD. This benchmark suite covers the dominant design choices in target-aware ligand generation, including sequential graph growth, iterative geometric denoising, continuous-time generative transport, and language-guided molecular proposal.

Autoregressive Models. Autoregressive methods formulate ligand generation as a sequence of local construction decisions conditioned on the protein pocket. AR (Luo et al., 2021) represents early graph-based atom-by-atom generation for SBDD, while Pocket2Mol (Peng et al., 2022) extends this direction with an equivariant pocket-aware representation for predicting atom types, coordinates, and bonds during molecular growth. These methods provide a representative comparison for discrete molecular construction under structural constraints.

3D Diffusion Models. Diffusion-based methods generate ligands by learning to denoise molecular states in three-dimensional space under pocket conditioning. TargetDiff (Guan et al., 2023a) serves as a canonical equivariant diffusion baseline for target-aware generation. DecompDiff (Guan et al., 2023b) incorporates decomposed ligand priors, BindDM (Huang et al., 2024) emphasizes binding-adaptive generation. Together, these baselines represent the current mainstream of 3D geometric generative modeling for SBDD.

Flow-Matching Models. Flow-matching methods learn continuous transport dynamics from simple priors to pocket-conditioned ligand distributions. We include PAFlow (Zhou et al., 2025) and DrugFlow (Schneuing et al., 2025) as representative continuous-time SBDD baselines. This group provides the most direct comparison for evaluating generative models based on learned vector fields and ODE-style sampling.

LLM-based SBDD Pipelines. LLM-based pipelines examine whether molecular language priors can support target-aware ligand design under structure-based evaluation. TamGen (Wu et al., 2024) represents target-conditioned molecular generation with language-modeling components. ELILLM variants, including ELILLM-Diff and ELILLM-Rand (Hu et al., 2026), are included

as recent LLM-driven SBDD baselines. This group evaluates how language-guided molecular proposal pipelines compare with explicitly 3D generative models under the same benchmark.

Replication protocol. We distinguish between full reproduction and unified re-evaluation. For DrugFlow and TamGen, we fully reproduce the baselines using the officially released configuration files. In both cases, we directly follow the provided YAML settings and do not modify model hyperparameters, sampling parameters, or post-processing options. DrugFlow is included as the closest flow-matching baseline to LiFT, while TamGen is re-run because its generated molecule files are not publicly released in a directly usable form.

For the remaining baselines, we use the publicly released generated molecules or official sample files corresponding to the authors’ recommended evaluation setting whenever available, and evaluate them with our unified evaluator. This choice avoids introducing additional implementation variance from re-training or re-sampling multiple heterogeneous models. All molecules, including those generated by LiFT, are processed with the same validity checking, docking/rescoring, PoseBusters auditing, RDKit/REOS filtering, and metric-computation pipeline. Therefore, for these methods, the comparison should be interpreted as a unified re-evaluation of released samples rather than a full reproduction of the original training and sampling procedure.

C.3 Detailed Metric Definitions

To ensure an explicit and auditable evaluation protocol, our metrics strictly reflect the experimental multi-objective tracks (Sec. 4.3), evaluating both absolute generative maximization (Means/Pass Rates) and empirical distribution alignment (Wasserstein Distance).

3D Binding Efficiencies. To rigorously assess target-aware affinity while penalizing physically unrealistic molecular inflation, both primary docking metrics are computed as efficiencies (normalized by the generated ligand’s heavy atom count). *Vina Efficiency* is derived from the raw AutoDock Vina v1.2.5 binding free energy (kcal/mol). Analogously, *Gnina Efficiency* is computed using the GNINA framework, leveraging its 3D CNN-based cross-docking scoring function to evaluate deep empirical pose quality per heavy atom.

1D Pharmacological & Topological Properties. Continuous and discrete molecular prop-

erties are computed deterministically via RDKit. This track evaluates *QED* (Quantitative Estimate of Drug-likeness) and *LogP* (partition coefficient) for pharmacological viability, alongside the *SA* (Synthetic Accessibility) score to assess synthesis feasibility. Geometric complexity and spatial flexibility are systematically quantified via *Rotatable Bond Counts* and *Topological Ring Frequencies* (categorized into > 0 , > 10 , and > 100 ring counts).

Structural Validity & MedChem Filters. Physical realism and medicinal chemistry safety are audited through comprehensive sub-filter suites. *PoseBusters* pass rates strictly verify 3D geometric constraints, covering critical sub-checks including Ring Flatness, Bond Angles, Double Bond stereochemistry, Internal Steric Clashes, and Protein-Ligand Overlap limits. Furthermore, 1D structural alerts are rigorously enforced via *REOS* (rapidly filtering BMS, Glaxo, Inpharmatica, and generic PAINS rule families) and *RDKit Structural Alerts* (comprehensively detecting NIH, PAINS-A/B/C, and ZINC violations) to exclude toxic or chemically reactive substructures.

Property Distribution Divergence. To quantify the macroscopic density shifts between the generated ensembles and the empirical target distribution across all aforementioned continuous metrics (*Vina*, *Gnina*, *QED*, *SA*, *LogP*, *Rotatable Bonds*, and *Ring Frequencies*), we explicitly employ the *Wasserstein Distance* (WD) computed via SciPy as our sole divergence metric. We deliberately avoid Kullback-Leibler (KL) divergence, Jensen-Shannon (JS) divergence, and Fréchet ChemNet Distance (FCD). This design choice is mathematically imperative for continuous-time flow matching: because empirical properties generated during unconstrained multi-objective optimization frequently exhibit disjoint supports relative to the natural distribution, KL divergence inherently suffers from numerical instability (∞), while JS divergence saturates, losing directional gradient sensitivity. Conversely, the Wasserstein metric (Earth Mover’s Distance) preserves a continuous, informative cost gradient even under non-overlapping supports, perfectly corresponding with our generative velocity field trajectory alignment.

C.4 Implementation and Hyperparameters

LiFT is implemented in *PyTorch* with PyTorch Lightning. We use a frozen SMI-TED encoder to provide the ligand-level semantic condition, and train a heterogeneous GVP-GNN as the

Table 7: Core implementation settings of LiFT, including the GVP-based geometric backbone, SMI-TED semantic conditioning, flow-matching training objective, and Euler-based sampling protocol.

Component	Hyperparameter	Value
Backbone	Geometric network	Heterogeneous GVP-GNN
	GVP layers	5
	Node / edge hidden dim.	(128, 32) / (128, 32)
	Self-conditioning	Enabled
Semantic control	Semantic encoder	Frozen SMI-TED
	SMILES embedding dim.	768
	Controller	SCDR/AdaLN
	Controller condition	SMILES embedding + flow time
Training	Objective	Flow matching over coords., atoms, and bonds
	Optimizer	AdamW with AMSGrad
	Learning rate	8×10^{-4}
	Batch size	48
	Training epochs	500
Sampling	ODE solver	Explicit Euler
	Training flow steps	5000
	Inference steps (NFE)	500
	Atom / bond priors	Marginal / uniform

pocket-conditioned geometric backbone. The SCDR/AdaLN controller incorporates the SMILES embedding and flow-time information into ligand-node updates. The model is optimized with a flow-matching objective over ligand coordinates, atom types, and bond types. During inference, samples are generated by integrating the learned vector field with an explicit Euler solver. We run all experiments on a single NVIDIA A800-SXM4-80GB GPU. On this hardware, one training epoch takes less than 10 minutes. For the 100-pocket evaluation setting with 100 generated ligands per pocket (10,000 molecules in total), LiFT 3D flow sampling takes 2 h 38 min 48 s; including language-condition generation and SMI-TED encoding, the complete measured pipeline takes 2 h 57 min 20 s. Docking and offline metric computation are excluded because they are evaluation procedures shared across methods rather than part of generation. The core implementation details and hyperparameter settings are summarized in Table 7.

Runtime and efficiency. To make the inference cost explicit, Table 8 reports component-wise wall-clock measurements under the same A800 setup. Language-condition generation produces three valid SMILES conditions per pocket, yielding 300 conditions over 100 pockets. The 3D stage then generates 100 ligands per pocket. SMI-TED timing is reported both with and without its one-time loading cost.

The 3D sampling stage accounts for approximately 89.5% of the measured LiFT runtime. Under the matched sampling setting, LiFT requires 2 h 38 min 48 s versus 1 h 59 min 20 s for DrugFlow,

corresponding to a 33.1% sampling-time overhead. This comparison isolates sampling cost rather than end-to-end system cost, since DrugFlow does not include corresponding language-generation or semantic-encoding stages.

C.5 Pocket-Level Uncertainty and Statistical Reporting

To quantify variability across targets, we report pocket-level sample standard deviation (SD), across-pocket variance, and pocket-bootstrap 95% confidence intervals (CIs). For each metric, we first aggregate generated molecules within each pocket and then compute SD and variance across test pockets. For CIs, we resample complete pockets with replacement, retain all molecules associated with each selected pocket, and recompute the original aggregate estimator. Using pockets rather than individual molecules as the resampling unit avoids treating molecules generated for the same target as statistically independent.

These quantities characterize across-pocket heterogeneity and sampling uncertainty; they should not be interpreted as retraining-seed variance. Because the reported results use fixed trained checkpoints, multi-seed retraining variance is a distinct source of uncertainty. Table 9 reports SD, variance, and 95% CIs for the continuous metrics most central to the main comparison. SD is reported in the original metric units, whereas variance is in squared units.

For the no-reference property-oriented variants, we additionally report uncertainty for representative filter and pose-validity metrics in Table 10. For percentage metrics, SD is measured in percentage points and variance in squared percentage points.

The uncertainty estimates support a cautious interpretation of small numerical differences. In particular, the QED-NoRef variant has a QED 95% CI of [0.723, 0.763] and an SA 95% CI of [2.602, 2.851], while confidence intervals for several docking and pose-validity comparisons overlap with strong 3D baselines. We therefore use these statistics to characterize stability and across-pocket heterogeneity rather than to claim significance from small point-estimate gaps.

C.6 Absolute Docking Scores

For direct comparability with studies reporting unnormalized docking scores, we additionally report raw Vina and Gnina scores under the same evaluation and failure-handling protocol. These absolute

Table 8: Measured runtime of the LiFT pipeline on a single NVIDIA A800-SXM4-80GB GPU. The matched DrugFlow comparison uses the same 100-pocket, 10,000-molecule sampling workload.

Component / Method	Timed Batch	Total Wall-Clock Time	Normalized Runtime
LLM condition generation	100 pockets; 300 conditions	14 min 48 s	8.88 s/pocket; 2.96 s/condition
SMI-TED encoding, incl. loading	300 conditions	224.08 s	0.747 s/condition
SMI-TED core encoding	300 conditions	208.17 s	0.694 s/condition
One-time SMI-TED loading	One initialization	14.87 s	Not amortized across future batches
LiFT 3D flow sampling	100 pockets $\times$ 100 ligands	2 h 38 min 48 s	0.953 s/molecule; 95.28 s/pocket
DrugFlow 3D flow sampling	Matched 10,000-molecule workload	1 h 59 min 20 s	0.716 s/molecule; 71.60 s/pocket
Complete measured LiFT pipeline	LLM + SMI-TED + 3D sampling	2 h 57 min 20 s	106.40 s/pocket

Table 9: Pocket-level uncertainty for continuous evaluation metrics. Each entry is **SD / variance / 95 % bootstrap CI**. Confidence intervals are computed by resampling complete test pockets.

Method	Gnina Eff.	Vina Eff.	QED	SA
AR	0.061/0.0038/[0.283,0.307]	0.083/0.0068/[-0.426,-0.393]	0.119/0.0140/[0.485,0.533]	0.895/0.8019/[4.109,4.470]
PAFlow	0.040/0.0016/[0.254,0.270]	0.092/0.0084/[-0.445,-0.414]	0.101/0.0102/[0.471,0.511]	0.830/0.6882/[4.720,5.058]
BindDM	0.044/0.0019/[0.240,0.257]	0.074/0.0055/[-0.376,-0.347]	0.100/0.0101/[0.489,0.527]	0.654/0.4281/[4.661,4.902]
DecompDiff	0.037/0.0014/[0.197,0.213]	0.069/0.0048/[-0.281,-0.252]	0.137/0.0187/[0.425,0.483]	0.551/0.3041/[4.407,4.649]
TargetDiff	0.043/0.0018/[0.236,0.253]	0.069/0.0047/[-0.357,-0.330]	0.111/0.0124/[0.458,0.500]	0.593/0.3516/[4.624,4.850]
DrugFlow	0.043/0.0018/[0.243,0.260]	0.081/0.0065/[-0.353,-0.322]	0.135/0.0184/[0.526,0.578]	0.701/0.4915/[3.295,3.570]
Pocket2Mol	0.063/0.0039/[0.285,0.310]	0.076/0.0058/[-0.440,-0.410]	0.091/0.0083/[0.555,0.591]	0.631/0.3976/[3.076,3.323]
TamGen	0.008/0.0001/[0.087,0.090]	0.013/0.0002/[-0.006,-0.001]	0.078/0.0060/[0.445,0.476]	0.387/0.1499/[7.450,7.602]
ELILLM-Diff	0.020/0.0004/[0.081,0.089]	0.010/0.0001/[-0.004,-0.000]	0.105/0.0111/[0.466,0.507]	0.660/0.4361/[8.434,8.694]
ELILLM-Rand	0.023/0.0005/[0.087,0.096]	0.012/0.0001/[-0.005,-0.001]	0.145/0.0210/[0.431,0.488]	0.624/0.3893/[7.988,8.234]
LiFT (Lig.-Ref)	0.075/0.0056/[0.241,0.270]	0.098/0.0096/[-0.365,-0.326]	0.181/0.0327/[0.444,0.516]	1.133/1.2843/[3.435,3.892]
LiFT (QED-Ref)	0.058/0.0034/[0.231,0.253]	0.092/0.0086/[-0.347,-0.311]	0.185/0.0342/[0.483,0.558]	1.097/1.2029/[3.464,3.922]
LiFT (Vina-Ref)	0.064/0.0040/[0.231,0.255]	0.091/0.0083/[-0.347,-0.311]	0.182/0.0331/[0.479,0.552]	1.075/1.1560/[3.537,3.981]
LiFT (Bal-Ref)	0.063/0.0040/[0.235,0.260]	0.096/0.0091/[-0.354,-0.317]	0.180/0.0324/[0.496,0.569]	1.086/1.1794/[3.416,3.876]
LiFT (QED-NoRef)	0.038/0.0014/[0.236,0.250]	0.082/0.0068/[-0.357,-0.325]	0.094/0.0089/[0.723,0.763]	0.662/0.4380/[2.602,2.851]
LiFT (Vina-NoRef)	0.043/0.0018/[0.248,0.264]	0.092/0.0085/[-0.363,-0.327]	0.099/0.0098/[0.708,0.750]	0.593/0.3511/[2.538,2.794]
LiFT (Bal-NoRef)	0.037/0.0014/[0.234,0.248]	0.101/0.0103/[-0.350,-0.311]	0.086/0.0074/[0.740,0.774]	0.574/0.3291/[2.552,2.774]

Table 10: Pocket-level uncertainty for representative filter and pose-validity metrics. Each entry is **SD / variance / 95 % bootstrap CI**.

Variant	RDKit (%)	REOS (%)	PoseB. (%)
QED-NoRef	10.92/119.2/[81.11,85.12]	16.94/287.1/[67.92,74.12]	21.52/462.9/[68.20,76.00]
Vina-NoRef	11.83/140.0/[80.83,85.74]	18.14/329.2/[70.17,77.25]	21.43/459.4/[69.40,77.30]
Bal-NoRef	13.76/189.3/[78.35,83.99]	19.52/381.1/[67.88,75.58]	21.53/463.5/[66.50,74.60]

scores complement the heavy-atom-normalized docking-efficiency metrics and the Wasserstein-distance analysis reported in the main paper.

All methods use the same docking and failure-handling protocol. Raw values near zero for some sequence-based baselines reflect unsuccessful or unscored post-hoc 3D poses under this unified pipeline rather than a different scoring function. Taken together, the reported docking results provide three complementary views: absolute Vina/Gnina scores for conventional comparability, heavy-atom-normalized efficiency for size-aware affinity evaluation, and Wasserstein Distance for distributional fidelity.

D Controlled SMILES-Condition Protocol

The main text uses the *Sense-Evolve-Assemble* agent to convert pocket descriptions, task directives, and optional ligand references into one-dimensional SMILES conditions for the downstream flow model. This appendix formalizes the operational interface between the language-based SMILES proposal stage and the geometric generator. It specifies the information available to the language model, the deterministic parsing and validation procedure, and the construction of the fixed SMILES-derived semantic latent used during flow sampling.

D.1 Semantic Intervention and Information Boundary

LiFT conditions the geometric generator through a semantic vector $\mathbf{z}_{sem}$ produced by a frozen SMI-TED encoder. The controlled variable is the source of the SMILES string used to instantiate this vector. When the semantic condition is instantiated from

Table 11: Absolute docking scores under the unified evaluation protocol. Lower raw Vina is better, while higher raw Gnina is better. Values are mean $\pm$ standard deviation.

Method	Raw Vina $\downarrow$	Raw Gnina $\uparrow$
AR	-6.756 ± 2.462	4.741 ± 1.391
PAFlow	-9.647 ± 3.503	5.765 ± 1.384
BindDM	-8.236 ± 3.066	5.593 ± 1.403
DecompDiff	-7.647 ± 2.283	5.849 ± 1.072
TargetDiff	-7.770 ± 2.761	5.475 ± 1.388
DrugFlow	-6.938 ± 2.908	5.118 ± 1.302
Pocket2Mol	-6.984 ± 2.907	4.687 ± 1.441
TamGen	-0.143 ± 1.840	3.880 ± 0.506
ELILLM-Diff	-0.071 ± 3.446	4.420 ± 1.182
ELILLM-Rand	-0.133 ± 2.608	4.275 ± 1.311
LiFT (Ligand-Ref Embedding)	-6.825 ± 2.555	4.902 ± 1.329
LiFT (QED-Reference)	-6.804 ± 2.477	4.910 ± 1.289
LiFT (Vina-Reference)	-6.782 ± 2.447	4.906 ± 1.310
LiFT (Balanced-Reference)	-6.797 ± 2.542	4.902 ± 1.305
LiFT (QED-No-Reference)	-6.748 ± 2.160	4.785 ± 1.057
LiFT (Vina-No-Reference)	-6.362 ± 2.177	4.645 ± 1.009
LiFT (Balanced-No-Reference)	-6.576 ± 2.601	4.791 ± 1.047

an LLM-generated SMILES string, the string is produced by the fixed protocol described below. In the *Ligand-Ref Embedding* condition, the available reference ligand SMILES is encoded directly as a reference embedding. This condition is used as a diagnostic, reference-guided setting to assess how reference semantic information affects the downstream generator. It is therefore reported separately from reference-free *de novo* variants.

The access boundary is fixed before any target is evaluated. In the reference-free *de novo* setting, the LLM receives only the pocket-derived profile and a task directive. It does not receive the test ligand, reference SMILES, hidden labels, docking scores, filter outcomes, or generated 3D structures. In the reference-guided setting, the LLM additionally receives the reference ligand SMILES and coarse reference-derived size constraints. We therefore report this setting separately as a *Ligand-Ref* variant. Across both settings, downstream metrics are used only for evaluation, not for selecting, revising, or ranking the SMILES condition. The corresponding prompt skeletons make this input separation explicit: Table 12 lists the reference-free fields, while Table 13 shows the additional reference-guided fields.

D.2 Operational Protocol

The protocol follows the three stages described in the methodology. First, the *Sense* stage summarizes the target pocket into a compact textual profile, including coarse geometry and physico-

chemical anchors from the local interaction shell. This profile is the only target-specific input in the reference-free setting. Second, the *Evolve* stage applies a fixed Pocket-of-Thought (PoT) schema. The schema asks the LLM to state a compact spatial and topological plan before emitting SMILES, making the intended SMILES condition inspectable without using the intermediate plan as a reward, filter, or selection signal. Third, the *Assemble* stage extracts tagged SMILES strings, applies deterministic cheminformatics verification, and passes the first valid, SMI-TED-encodable molecule to the semantic encoder.

Prompt variants are fixed at the script level. Distribution-oriented, balanced, Vina-oriented, and QED/safety-oriented directives modify only the task text supplied to the LLM. They do not change the generator weights, invoke an additional guidance model, or introduce feedback from downstream evaluation. The same parsing, validation, and semantic extraction rule is used for all prompt families and LLM backbones considered in the experiments.

D.3 Validation and Candidate Accounting

Candidate acceptance is based only on molecular string validity and successful SMI-TED encoding. RDKit is used as an unweighted validity check rather than as a ranking surrogate. If a generated string fails to parse, the repair routine is restricted to local SMILES syntax correction using the parser diagnostic. It does not receive docking scores, property scores, reference-similarity scores, filter results, or sampled 3D coordinates. Thus, the repair step provides syntax-level recovery only and does not introduce score-feedback optimization over evaluation metrics. The repair interface is specified in Table 14, which exposes only the parser diagnostic, invalid SMILES string, and captured PoT context.

When multiple valid candidates are available, their chronological order is preserved and the first SMI-TED-encodable candidate instantiates $\mathbf{z}_{sem}$. If no valid encodable candidate is produced under the fixed protocol, the target is logged as a semantic-instantiation failure rather than being replaced through post-hoc search. This accounting rule prevents downstream scores from influencing which semantic condition is passed to the 3D generator.

D.4 Semantic Feature Extraction and Claim Boundary

After validation, the selected SMILES string is embedded by the frozen SMI-TED encoder:

$$\mathbf{z}_{sem} = \text{Encoder}_{\text{SMI-TED}}(S_{smiles}).$$

The resulting vector is constructed before any 3D geometric generation and is held fixed throughout flow sampling. The downstream flow model does not call the LLM, rerun repair, inspect candidate rationales, or revise $\mathbf{z}_{sem}$ using generated coordinates or evaluation metrics.

This protocol supports a deliberately narrow operational claim: LiFT uses language to instantiate a fixed SMILES-derived semantic condition under predefined information boundaries. It does not claim that PoT rationales are faithful explanations, that the LLM independently performs molecular optimization, or that *Ligand-Ref Embedding* represents a reference-free deployment condition. Instead, the protocol separates SMILES-condition construction from 3D generation and downstream metric evaluation.

D.5 Prompt Skeletons and Structural Schema

For reproducibility, the accompanying prompt tables define the input fields, required output tags, parsing schema, reference-guided additions, and repair interface used by the implementation. Runtime variable interpolations are shown in square brackets, e.g., [Target ID]. These templates are fixed before evaluation and are not edited per target after observing generated molecules or benchmarking scores. Specifically, Table 12 defines the reference-free *de novo* template, Table 13 defines the *Ligand-Ref* scaffold-hopping template, and Table 14 defines the syntax-repair interface.

D.6 Task-Specific Property Optimization Directives

To evaluate controllable property steering, we construct task-specific SMILES-derived semantic conditions by varying only the directive given to the LLM. All other components are kept unchanged: the information boundary, deterministic SMILES parsing, validation procedure, candidate-accounting rule, SMI-TED semantic extraction, and the downstream 3D generator. As a result, the balanced, Vina-oriented, and QED/safety-oriented settings isolate the effect of semantic task directives under the same retraining-free generation in-

terface, rather than introducing separate optimization pipelines. The concrete directive blocks are reported in Table 15, Table 16, and Table 17.

E Rule-Level Structural, MedChem, and Pose Validity Analysis

Table 2 reports the aggregate RDKit, REOS, and PoseBusters pass rates together with the main optimization metrics. These aggregate scores summarize whether generated molecules satisfy broad structural-alert, medicinal-chemistry, and 3D pose-validity checks, but they do not reveal which rule families or geometric constraints drive the result. In this appendix, we therefore unpack the aggregate filter outcomes into rule-level diagnostics: RDKit and REOS are decomposed into chemical alert families, while PoseBusters is decomposed into 3D validity sub-checks.

E.1 Fine-grained RDKit and REOS Checks

Tables 18 and 19 provide the chemical rule-level audit. Table 18 reports the RDKit structural-alert sub-filters, including NIH, PAINS, the PAINS-A/B/C subclasses, and ZINC. Table 19 reports the REOS medicinal-chemistry rule families, including BMS, Glaxo, Inpharmatica, and REOS-PAINS. All entries are pass rates, so higher values indicate fewer violations of the corresponding rule. The two tables explain the aggregate RDKit/REOS compliance results in Table 2, rather than replacing the main comparison.

The two chemical-filter tables show that different model families have different localized alert profiles. Several baselines are strongest on individual NIH or PAINS columns, while LiFT no-reference variants are strong on ZINC and remain competitive on Glaxo and Inpharmatica while maintaining high PAINS-family pass rates. Together with the aggregate RDKit/REOS scores in Table 2, these sub-filters support a cautious interpretation: LiFT’s favorable chemical-filter behavior is not tied to a single rule family, but is reflected across multiple structural-alert and medicinal-chemistry checks.

E.2 Fine-grained PoseBusters Checks

RDKit and REOS operate mainly on molecular graph and substructure rules. PoseBusters provides a complementary 3D validity check by testing whether the predicted ligand pose is geometrically plausible and physically compatible with the binding pocket. This matters for SBDD evaluation because a model can otherwise appear competitive

[HTML]F3F3F3 Box 1: Blind <i>De Novo</i> Generation Template (No-Reference Setting) [System Role] You are a premier De Novo Drug Designer. Your task is to generate novel lead-like molecules from scratch that perfectly complement the provided 3D protein pocket environment. ### GLOBAL DATASET BASELINE (Context Only) To calibrate your chemical intuition, active ligands in our training set have the following average properties: QED $\approx$ 0.53, SA Score $\approx$ 3.2, LogP $\approx$ 1.76, Heavy Atoms $\approx$ 24, Rotatable Bonds $\approx$ 4.6, and Vina Score $\approx$ -8.3. Treat these values strictly as a preliminary global reference. You must flexibly adapt or even drastically deviate from these averages depending on the specific geometric shape, size, and chemical environment of the target pocket. ### STEP 1: POCKET-OF-THOUGHT (PoT) REASONING [CRITICAL] Before generating SMILES, you MUST provide a brief 'Reasoning' section (max 4 sentences). 1. Explicitly explain your spatial reasoning strategy based on the 8.0Å anchors. 2. INTERACTION MAPPING: You MUST specify which part of your proposed molecule (e.g., 'the secondary amine', 'the phenyl ring') corresponds to each anchor residue (e.g., 'to align with Asp102'). 3. Explain how the molecular topology (elongated, branched, or compact) fits the pocket's geometric shape. ### STEP 2: DE NOVO GENERATION REQUIREMENTS 1. Scaffold Innovation: Generate entirely novel scaffolds from scratch. Prioritize chemical diversity and high 3D complementarity. 2. Guided Alignment (Soft Constraints): Treat the provided Macro 3D constraints as highly informed, important references to guide your design. [Insert Task-Specific Optimization Directive Here]. 3. Chemical Rigor: Ensure perfect SMILES syntax, proper valency, and strictly avoid problematic motifs (e.g., PAINS or highly reactive groups). ### OUTPUT FORMAT: Your response must follow this EXACT structure: [Reasoning] Your spatial and property reasoning analysis here... <smiles>SMILES_1</smiles> <smiles>SMILES_2</smiles> <smiles>SMILES_3</smiles> [User Input] Target ID: [Target ID] 3D Local Environment Analysis (8.0Å Interaction Shell) - Description: [Pocket Profile Description] CRITICAL INSTRUCTION: You are performing a blind de novo drug design. Your PoT should logically deduce a suitable novel chemical scaffold from scratch that strictly complements the pocket's 3D environment. Now, perform Step 1 (PoT Reasoning) and Step 2 (Generation) to provide 3 de novo candidates.

Table 12: System prompt template for blind *de novo* ligand generation (no-reference setting).

[HTML]F3F3F3 Box 2: Reference-Guided Optimization Template (Ref Setting)
[System Role] You are a premier Medicinal Chemist specializing in Scaffold Hopping. Your task is to design novel lead-like molecules by replacing the core scaffold of a known active ligand, while preserving its key pharmacophore interactions within the 3D pocket. ### STEP 1: POCKET-OF-THOUGHT (PoT) REASONING [CRITICAL] Before generating SMILES, you MUST provide a brief 'Reasoning' section (max 4 sentences). 1. Explicitly explain your spatial reasoning strategy based on the 8.0Å anchors. 2. INTERACTION MAPPING: You MUST specify which part of your proposed molecule (e.g., 'the secondary amine', 'the phenyl ring') corresponds to each anchor residue (e.g., 'to align with Asp102'). 3. Explain how the molecular topology (elongated, branched, or compact) fits the pocket's geometric shape. ### STEP 2: LEAD OPTIMIZATION REQUIREMENTS 1. Scaffold Enhancement: Retain the core structural backbone of the reference ligand. Focus your innovation on decorating the core, optimizing side chains, or making bioisosteric replacements. Do not completely delete the core scaffold unless necessary to resolve a severe spatial clash. [Insert Task-Specific Optimization Directive Here]. 2. Physical Alignment: Ensure the new molecule stays within the target molecular weight and atom count ranges derived from the reference. 3. Chemical Rigor: Ensure perfect SMILES syntax, proper valency, and strictly avoid PAINS motifs or reactive functional groups. ### OUTPUT FORMAT: Your response must follow this EXACT structure: [Reasoning] Your spatial and property reasoning analysis here... <smiles>SMILES_1</smiles> <smiles>SMILES_2</smiles> <smiles>SMILES_3</smiles>
[User Input] Target ID: [Target ID] 3D Local Environment Analysis (8.0Å Interaction Shell) - Description: [Pocket Profile Description] Physical Boundary Constraints - Target Molecular Weight: [Weight Range] - Heavy Atom Count: [Atom Count Range] Reference Ligand Pattern (Ground Truth) SMILES: [Reference SMILES] CRITICAL INSTRUCTION: The Reference Ligand is known to bind effectively; your PoT should explain how to enhance its existing interactions rather than fundamentally deleting its core scaffolds unless there is a severe spatial clash. Now, perform Step 1 (PoT Reasoning) and Step 2 (Generation) to provide 3 optimized candidates.

Table 13: System prompt template for reference-guided lead optimization and scaffold hopping.

[HTML]F3F3F3 Box 3: Knowledge-Guided Decoding (KGD) Repair Prompt [System Role] You are the 'KGD Repair Engine'. You fix SMILES based on specific parser error messages. ### DIAGNOSTIC FEEDBACK: The chemical parser reported this error: [RDKit Error Message] ### REPAIR RULES: 1. If 'unclosed ring' is reported, check your ring closure numbers (e.g., '1', '2') and ensure they all have pairs. 2. If 'valence' is reported, fix the atom's bond count or add appropriate charges. 3. Maintain the pharmacophore described in the reasoning. OUTPUT: Provide ONLY the repaired SMILES in <smiles></smiles> tags. [User Input] Target Reasoning: [Captured PoT Context] Broken SMILES: [Invalid SMILES String] Please fix the error indicated in the diagnostic feedback.
--

Table 14: System prompt template for Knowledge-Guided Decoding (KGD) chemical syntax repair.

[HTML]F3F3F3 Box 4: Property Optimization Template — Balanced Mode (No-Reference Setting) [System Role] You are a premier De Novo Drug Designer. Your task is to generate novel lead-like molecules from scratch that perfectly complement the provided 3D protein pocket environment. ### STEP 1: POCKET-OF-THOUGHT (PoT) REASONING [CRITICAL] Before generating SMILES, you MUST provide a brief 'Reasoning' section (max 4 sentences). 1. Explicitly explain your spatial reasoning strategy based on the 8.0Å anchors. 2. INTERACTION MAPPING: You MUST specify which part of your proposed molecule (e.g., 'the secondary amine', 'the phenyl ring') corresponds to each anchor residue (e.g., 'to align with Asp102'). 3. Explain how the molecular topology (elongated, branched, or compact) fits the pocket's geometric shape. ### STEP 2: MOLECULAR GENERATION REQUIREMENTS (BALANCED OPTIMIZATION) 1. De Novo Innovation: Generate entirely novel scaffolds from scratch. Do not rely on any known reference inhibitors. 2. Balanced Maximization: Explicitly optimize BOTH drug-likeness (QED > 0.6, SA < 3.0) AND theoretical binding affinity (e.g., highly negative Vina/Gnina scores). CRITICALLY: You must also optimize Ligand Efficiency (binding energy per heavy atom). To achieve this, do NOT artificially inflate the molecular size just to accumulate weak van der Waals forces. Every heavy atom must count. 3. Realism & Safety: Ensure the molecule maintains highly realistic 3D geometry without internal steric clashes or absurdly twisted bond lengths. Strictly avoid Pan-Assay Interference Compounds (PAIS) and reactive/toxic groups. 4. GRAMMAR & SYNTAX: The SMILES must be strictly valid. Ensure all ring closures (numbers) are perfectly paired, do not generate fake aromatic rings (Kekulization), and ensure no atom exceeds its maximum allowed valence. ### OUTPUT FORMAT: Your response must follow this EXACT structure: [Reasoning] Your spatial reasoning analysis here... <smiles>SMILES_1</smiles> <smiles>SMILES_2</smiles> <smiles>SMILES_3</smiles>

Table 15: System prompt protocol for target-driven Balanced Optimization.

[HTML]F3F3F3 Box 5: Property Optimization Template — Medicinal Chemistry / Safety Mode
[System Role] You are a premier De Novo Drug Designer. Your task is to generate novel lead-like molecules from scratch that perfectly complement the provided 3D protein pocket environment. ### STEP 1: POCKET-OF-THOUGHT (PoT) REASONING [CRITICAL] Before generating SMILES, you MUST provide a brief 'Reasoning' section (max 4 sentences). 1. Explicitly explain your spatial reasoning strategy based on the 8.0Å anchors. 2. INTERACTION MAPPING: You MUST specify which part of your proposed molecule (e.g., 'the secondary amine', 'the phenyl ring') corresponds to each anchor residue (e.g., 'to align with Asp102'). 3. Explain how the molecular topology (elongated, branched, or compact) fits the pocket's geometric shape. ### STEP 2: MOLECULAR GENERATION REQUIREMENTS 1. De Novo Innovation: Generate entirely novel scaffolds from scratch. Do not rely on any known reference inhibitors. 2. Medicinal Chemistry Filters: Explicitly optimize for drug-likeness (QED > 0.6) and synthetic accessibility (SA < 3.0). Strictly avoid Pan-Assay Interference Compounds (PAINS) and reactive/toxic groups. 3. GRAMMAR & SYNTAX: The SMILES must be strictly valid. Ensure all ring closures (numbers) are perfectly paired, do not generate fake aromatic rings (Kekulization), and ensure no atom exceeds its maximum allowed valence. ### OUTPUT FORMAT: Your response must follow this EXACT structure: [Reasoning] Your spatial reasoning analysis here... <smiles>SMILES_1</smiles> <smiles>SMILES_2</smiles> <smiles>SMILES_3</smiles>

Table 16: System prompt protocol for targeted Medicinal Chemistry and structural safety screening.

[HTML]F3F3F3 Box 6: Property Optimization Template — Affinity Mode (No-Reference Setting)
[System Role] You are a premier De Novo Drug Designer. Your task is to generate novel lead-like molecules from scratch that perfectly complement the provided 3D protein pocket environment. ### STEP 1: POCKET-OF-THOUGHT (PoT) REASONING [CRITICAL] Before generating SMILES, you MUST provide a brief 'Reasoning' section (max 4 sentences). 1. Explicitly explain your spatial reasoning strategy based on the 8.0Å anchors. 2. INTERACTION MAPPING: You MUST specify which part of your proposed molecule (e.g., 'the secondary amine', 'the phenyl ring') corresponds to each anchor residue (e.g., 'to align with Asp102'). 3. Explain how the molecular topology (elongated, branched, or compact) fits the pocket's geometric shape. ### STEP 2: MOLECULAR GENERATION REQUIREMENTS (AFFINITY OPTIMIZATION) 1. De Novo Innovation: Generate entirely novel scaffolds from scratch. Do not rely on any known reference inhibitors. 2. Affinity & Efficiency Maximization: Explicitly optimize the theoretical binding affinity (e.g., highly negative Vina/Gnina scores). CRITICALLY: You must also optimize Ligand Efficiency (binding energy per heavy atom). To achieve this, do NOT artificially inflate the molecular size just to accumulate weak van der Waals forces. Every heavy atom must count. 3. PoseBusters Realism: Ensure the molecule maintains highly realistic 3D geometry without internal steric clashes or absurdly twisted bond lengths. 4. GRAMMAR & SYNTAX: The SMILES must be strictly valid. Ensure perfectly paired ring closures and no hypervalency. ### OUTPUT FORMAT: Your response must follow this EXACT structure: [Reasoning] Your spatial reasoning analysis here... <smiles>SMILES_1</smiles> <smiles>SMILES_2</smiles> <smiles>SMILES_3</smiles>

Table 17: System prompt protocol for targeted Binding Affinity maximization.

Table 18: RDKit structural-alert sub-filter pass rates on CrossDocked2020. All entries are percentages, and higher values indicate fewer violations. Best results among generated methods are highlighted in **bold**, and the second best are underlined.

Category	Method	RDKit Structural Alerts (%) $\uparrow$					
		NIH	PAINS	PAINS-A	PAINS-B	PAINS-C	ZINC
3D-based Models	AR (Luo et al., 2021)	54.80	96.69	<u>92.13</u>	97.55	99.88	95.92
	PAFlow (Zhou et al., 2025)	77.80	98.50	98.83	<u>99.80</u>	99.85	94.65
	BindDM (Huang et al., 2024)	73.82	97.64	98.41	99.44	99.73	93.98
	DecompDiff (Guan et al., 2023b)	64.02	95.73	98.15	97.83	99.64	97.15
	TargetDiff (Guan et al., 2023a)	69.11	97.91	99.07	98.98	99.76	93.11
	DrugFlow (Schneuing et al., 2025)	79.69	96.16	98.22	97.99	<u>99.86</u>	96.58
	Pocket2Mol (Peng et al., 2022)	88.95	95.05	98.64	96.71	99.53	97.40
	TamGen (Wu et al., 2024)	<u>89.72</u>	95.70	99.28	97.85	98.52	38.93
LLM-based Models	ELILLM-Diff (Hu et al., 2026)	89.98	<u>98.15</u>	98.52	99.87	99.73	25.99
	ELILLM-Rand (Hu et al., 2026)	87.81	96.70	98.68	98.22	99.72	32.47
	Ligand-Ref Embedding	63.21	95.32	99.05	96.42	99.83	94.29
LiFT Variants	LiFT (Bal-Reference)	67.74	95.33	98.97	96.48	99.76	96.00
	LiFT (QED-Reference)	64.75	94.34	98.24	96.21	99.68	95.40
	LiFT (Vina-Reference)	64.84	95.79	98.54	97.34	99.85	95.76
	LiFT (Vina-No-Reference)	86.08	96.69	97.76	99.17	99.65	<u>92.48</u>
	LiFT (Vina-No-Reference)	86.86	95.85	98.09	97.89	99.77	99.21
	LiFT (Bal-No-Reference)	83.55	97.05	98.13	99.11	99.78	99.62
	LiFT (Bal-No-Reference)	83.55	97.05	98.13	99.11	99.78	99.62

Table 19: REOS medicinal-chemistry rule-family pass rates on CrossDocked2020. All entries are percentages, and higher values indicate fewer violations. Best results among generated methods are highlighted in **bold**, and the second best are underlined.

Category	Method	REOS Rule Families (%) $\uparrow$			
		BMS	Glaxo	Inpharmatica	PAINS
3D-based Models	AR (Luo et al., 2021)	54.80	61.94	54.22	97.03
	PAFlow (Zhou et al., 2025)	77.80	87.96	70.25	98.62
	BindDM (Huang et al., 2024)	73.82	93.37	67.62	98.00
	DecompDiff (Guan et al., 2023b)	64.02	80.91	65.40	96.19
	TargetDiff (Guan et al., 2023a)	69.11	90.93	69.86	<u>98.19</u>
	DrugFlow (Schneuing et al., 2025)	79.69	89.75	79.34	96.61
	Pocket2Mol (Peng et al., 2022)	88.95	96.27	72.91	94.98
	TamGen (Wu et al., 2024)	<u>89.72</u>	94.07	62.51	95.65
LLM-based Models	ELILLM-Diff (Hu et al., 2026)	89.98	89.47	82.35	98.14
	ELILLM-Rand (Hu et al., 2026)	87.81	83.96	79.04	96.43
	Ligand-Ref Embedding	63.21	79.97	70.37	95.78
LiFT Variants	LiFT (Bal-Reference)	67.74	84.49	71.26	95.63
	LiFT (QED-Reference)	64.75	83.36	71.16	94.67
	LiFT (Vina-Reference)	64.84	84.91	70.47	96.20
	LiFT (QED-No-Reference)	86.08	<u>95.70</u>	77.88	97.51
	LiFT (Vina-No-Reference)	86.86	95.16	<u>80.67</u>	97.21
	LiFT (Bal-No-Reference)	83.55	95.51	79.57	98.08
	LiFT (Bal-No-Reference)	83.55	95.51	79.57	98.08

on docking-oriented metrics while producing poses with broken connectivity, strained bond geometry, internal clashes, or unrealistic protein-ligand contacts.

Table 20 reports PoseBusters sub-checks that complement the overall PoseBusters score in Table 2. These columns separate local ligand geometry from protein-ligand pose compatibility: aromatic-ring flatness, bond-angle geometry, double-bond flatness, internal steric clashes, protein-ligand maximum distance, and protein-volume overlap. As above, all values are pass rates in percentages, so higher values indicate fewer violations.

The PoseBusters sub-checks add a complementary view to the aggregate pose-validity score in Table 2. Language-only baselines can satisfy several local geometry checks while still having very low overall PoseBusters pass rates, indicat-

ing that isolated sub-checks do not guarantee a valid 3D pose. In contrast, LiFT variants retain stable protein-overlap behavior and competitive geometry-related sub-checks while also maintaining the RDKit/REOS compliance patterns above. This supports that LiFT’s aggregate gains remain consistent with both chemical rule filters and 3D pose-validity checks.

F Qualitative Case Study on Representative Targets

To visually demonstrate the binding modes and geometric complementarity of the generated molecules, we select four representative protein targets (1E8H, 1JN2, 2AZY, and 3AF2) for qualitative analysis. These targets encompass a diverse range of challenging spatial environments, varying from deep, narrow binding pockets to wide, flat surface grooves. To ensure a fair and rigorous comparison, we establish baselines representing state-of-the-art autoregressive (AR), diffusion (DecompDiff), and flow-matching (PAFlow) paradigms, alongside the Ground Truth (GT) standard molecules from the test set. These specific targets were deliberately chosen due to their challenging nature, distinct physicochemical properties, and the high performance variance observed in their corresponding GT molecules. For each target, we generated a library of dozens of candidate ligands across all generative methods. Subsequently, these candidates were rigorously filtered based on a comprehensive evaluation of AutoDock Vina binding affinity, Quantitative Estimate of Drug-likeness (QED), and Synthetic Accessibility (SA). The selected ligand-protein complexes represent the top-performing candidates that achieve an optimal balance across these critical metrics, allowing for a rigorous visual comparison against the GT.

As illustrated in the 3D visualizations and quantitative annotations in Figure 7, the molecules generated by LiFT exhibit superior spatial-semantic fidelity. Crucially, distinguishing itself from approaches that merely game the Vina scoring function, LiFT achieves an exceptional equilibrium: it not only excels in geometric complementarity but also demonstrates outstanding intrinsic molecular properties (QED and SA). Compared to baseline models—which often suffer from severe conformational collapse (e.g., DecompDiff in 1JN2)—the spatial extension and receptor complementarity of LiFT closely parallel, and occasionally surpass,

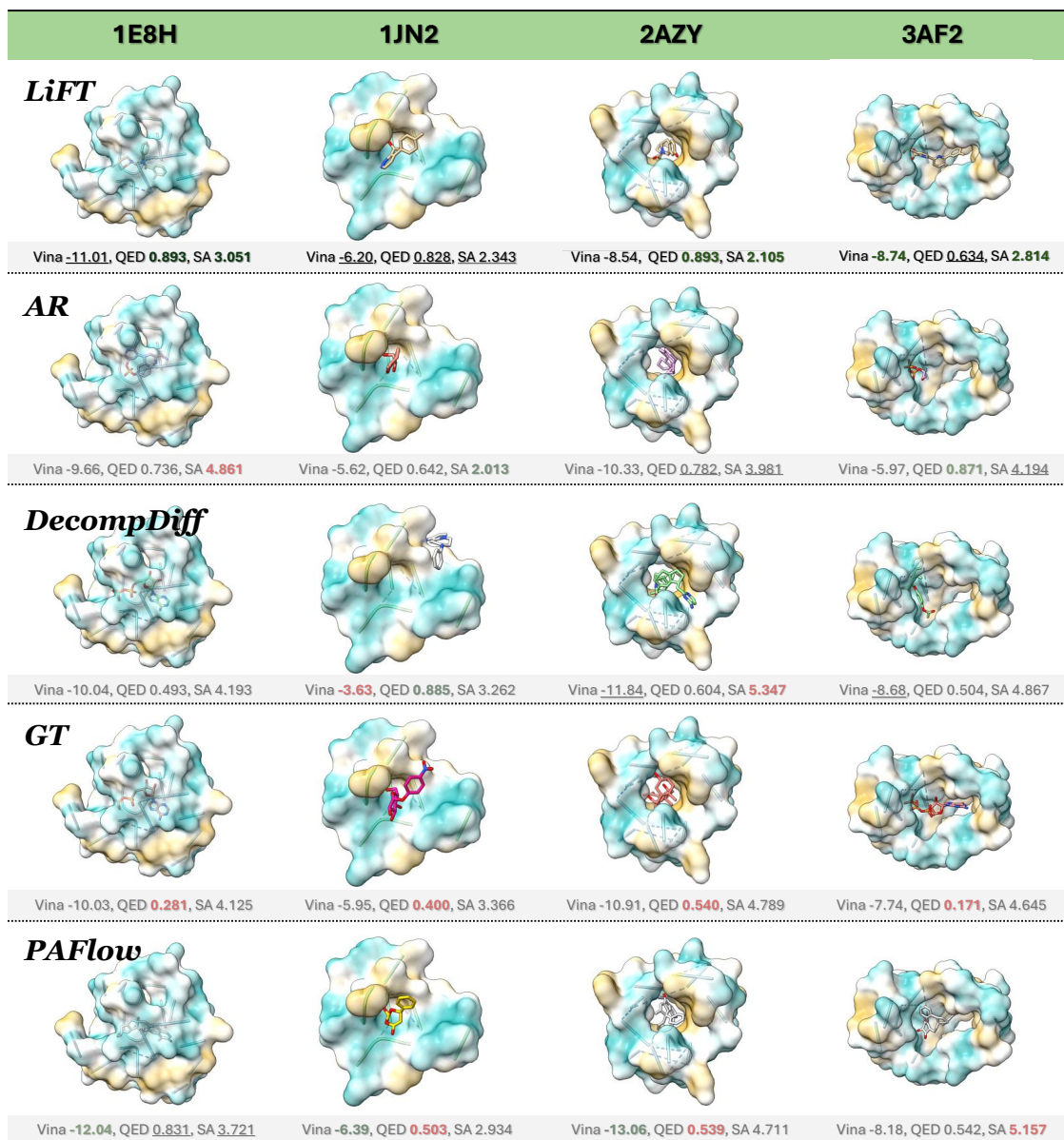


Figure 7: **Qualitative 3D visualization of generated ligands within four representative target pockets (1E8H, 1JN2, 2AZY, and 3AF2).** The grid compares the binding poses and geometric complementarity of molecules generated by *LiFT*, AR, DecompDiff, and PAFlow, alongside the Ground Truth (GT) references. Key pharmacological metrics (AutoDock Vina binding affinity, QED, and SA) are annotated below each complex. Green bold text signifies superior performance or highly desirable drug-like properties, whereas red text highlights severe metric degradation (e.g., collapsed QED or SA scores despite artificially high Vina affinity). *LiFT* uniquely achieves an exceptional equilibrium, matching or exceeding GT geometric complementarity while strictly maintaining optimal synthesizability and drug-likeness.

Table 20: PoseBusters sub-filter pass rates on CrossDocked2020. The table reports diagnostic 3D validity checks, excluding the overall PoseBusters score already reported in Table 2. All entries are percentages, and higher values indicate fewer geometric or physical violations. Best generated-method results are highlighted in **bold**, and second best are underlined; ties share the same mark.

Category	Method	PoseBusters 3D Validity Checks (%) $\uparrow$					
		Ring Flat	Bond Ang.	Double Bond	Int. Clash	Max. Lig.-Prot. Dist.	Protein Overlap
3D-based Models	AR (Luo et al., 2021)	90.56	85.28	94.80	91.03	98.93	98.79
	PAFlow (Zhou et al., 2025)	99.76	32.14	99.47	72.45	98.78	98.24
	BindDM (Huang et al., 2024)	99.98	45.64	100.00	86.63	98.91	97.85
	DecompDiff (Guan et al., 2023b)	99.97	92.82	96.52	93.92	98.81	98.81
	TargetDiff (Guan et al., 2023a)	<u>99.99</u>	74.13	<u>99.99</u>	91.18	98.95	98.43
	DrugFlow (Schneuing et al., 2025)	99.96	96.44	99.01	97.83	100.00	98.98
	Pocket2Mol (Peng et al., 2022)	99.60	99.01	99.31	99.19	<u>99.00</u>	98.55
LLM-based Models	TamGen (Wu et al., 2024)	100.00	99.85	100.00	100.00	8.40	97.95
	ELILLM-Diff (Hu et al., 2026)	99.96	<u>99.90</u>	94.58	<u>99.51</u>	5.65	98.19
	ELILLM-Rand (Hu et al., 2026)	99.97	99.94	97.76	99.35	6.58	98.02
LiFT Variants	Ligand-Ref Embedding	<u>99.99</u>	94.51	99.63	95.78	100.00	99.97
	LiFT (Bal-Reference)	99.97	94.03	99.24	95.84	100.00	98.94
	LiFT (QED-Reference)	99.97	93.85	99.49	95.52	100.00	99.66
	LiFT (Vina-Reference)	99.97	93.73	99.49	95.48	100.00	<u>99.93</u>
	LiFT (QED-No-Reference)	99.87	95.65	99.27	95.59	100.00	99.61
	LiFT (Vina-No-Reference)	99.93	95.96	99.59	97.01	100.00	98.35
	LiFT (Bal-No-Reference)	99.81	95.57	99.50	94.97	100.00	99.23

the GT ligands. LiFT successfully captures the authentic geometric binding modes while dramatically elevating the pharmacological properties (e.g., achieving exceptional QED scores of 0.893 in both 1E8H and 2AZY). Furthermore, LiFT maintains highly competitive Vina affinity (e.g., -11.01 in 1E8H) alongside highly reasonable SA values. This substantiates that our SCDR mechanism guarantees physical complementarity without sacrificing synthesizability, elegantly avoiding the common pitfall of trading drug-like properties for artificially inflated affinity scores—a catastrophic trade-off explicitly observed in PAFlow, which forces high Vina scores (e.g., -13.06 in 2AZY) at the severe expense of QED and SA degradation.

G Detailed Mathematical Formulations

This appendix gives an implementation-faithful verification of LiFT’s heterogeneous conditional generator. We make explicit the continuous coordinate flow, the categorical atom/bond bridges, and the conditions under which SCDR injects semantic control without changing the underlying transition families or breaking the geometric structure of the GVP backbone.

G.1 Problem Setup and Heterogeneous State Space

We formulate target-aware ligand generation as a heterogeneous conditional generative process. Given a protein pocket $\mathcal{P}$, a semantic condition

$\mathbf{z}_{sem} \in \mathbb{R}^{d_{sem}}$, and a ligand size N fixed before the generative chain starts, LiFT learns

$$p_{\theta}(z_1 \mid \mathcal{P}, \mathbf{z}_{sem}, N), \quad (13)$$

with $\mathbf{z}_{sem}$ denoting the precomputed molecule-level semantic embedding routed by SCDR. It is a fixed conditioning vector for the generative trajectory, not a generated state variable and not an evaluator-side score. A ligand state is written as

$$\mathbf{z}_t = (\mathbf{X}_t, \mathbf{H}_t, \mathbf{E}_t). \quad (14)$$

The three components are

$$\begin{aligned} \mathbf{X}_t &\in \mathbb{R}^{N \times 3}, \\ \mathbf{H}_t &\in \{0, 1\}^{N \times K_h}, \\ \mathbf{E}_t &\in \{0, 1\}^{M \times K_e}. \end{aligned} \quad (15)$$

corresponding to coordinates, atom types, and bond types. N is the number of ligand atoms. For each forward pass, M is determined by a fixed candidate edge set chosen before the bridge trajectory starts. In the reported setting, this set contains all unordered ligand atom pairs, so $M = N(N-1)/2$ for a single molecule. The same fixed edge set is used throughout training and sampling for that trajectory. K_h, K_e are the atom and bond class counts.

The components live in different state spaces: $\mathbf{X}_t$ is Euclidean, whereas $\mathbf{H}_t$ and $\mathbf{E}_t$ are categorical. LiFT therefore uses a heterogeneous process:

$$\begin{aligned} \mathbf{X}_t &: \text{continuous flow matching,} \\ (\mathbf{H}_t, \mathbf{E}_t) &: \text{discrete Markov bridges.} \end{aligned} \quad (16)$$

The semantic condition enters the generative process through the neural parameterization of the dynamics. We write

$$\chi_t = (z_t, \mathcal{P}, t, \mathbf{z}_{sem}). \quad (17)$$

The learned dynamics return

$$\hat{\mathbf{v}}_\theta^X(\chi_t), \hat{p}_\theta^H(\chi_t), \hat{p}_\theta^E(\chi_t). \quad (18)$$

The two distributions $\hat{p}_\theta^H$ and $\hat{p}_\theta^E$ serve as terminal atom-type and bond-type predictors. SCDR modulates this parameterization while leaving the coordinate and categorical state spaces intact. The heterogeneous process fixes the transition families; semantic guidance acts through the learned dynamics.

G.2 Continuous Coordinate Flow Matching

The coordinate component uses Euclidean flow matching. Given a prior sample $\mathbf{X}_0$ and a data sample $\mathbf{X}_1$, LiFT uses, for $t \in [0, 1]$,

$$\mathbf{X}_t = (1 - t)\mathbf{X}_0 + t\mathbf{X}_1. \quad (19)$$

Coordinates use the pocket-centered frame implemented by `center_data`. Let $\mathbf{c}_\mathcal{P}$ be the pocket center before normalization. Training subtracts this same center from pocket and ligand coordinates before constructing the path:

$$\mathbf{P} \leftarrow \mathbf{P} - \mathbf{c}_\mathcal{P}, \mathbf{X}_1 \leftarrow \mathbf{X}_1 - \mathbf{c}_\mathcal{P}. \quad (20)$$

For pockets without explicit atoms, the same construction is anchored by the ligand center. The coordinate prior is sampled around the pocket center in the current frame,

$$\mathbf{X}_0 \sim \mathcal{N}(\mathbf{c}_\mathcal{P}^{\text{norm}}, \mathbf{I}), \quad (21)$$

where $\mathbf{c}_\mathcal{P}^{\text{norm}} = \mathbf{0}$ after pocket centering in the standard non-empty-pocket case. Thus $\mathbf{X}_0$ and $\mathbf{X}_1$ are expressed in one reference frame. This centering removes global translation. In the reported configuration, the model uses reflection-sensitive chiral edge features, so the architectural equivariance guarantee is stated for proper rotations: if pocket and ligand coordinates are jointly transformed by any $R \in SO(3)$, the GVP vector field transforms accordingly. The path velocity is

$$u_t^X = \frac{d\mathbf{X}_t}{dt} = \mathbf{X}_1 - \mathbf{X}_0. \quad (22)$$

The neural field predicts a scaled coordinate velocity:

$$\hat{\mathbf{v}}_\theta^X(\chi_t) \approx \frac{\mathbf{X}_1 - \mathbf{X}_0}{c_x}, \quad (23)$$

where $c_x > 0$ is the coordinate scaling constant used during training and sampling; we use $c_x = 2.7$. The atom-wise coordinate loss is

$$\ell_i^X = \frac{1}{3} \left\| \hat{\mathbf{v}}_\theta^X(\chi_t)_i - \frac{\mathbf{X}_{1,i} - \mathbf{X}_{0,i}}{c_x} \right\|_2^2. \quad (24)$$

The molecule-wise objective uses the mean over ligand atoms:

$$\mathcal{L}_X = \mathbb{E}_{t, \mathbf{X}_0, \mathbf{X}_1} \left[\frac{1}{N} \sum_{i=1}^N \ell_i^X \right]. \quad (25)$$

The same scaling appears in inference. With sampler step size

$$\delta_{\text{sam}} = \frac{1}{N_{\text{sam}}}, \quad (26)$$

Euler integration updates coordinates as

$$\mathbf{X}_{k+1} = \mathbf{X}_k + \delta_{\text{sam}} c_x \hat{\mathbf{v}}_\theta^X(\chi_k). \quad (27)$$

Sampling inverts the training scale. The learned coordinate dynamics approximate the velocity field of the linear path while remaining conditioned on the pocket and semantic prior through the neural parameterization.

G.3 Discrete Markov Bridges for Atom and Bond Types

Atom and bond identities are categorical; we therefore model them with discrete Markov bridges. The dynamics below are conditional on the ligand size N . For bond variables, the candidate edge set is fixed along one trajectory, so M defines a fixed categorical state space for that forward pass. Let

$$\mathbf{Y} \in \{\mathbf{H}, \mathbf{E}\} \quad (28)$$

denote either atom-type or bond-type variables. We write $\mathcal{I}_Y$ for its instance index set, with $|\mathcal{I}_H| = N$ and $|\mathcal{I}_E| = M$. Each $\mathbf{Y}_{t,i} \in \{0, 1\}^K$ is a one-hot categorical state, where $K = K_h$ for atoms and $K = K_e$ for bonds.

The initial state $\mathbf{Y}_{0,i}$ is sampled from a categorical prior π_0^Y , and the terminal state $\mathbf{Y}_{1,i}$ is the data one-hot label. The categorical priors are fixed before training and are not recomputed from test targets. In the reported configuration, π_0^H uses the empirical atom-type marginal and π_0^E uses a uniform bond prior. For each instance i , the bridge marginal from 0 to t is defined by

$$\bar{\mathbf{Q}}_t(\mathbf{Y}_{1,i}) = (1 - t)\mathbf{I} + t\mathbf{1}\mathbf{Y}_{1,i}^\top. \quad (29)$$

The induced marginal is

$$q(\mathbf{Y}_{t,i} \mid \mathbf{Y}_{0,i}, \mathbf{Y}_{1,i}) = \mathbf{Y}_{0,i} \bar{\mathbf{Q}}_t(\mathbf{Y}_{1,i}). \quad (30)$$

The graph-level bridge factorizes over atom or edge instances:

$$q(\mathbf{Y}_t \mid \mathbf{Y}_0, \mathbf{Y}_1) = \prod_{i \in \mathcal{I}_Y} q(\mathbf{Y}_{t,i} \mid \mathbf{Y}_{0,i}, \mathbf{Y}_{1,i}). \quad (31)$$

The product form defines the elementary categorical kernels. Coupling among atom and edge variables is introduced through the shared state-dependent context χ_t , which parameterizes the terminal distributions below. The transition keeps each variable categorical while increasing probability mass toward the terminal data state.

For a transition from time s to t , $0 \leq s < t \leq 1$, define

$$\beta_{t|s} = \frac{1-t}{1-s}. \quad (32)$$

The corresponding one-step bridge transition is

$$\mathbf{Q}_{t|s}(\mathbf{Y}_{1,i}) = \beta_{t|s} \mathbf{I} + (1 - \beta_{t|s}) \mathbf{1} \mathbf{Y}_{1,i}^\top. \quad (33)$$

The denominator is only evaluated for $s < 1$. At the final sampling step, $t = 1$ is allowed with $s < 1$, yielding $\beta_{1|s} = 0$ and a terminal transition fully determined by the predicted clean-state distribution. For a current one-hot state $\mathbf{Y}_{s,i}$, the true bridge transition is

$$q(\mathbf{Y}_{t,i} \mid \mathbf{Y}_{s,i}, \mathbf{Y}_{1,i}) = \mathbf{Y}_{s,i} \mathbf{Q}_{t|s}(\mathbf{Y}_{1,i}). \quad (34)$$

The model predicts a terminal distribution for every instance,

$$\hat{\pi}_{t,i}^Y := \hat{p}_{\theta,i}^Y(\chi_t) = \text{softmax}(f_{\theta,i}^Y(\chi_t)). \quad (35)$$

The model transition is obtained by replacing the unknown terminal one-hot state with this predicted distribution:

$$\hat{\mathbf{Q}}_{t|s,i}^\theta = \beta_{t|s} \mathbf{I} + (1 - \beta_{t|s}) \mathbf{1} (\hat{\pi}_{s,i}^Y)^\top. \quad (36)$$

$$p_\theta(\mathbf{Y}_{t,i} \mid \mathbf{Y}_{s,i}, \chi_s) = \mathbf{Y}_{s,i} \hat{\mathbf{Q}}_{t|s,i}^\theta. \quad (37)$$

The transition is row-stochastic: all entries are non-negative, and every row sums to

$$\beta_{t|s} + (1 - \beta_{t|s}) \sum_{k=1}^K \hat{\pi}_{s,i,k}^Y = 1. \quad (38)$$

LiFT optimizes the variational bridge objective over sampled bridge states. During training, let

r denote the sampled current time. The bridge step size is

$$\delta_{\text{tr}} = \frac{1}{N_{\text{tr}}}, \quad (39)$$

where the reported runs use $N_{\text{tr}} = 5000$. We set

$$r^+ = \min(r + \delta_{\text{tr}}, 1), \quad (40)$$

with $r < 1$ almost surely under the continuous time sampler. The KL term compares the true and learned one-step transitions from r to r^+ :

$$\mathcal{L}_Y^{VLB} = \mathbb{E}_{r,z_0,z_1,\mathbf{Y}_r} \left[\frac{1}{|\mathcal{I}_Y|} \sum_{i \in \mathcal{I}_Y} D_{\text{KL}}(q_i^Y \parallel p_{\theta,i}^Y) \right]. \quad (41)$$

The two distributions inside the KL are

$$\begin{aligned} q_i^Y &= q(\mathbf{Y}_{r^+,i} \mid \mathbf{Y}_{r,i}, \mathbf{Y}_{1,i}), \\ p_{\theta,i}^Y &= p_\theta(\mathbf{Y}_{r^+,i} \mid \mathbf{Y}_{r,i}, \chi_r). \end{aligned} \quad (42)$$

Equivalently, each instance-level KL is

$$D_{\text{KL}}(q_i^Y \parallel p_{\theta,i}^Y) = \sum_{k=1}^K q_{i,k}^{r^+} \log \frac{q_{i,k}^{r^+}}{p_{\theta,i,k}^{r^+}}, \quad (43)$$

with

$$\begin{aligned} q_{i,k}^{r^+} &= q(Y_{r^+,i} = k \mid Y_{r,i}, Y_{1,i}), \\ p_{\theta,i,k}^{r^+} &= p_\theta(Y_{r^+,i} = k \mid Y_{r,i}, \chi_r). \end{aligned} \quad (44)$$

The main objective uses this molecule-wise mean. The corresponding terminal cross-entropy form at the sampled time r is

$$\mathcal{L}_Y^{CE} = -\frac{1}{|\mathcal{I}_Y|} \sum_{i \in \mathcal{I}_Y} \langle \mathbf{Y}_{1,i}, \log \hat{\pi}_{r,i}^Y \rangle. \quad (45)$$

We use the VLB objective for the discrete bridge losses.

During sampling, for a step from t to $t' = t + \delta_{\text{sam}}$, the predicted terminal distribution defines

$$\hat{\mathbf{Q}}_{t'|t,i}^\theta = \beta_{t'|t} \mathbf{I} + (1 - \beta_{t'|t}) \mathbf{1} (\hat{\pi}_{t,i}^Y)^\top. \quad (46)$$

The next categorical state follows

$$p_\theta(\mathbf{Y}_{t',i} \mid \mathbf{Y}_{t,i}, \chi_t) = \mathbf{Y}_{t,i} \hat{\mathbf{Q}}_{t'|t,i}^\theta. \quad (47)$$

Sampling gives

$$\mathbf{Y}_{t',i} \sim \text{Cat}(p_\theta(\mathbf{Y}_{t',i} \mid \mathbf{Y}_{t,i}, \chi_t)). \quad (48)$$

Discrete variables therefore remain categorical during both training and inference.

G.4 Joint LiFT Objective

The heterogeneous training objective combines the coordinate flow loss with the discrete bridge losses for atom and bond variables:

$$\mathcal{L}_{\text{LiFT}} = \lambda_x \mathcal{L}_X + \lambda_h \mathcal{L}_H + \lambda_e \mathcal{L}_E. \quad (49)$$

$\mathcal{L}_X$ denotes the coordinate flow-matching loss; $\mathcal{L}_H$ and $\mathcal{L}_E$ denote the Markov bridge losses for atom and bond types. In the reported configuration, $\lambda_x = 1.0$, $\lambda_h = 50.0$, and $\lambda_e = 50.0$.

Some implementation variants can add auxiliary torsion, rotation, or translation losses for flexible pocket variables outside the ligand flow/bridge construction. These terms are not part of the ligand objective above; the continuous-discrete ligand objective remains the base generative objective.

The semantic condition $\mathbf{z}_{sem}$ enters these probability paths through the shared neural parameterization:

$$(\hat{\mathbf{v}}_\theta^X, \hat{\pi}_\theta^H, \hat{\pi}_\theta^E) = G_\theta(\chi_t). \quad (50)$$

This gives a clean separation between transition design and semantic modulation. The flow and bridge mechanisms define the continuous and categorical transition families; SCDR routes $\mathbf{z}_{sem}$ into their shared GVP parameterization. Semantic modulation changes how transition parameters are predicted while keeping the transition families fixed. Thus this appendix establishes a mathematically defined intervention channel for $\mathbf{z}_{sem}$; claims about controllable property changes are empirical intervention claims evaluated through GT/LLM and ablation protocols, not theorem-level consequences of the objective alone.

G.5 SCDR Equivariant Routing

We formulate SCDR as an equivariant semantic routing module for a GVP backbone. The main text denotes the router latent by $\mathbf{h}$; here we use $\mathbf{q}^{(l)}$ to emphasize that the same mechanism is instantiated at each GVP layer. At layer l , ligand node features are

$$\mathcal{H}^{(l)} = \{\mathbf{h}_i^{(l)}\}_{i=1}^N, \mathbf{h}_i^{(l)} = (\mathbf{s}_i^{(l)}, \mathbf{V}_i^{(l)}). \quad (51)$$

$\mathbf{s}_i^{(l)} \in \mathbb{R}^{d_s}$ and $\mathbf{V}_i^{(l)} \in \mathbb{R}^{d_v \times 3}$ are scalar and vector channels. An SCDR routing map

$$\mathcal{R}_{\theta_{\text{scdr}}}^{(l)} = \mathcal{R}_{\theta_{\text{scdr}}}^{(l)}(\mathcal{H}^{(l)}, \mathbf{z}_{sem}, t) \quad (52)$$

is constructed around four requirements: permutation invariance over ligand atoms, geometric invariance of the routing signal, bounded perturbation of the backbone, and stable initialization behavior.

Invariant node descriptor. A scalar routing network cannot consume raw vector channels $\mathbf{V}_i^{(l)}$ without fixing a coordinate frame. SCDR therefore accesses vector features through rotation-invariant quantities. We use the channel-wise vector norm:

$$\|\mathbf{V}_i^{(l)} R\|_2 = \|\mathbf{V}_i^{(l)}\|_2 \quad (\forall R \in SO(3)). \quad (53)$$

We define the invariant node descriptor as

$$\mathbf{u}_i^{(l)} = [\text{LN}_s(\mathbf{s}_i^{(l)}), \text{LN}_v(\|\mathbf{V}_i^{(l)}\|_2)]. \quad (54)$$

The descriptor remains state-dependent while being invariant to proper rotations. The reported backbone may use reflection-sensitive chiral edge features; the SCDR router itself does not introduce additional frame-dependent vector inputs.

DeepSets sensing. Since $\mathbf{z}_{sem}$ is molecule-level, the SCDR readout must be invariant to atom ordering. The DeepSets (Zaheer et al., 2017) representation principle gives a natural admissible family of permutation-invariant set functions:

$$F(\{\mathbf{u}_i\}_{i=1}^N) = \rho \left(\sum_{i=1}^N \varphi(\mathbf{u}_i) \right). \quad (55)$$

SCDR follows this family by first computing

$$\mathbf{r}_i^{(l)} = \varphi^{(l)}(\mathbf{u}_i^{(l)}), \quad (56)$$

then aggregating node states into an invariant structural summary:

$$\Psi_{str}^{(l)} = [\text{mean}_i \mathbf{r}_i^{(l)}, \max_i \mathbf{r}_i^{(l)}, \text{std}_i \mathbf{r}_i^{(l)}, \log N]. \quad (57)$$

The mean, maximum, and standard deviation encode global level, extreme local responses, and structural heterogeneity, while $\log N$ preserves molecular scale. The resulting $\Psi_{str}^{(l)}$ is a compact DeepSets-motivated statistic of the current 3D molecular state, and is both permutation-invariant and geometrically invariant.

SCDR routing. Let $\tau(t)$ denote the integration-time embedding used by the backbone. We instantiate it as an RBF embedding of t when time embedding is enabled, and as the scalar t otherwise. This integration-time signal is distinct from the training-progress warmup variable p used below to activate SCDR parameters smoothly during optimization. SCDR uses the routing inputs

$$[\mathbf{z}_{sem}, \tau(t), \Psi_{str}^{(l)}]. \quad (58)$$

The semantic–time branch is

$$\mathbf{c}^{(l)} = \psi_c^{(l)}([\mathbf{z}_{sem}, \boldsymbol{\tau}(t)]), \quad (59)$$

and the structural branch is

$$\mathbf{g}^{(l)} = \psi_s^{(l)}(\Psi_{str}^{(l)}). \quad (60)$$

After normalization,

$$\hat{\mathbf{c}}^{(l)} = \text{LN}(\mathbf{c}^{(l)}), \hat{\mathbf{g}}^{(l)} = \text{LN}(\mathbf{g}^{(l)}). \quad (61)$$

SCDR uses two gates:

$$a_{\text{time}}^{(l)} = \sigma\left(\eta_t^{(l)}(\boldsymbol{\tau}(t))\right), \quad (62)$$

$$a_c^{(l)} = \sigma\left(\eta_c^{(l)}([\hat{\mathbf{c}}^{(l)}, \hat{\mathbf{g}}^{(l)}])\right). \quad (63)$$

The shared router latent is

$$\mathbf{q}^{(l)} = \text{LN}\left(\rho^{(l)}\left(\hat{\mathbf{c}}^{(l)} + a_{\text{time}}^{(l)} a_c^{(l)} \hat{\mathbf{g}}^{(l)}\right)\right). \quad (64)$$

Because both branches are built from invariant inputs, $\mathbf{q}^{(l)}$ is an invariant molecule-level routing variable.

Decoupled bounded dispatching. Since $\mathbf{q}^{(l)}$ is invariant, it can generate scalar modulation coefficients without breaking equivariance. We denote the SCDR parameter subset by $\theta_{\text{schr}} \subset \theta$. For scalar states, SCDR first applies the semantic AdaLN path. Write $\bar{\gamma}^{(l)} = \tanh \gamma^{(l)}(\mathbf{z}_{sem})$ and $\bar{\beta}^{(l)} = \beta^{(l)}(\mathbf{z}_{sem})$. Then

$$\tilde{\mathbf{s}}_i^{(l)} = \text{LN}_s(\mathbf{s}_i^{(l)}), \mathbf{s}_{base,i}^{(l)} = \tilde{\mathbf{s}}_i^{(l)} \odot (1 + \bar{\gamma}^{(l)}) + \bar{\beta}^{(l)}. \quad (65)$$

This base AdaLN path uses the semantic condition alone. The full SCDR routing is obtained by combining this semantic state modulation with the structure-aware latent $\mathbf{q}^{(l)}$ in the gates below. For vector updates, equivariance permits scalar gating but forbids arbitrary vector bias. SCDR therefore separates scalar state modulation from residual update modulation:

$$\begin{aligned} \alpha_{s,state}^{(l)} &= 1 + w_{\text{prog}}(p) b_{a_s}(W_{s,state}^{(l)} \mathbf{q}^{(l)}), \\ \alpha_{s,upd}^{(l)} &= 1 + w_{\text{prog}}(p) b_{a_u}(W_{s,upd}^{(l)} \mathbf{q}^{(l)}), \\ \alpha_{v,upd}^{(l)} &= 1 + w_{\text{prog}}(p) b_{a_v}(W_{v,upd}^{(l)} \mathbf{q}^{(l)}) \end{aligned} \quad (66)$$

where

$$b_a(x) = a \frac{x}{1 + |x|}. \quad (67)$$

For vector inputs, b_a is applied componentwise. In the reported implementation, $a_s = 0.5$ for

the scalar state gate and $a_u = a_v = 0.9$ for the scalar/vector update gates. The variable $p \in [0, 1]$ denotes training progress, and $w_{\text{prog}}(p) \in [0, 1]$ is the cosine warmup coefficient used when an explicit progress schedule is supplied. This training-progress gate is separate from the integration-time gate $a_{\text{time}}(t)$. The routed scalar state is

$$\hat{\mathbf{s}}_i^{(l)} = \mathbf{s}_{base,i}^{(l)} + \tilde{\mathbf{s}}_i^{(l)} \odot (\alpha_{s,state}^{(l)} - 1), \quad (68)$$

and the routed feed-forward update is

$$\Delta \hat{\mathbf{s}}_i^{(l)} = \alpha_{s,upd}^{(l)} \odot \Delta \mathbf{s}_i^{(l)}, \Delta \hat{\mathbf{V}}_i^{(l)} = \alpha_{v,upd}^{(l)} \odot \Delta \mathbf{V}_i^{(l)}. \quad (69)$$

The vector gate is channel-wise scalar broadcast: if $\alpha_{v,upd}^{(l)} \in \mathbb{R}^{d_v}$ and $\Delta \mathbf{V}_i^{(l)} \in \mathbb{R}^{d_v \times 3}$, then

$$\left(\Delta \hat{\mathbf{V}}_i^{(l)}\right)_{c,:} = \alpha_{v,upd,c}^{(l)} \left(\Delta \mathbf{V}_i^{(l)}\right)_{c,:}, \quad c = 1, \dots, d_v. \quad (70)$$

The routed form can be summarized as

$$[\mathbf{z}_{sem}, \boldsymbol{\tau}(t), \Psi_{str}^{(l)}] \mapsto (\alpha_{s,state}^{(l)}, \alpha_{s,upd}^{(l)}, \alpha_{v,upd}^{(l)}) \quad (71)$$

where $\Psi_{str}^{(l)}$ is the invariant structural statistic defined above. SCDR is a bounded, permutation-invariant, geometry-preserving routing module for semantic modulation of GVP dynamics.

G.6 Equivariance and Stability Proofs

We verify that the SCDR-modulated GVP block preserves the required geometric symmetries and remains a bounded perturbation of the original backbone.

Proposition 1: equivariance preservation. Assume the base GVP block is permutation equivariant over nodes and $SO(3)$ -equivariant over vector channels under the reported reflection-sensitive configuration. For any node permutation Π and any proper rotation $R \in SO(3)$, the base block satisfies

$$\begin{aligned} F_s(\Pi \mathbf{s}, \Pi \mathbf{V} R) &= \Pi F_s(\mathbf{s}, \mathbf{V}), \\ F_V(\Pi \mathbf{s}, \Pi \mathbf{V} R) &= \Pi F_V(\mathbf{s}, \mathbf{V}) R. \end{aligned} \quad (72)$$

Then the SCDR-modulated block preserves the same transformation laws.

Proof. SCDR reads vector channels only through norms:

$$\|\mathbf{V}_i R\|_2 = \|\mathbf{V}_i\|_2. \quad (73)$$

Each node descriptor $\mathbf{u}_i$ is therefore $SO(3)$ -invariant. Since Ψ_{str} is built from permutation-invariant pooling operations (mean, max, std, and

count),

$$\Psi_{str}(\Pi s, \Pi V R) = \Psi_{str}(s, V). \quad (74)$$

Consequently, the router latent and all routing gates are invariant:

$$\alpha(\Pi s, \Pi V R) = \alpha(s, V), \quad (75)$$

for

$$\alpha \in \{\alpha_{s,state}, \alpha_{s,upd}, \alpha_{v,upd}\}. \quad (76)$$

The AdaLN coefficients $\gamma(\mathbf{z}_{sem})$ and $\beta(\mathbf{z}_{sem})$ depend only on the molecule-level semantic condition and are broadcast to ligand nodes by the batch index. They remain invariant under node permutation and under $SO(3)$ transformations, so scalar modulation preserves permutation equivariance. For the vector update, since $\alpha_{v,upd}$ is an invariant channel-wise scalar broadcast over the Cartesian dimension,

$$\begin{aligned} \Delta \hat{\mathbf{V}}(\Pi s, \Pi V R) &= \alpha_{v,upd} \odot (\Pi \Delta \mathbf{V} R) \\ &= \Pi (\alpha_{v,upd} \odot \Delta \mathbf{V}) R \\ &= \Pi \Delta \hat{\mathbf{V}}(s, V) R. \end{aligned} \quad (77)$$

Hence the SCDR-modulated block remains permutation equivariant and $SO(3)$ -equivariant under the reported configuration.

Proposition 2: bounded routing. SCDR yields a bounded local amplification of the original GVP update.

Proof. The dispatching nonlinearity is

$$b_a(x) = a \frac{x}{1 + |x|}. \quad (78)$$

This implies

$$|b_a(x)| < a. \quad (79)$$

Since

$$\alpha = 1 + w_{\text{prog}}(p) b_a(W \mathbf{q}), \quad 0 \leq w_{\text{prog}}(p) \leq 1, \quad (80)$$

we have

$$1 - a < \alpha < 1 + a. \quad (81)$$

The vector update is consequently bounded:

$$\|\Delta \hat{\mathbf{V}}\| = \|\alpha_{v,upd} \odot \Delta \mathbf{V}\| \leq (1 + a_v) \|\Delta \mathbf{V}\|. \quad (82)$$

The same argument applies to scalar updates, so each routed block remains a finite multiplicative perturbation of the corresponding backbone update.

Proposition 3: initialization behavior. At initialization, the AdaLN state path recovers the scalar-normalized ligand state of the replaced normalization path. The vector state is passed unchanged, and either the warmup coefficient or the small-variance update initialization keeps the residual gates near identity at startup.

Proof. The AdaLN projection is zero-initialized:

$$\gamma(\mathbf{z}_{sem}) = 0, \beta(\mathbf{z}_{sem}) = 0. \quad (83)$$

At the start of an explicit training-progress warmup schedule,

$$w_{\text{prog}}(p) = 0. \quad (84)$$

The routing gates then satisfy

$$\alpha_{s,state} = \alpha_{s,upd} = \alpha_{v,upd} = 1. \quad (85)$$

Substituting these values gives

$$\hat{s} = \text{LN}_s(s), \Delta \hat{s} = \Delta s, \Delta \hat{\mathbf{V}} = \Delta \mathbf{V}. \quad (86)$$

The scalar AdaLN projection starts exactly from the scalar-normalized ligand state, while the vector state path is unchanged. When the implementation runs with active gates from the beginning, the gate heads are initialized with small variance and zero bias; because $b_a(0) = 0$ and b_a is Lipschitz around the origin, the multiplicative gates remain near one at startup. Proposition 2 controls their range throughout training.

Together, these propositions characterize SCDR as an invariant, bounded, and stably initialized routing mechanism. It preserves the geometric structure of the GVP backbone while enabling semantic modulation.

G.7 Inference Consistency

Conditional sampling uses the transition families optimized during training and keeps the same semantic-conditioning interface as the training dynamics. Before the chain starts, the ligand size N , candidate edge set, categorical priors, and semantic condition $\mathbf{z}_{sem}$ are fixed for the entire trajectory; they are not changed by downstream docking, property scores, reference molecules, or filtering. The generative chain starts from

$$\mathbf{X}_0 \sim \mathcal{N}(\mathbf{c}_p^{\text{norm}}, \mathbf{I}), \mathbf{H}_0 \sim \pi_0^H, \mathbf{E}_0 \sim \pi_0^E, \quad (87)$$

in the same pocket-centered frame used during training. Let $\delta_{\text{sam}} = 1/N_{\text{sam}}$; the reported runs

use $N_{\text{sam}} = 500$. For $k = 0, \dots, N_{\text{sam}} - 1$ with $t_k = k/N_{\text{sam}}$, the model forms

$$z_k = (\mathbf{X}_k, \mathbf{H}_k, \mathbf{E}_k), \chi_k = (z_k, \mathcal{P}, t_k, \mathbf{z}_{\text{sem}}) \quad (88)$$

and predicts

$$\hat{\mathbf{v}}_{\theta}^X(\chi_k), \hat{\pi}_k^H, \hat{\pi}_k^E. \quad (89)$$

At every solver step, $\mathbf{z}_{\text{sem}}$ is passed to the same SCDD-modulated GVP parameterization used during training; coordinate and categorical components then evolve according to their native transition rules.

For coordinates, the model is trained to predict the scaled velocity

$$\hat{\mathbf{v}}_{\theta}^X \approx \frac{\mathbf{X}_1 - \mathbf{X}_0}{c_x}. \quad (90)$$

Euler sampling applies the inverse scaling:

$$\mathbf{X}_{k+1} = \mathbf{X}_k + \delta_{\text{sam}} c_x \hat{\mathbf{v}}_{\theta}^X(\chi_k). \quad (91)$$

The coordinate sampler thereby integrates the velocity field learned by the flow-matching objective.

For each discrete variable

$$\mathbf{Y} \in \{\mathbf{H}, \mathbf{E}\}, \quad (92)$$

the terminal distribution for instance $i \in \mathcal{I}_Y$ has the softmax form

$$\hat{\pi}_{k,i}^Y = \text{softmax}(f_{\theta,i}^Y(\chi_k)). \quad (93)$$

The corresponding one-step bridge matrix is

$$\hat{\mathbf{Q}}_{k+1|k,i}^{\theta} = \beta_k \mathbf{I} + (1 - \beta_k) \mathbf{1}(\hat{\pi}_{k,i}^Y)^{\top}, \quad (94)$$

where

$$\beta_k = \frac{1 - t_{k+1}}{1 - t_k}. \quad (95)$$

The next state is sampled as

$$\mathbf{Y}_{k+1,i} \sim \text{Cat}(\mathbf{Y}_{k,i} \hat{\mathbf{Q}}_{k+1|k,i}^{\theta}). \quad (96)$$

Inference therefore follows the same transition families used by the training objectives. Coordinates use the learned flow-matching velocity; atom and bond updates use learned terminal distributions inside the Markov bridge family. SCDD affects inference only through the shared neural parameterization, leaving the continuous and categorical transition structures unchanged.

G.8 Summary of the Verification

The formulation positions LiFT as a structured heterogeneous continuous-discrete generative model. Coordinates follow a flow-matching vector field, while atom and bond identities follow categorical Markov bridges. The semantic condition $\mathbf{z}_{\text{sem}}$ modulates the shared GVP parameterization while preserving these state spaces.

SCDD is formulated as an equivariant routing module with four preserved properties: permutation invariance, proper-rotation equivariance under the reported configuration, bounded local amplification, and stable initialization. Its DeepSets-motivated structural sensing provides an invariant summary of the current 3D molecular state, and its decoupled scalar gates inject semantic modulation through equivariant vector-channel scaling.

Overall, LiFT adds SCDD routing to the heterogeneous flow/bridge process while preserving boundedness, near-identity initialization, and equivariance. The appendix therefore establishes structural compatibility between semantic routing and the heterogeneous generative dynamics. Chemical validity, affinity gains, and semantic controllability are established empirically by the experiments and intervention ablations rather than by the symmetry proof alone.

H Extended Comparison with Related Work

This appendix provides a structured comparison between LiFT and representative approaches spanning 1D–3D molecular generation, protein-conditioned sequence generation, LLM-assisted molecular design, and native pocket-conditioned 3D SBDD. The comparison is organized around neutral methodological dimensions—language input, target representation, the role of the 1D representation, native output, trainable components, interaction stage, and trajectory-time control—rather than a LiFT-specific ranking criterion. The goal is to clarify which interface each method addresses and where language- or sequence-derived information enters the molecular design pipeline.

H.1 Systematic Methodological Comparison

Table 21 summarizes the main methodological distinctions. Existing approaches cover several complementary interfaces. NExT-Mol (Liu et al., 2025) and MolSculpt (Chen et al., 2025) connect molecular-language representations with target-

free 3D molecular generation. MolChord (Zhang et al., 2025) and TamGen (Wu et al., 2024) use protein information to condition molecular-sequence generation, with the generated molecular string serving as the primary molecular proposal. ELILLM (Hu et al., 2026) searches a pretrained LLM latent space using target-dependent optimization feedback, while CIDD (Gao et al., 2025) applies LLM-based refinement after an initial SBDD generation stage. Native 3D SBDD methods such as AR (Luo et al., 2021), Pocket2Mol (Peng et al., 2022), TargetDiff (Guan et al., 2023a), Decom-pDiff (Guan et al., 2023b), PAFflow (Zhou et al., 2025), and DrugFlow (Schneuing et al., 2025) directly model pocket-conditioned ligand geometry without an open-ended language-derived semantic condition.

LiFT occupies a different interface within this landscape: open-ended design preferences and pocket-derived specifications are converted into a SMILES-derived soft semantic prior rather than a final ligand, while the final molecular topology and pose are generated in the native pocket-conditioned 3D space. The semantic condition acts inside the flow trajectory, and SCDR adjusts its influence according to flow time and the evolving ligand state. This comparison concerns the location and role of cross-modal conditioning; it does not imply that the individual components used by LiFT are themselves new molecular or geometric primitives.

H.2 Focused Comparison with ELILLM

ELILLM (Hu et al., 2026) is particularly relevant because both methods connect language-model representations with structure-based molecular design, but the representations play different roles and act at different stages. Table 22 therefore separates the two pipelines along the dimensions most directly related to their problem formulation.

In short, ELILLM searches and optimizes a final 1D candidate in an LLM latent space and subsequently obtains its 3D evaluation through conventional conformer construction and docking. LiFT instead uses language-derived chemical semantics as a soft condition that directly operates inside a native pocket-conditioned 3D generation trajectory. Accordingly, the two methods address related SBDD goals through different interfaces: ELILLM focuses on molecular-candidate optimization in language-model latent space, whereas LiFT focuses on state-aware cross-modal conditioning of an evolving 3D generator.

Table 21: Systematic methodological comparison of LiFT with representative related approaches. The dimensions describe where language, molecular strings, protein information, and 3D generation interact. “Native output” denotes the direct output of the generative model before optional post-hoc conformer construction, docking, or refinement.

Method	Language Input	Target representation	Rep-resentation	Role of 1D Repr-entation	Native Output	Trainable Components / Objective	Interaction Stage	Trajectory-Time Control
LiFT	Open-ended task preferences, pocket descriptions, property trade-offs, and optional references	Native protein pocket	3D	Intermediate SMILES-derived soft semantic prior , not the final ligand	Pocket-conditioned 3D ligand and aligned pose	Lightweight semantic projector and SCDR are optimized under the existing flow-matching objective while retaining the standard geometric backbone	Semantic information acts inside the 3D flow trajectory	Dynamic routing according to flow time and the evolving 3D ligand state
NExT-Mol (Liu et al., 2025)	No open-ended human design instruction; molecular string is provided	No target	protein	SELFIES specifies the molecular identity to be reconstructed in 3D	A conformer of the specified molecule	1D-3D representation alignment and diffusion-based conformer generation	Molecular-string information is aligned with the corresponding 3D molecule during conformer generation	Fixed molecule-level alignment; no pocket-conditioned semantic routing
MolSculpt (Chen et al., 2025)	No open-ended task-language interface	No target	protein	SELFIES-derived molecular-language features provide generic chemical knowledge	Target-free 3D molecule	Learnable queries/projector and a 3D diffusion objective	Molecular-language features condition diffusion	No target-pocket state or language-guided state-dependent routing
MolChord (Zhang et al., 2025)	Text-aligned molecular representations, but not open-ended LiFT-style task control	Protein structure representation		Autoregressively decoded molecular string is the final molecular proposal	SMILES; 3D conformer and docking are obtained afterward	Structure adapter and molecular sequence decoder trained with a sequence-generation objective	Protein representation conditions 1D decoding	No native 3D generation trajectory
TamGen (Wu et al., 2024)	No open-ended human design-language interface	Protein pocket representation	or	Generated SMILES is the final molecular proposal	SMILES; conformer construction and docking are post hoc	Target-conditioned chemical language model with an autoregressive objective	Target information conditions sequence decoding	No native 3D generation trajectory
ELILLM (Hu et al., 2026)	Optimization objectives rather than an open-ended 3D design interface	Target enters primarily through a docking oracle		Optimized latent is decoded into the final SMILES candidate	SMILES; conformer generation and docking follow	Bayesian optimization in a pretrained LLM latent space	Control during latent-space exploration before 3D evaluation	No interaction with an evolving 3D ligand state
CIDD (Gao et al., 2025)	LLM refinement instructions informed by interaction analysis	Protein pocket and docking-derived feedback		The molecular string represents an already generated candidate to be edited	Refined molecular proposal after an initial SBDD stage	Prompted or iterative LLM refinement using docking/interaction feedback	Language intervenes after initial molecule generation	Post-generation refinement rather than trajectory-internal control
AR / Pocket2Mol / TargetDiff / Decom-pDiff / PAFLOW / DrugFlow (Luo et al., 2021; Peng et al., 2022; Guan et al., 2023a,b; Zhou et al., 2025; Schneuing et al., 2025)	No open-ended human language instruction	Native protein pocket	3D	Usually no language-derived 1D semantic condition	Pocket-conditioned 3D ligand and pose	Autoregressive-, diffusion-, or flow-based generation objectives	Pocket geometry participates throughout native 3D generation	State-dependent geometric generation, but no language-derived semantic routing

Table 22: Focused methodological comparison between ELILLM and LiFT. The distinction is primarily the role of the language-derived representation and the space in which target-aware control occurs.

Comparison Dimension	ELILLM (Hu et al., 2026)	LiFT
Core problem	Searches for molecular candidates with improved docking objectives in a pretrained LLM latent space	Enables language design intent to participate in native pocket-conditioned 3D generation
Role of language/latent representation	Optimized and decoded into the final SMILES candidate	Forms a chemical-semantic prior that is not the final output
How target information enters	Primarily through a docking oracle and Bayesian-optimization feedback	The 3D pocket directly conditions the flow backbone and is also interpreted by the agent to form a pocket-derived specification
Native generation space	LLM latent / 1D sequence space	3D ligand state space in the target-pocket coordinate frame
Native output	SMILES	Atom types, 3D coordinates, and a pocket-aligned pose
How the 3D structure is obtained	Conformer embedding, optimization, and docking after sequence generation	Native generation within a pocket-conditioned flow trajectory
Where control occurs	Before generation, through search in the LLM latent space	During generation, by dynamically modulating the velocity field according to the intermediate 3D state
State-aware 3D routing	No	Yes, through SCDR
Adaptation to a new task	Changes the optimization objective and reruns latent-space search	Changes the inference-time language instruction without retraining the generator