

Decomposing Visual Histories with Vision-Language Agents: Hierarchical Temporal Guidance for Compositional Image Generation

Lujundong Li^{1*}, Mingxu Zhang¹, Songning Lai², Jiayu Yang^{1,3}, Yuxuan Fan⁴,
Ziwei Xie⁵, Feng Liu⁵, Changwang Zhang⁵, Jun Wang⁵, Dazhong Shen^{6†}, Ying Sun^{7†}

¹AI Thrust, Information Hub, The Hong Kong University of Science and Technology (Guangzhou)

²JoinQuant ³Tencent Hunyuan

⁴College of Computing and Data Science, Nanyang Technological University ⁵OPPO Research Institute

⁶College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics

⁷National University of Defense Technology

edleebit@gmail.com, shendazhong@nuaa.edu.cn, sunyinggilly@gmail.com

Abstract

Reference-conditioned image generation increasingly relies on visual histories that rarely speak with one voice: one example pins the subject, another fixes a layout, a third only suggests a palette. Current image-conditioned diffusion methods encode these references once, average them into a single vector, and inject it at every denoising step, so conflicting cues collide and the output is faithful to none. We propose the Hierarchical Temporal Guidance Framework (HTGF), a training-free pipeline that reframes multi-reference conditioning as a *temporal routing* problem. A semantic decomposer reads each reference and emits a soft routing distribution over three axes (Subject, Structure, Detail), and a closed-form SNR-sensitivity argument places each axis in the denoising window where it has the most leverage. A short manifold-aware corrector then smooths the trajectory at the stage boundaries where the active condition changes. HTGF adds no training and no architectural change to the diffusion backbone. On three datasets it outperforms strong VLM- and diffusion-based baselines, gaining +11.7 CIS over LaVIT on Movie Poster and FID 38.87 vs. 50.12 for the best baseline under noisy histories, and degrading gracefully down to the zero-reference limit.

1 Introduction

Modern image generation increasingly leans on visual references. A designer iterates against a mood board; a personalized model conditions on a user’s interaction history; a multi-modal interface stitches together prompts of mixed text and images. In each of these settings the user supplies several reference

images, and in each of them those references rarely speak with one voice: one example pins the subject, another suggests a layout, a third only fixes a palette, and a fourth may be incidental. The dominant practice in image-conditioned diffusion is to ignore this disagreement entirely. References are encoded once, the embeddings are summed or averaged, and the resulting static vector is injected at every denoising step (Ye et al., 2023; Jin et al., 2024b; Xu et al., 2025a). The compromise is destructive: a layout cue from one reference fights a texture cue from another and the output is faithful to neither; a single stylistic outlier drags the entire generation off-prompt.

The natural diagnosis is that two kinds of structure are being thrown away at once. References are not interchangeable: some are most informative about *what* the image depicts, others about *where* things are arranged, others about *how* surfaces look. And denoising is not a stationary function but a trajectory that lays down coarse semantics under high noise, then layout, then surface detail. The harder problem, and the one this paper addresses, is that these two structures cannot be exploited in isolation. A vision-language model can decompose references into semantic axes, but decomposition without temporal placement still merges every axis at every step. A hierarchical sampler can stage the denoising schedule, but staging without decomposition has nothing to route. Each half on its own collapses back into flat fusion. The two structures are *coupled*, and the open question is what couples them: what tells you that *this* reference, scored on *this* axis, belongs in *this* denoising window. Without a principled answer, the schedule reduces to another tuned hyperparameter that has to be re-discovered on every dataset.

This paper develops a principled answer. We observe that an image-conditioning vector’s in-

* Work done during an internship at OPPO Research Institute.

† Corresponding authors.

fluence on the predicted clean image follows in closed form from the denoiser’s noise level: under the standard Tweedie formulation (§3.3, Proposition 1), high-noise steps have globally amplified leverage and low-noise steps have only local leverage. This SNR-sensitivity motivates aligning each semantic axis with a specific denoising window, placing Subject in the high-noise regime and Detail in the low-noise regime. A semantic decomposer scores each reference along the three axes, a soft routing distribution then routes each reference, and the SNR-derived window supplies the placement. The remaining failure mode is mechanical: stage-aligned injection introduces discontinuities at the boundaries between windows, which is precisely the regime where any scheduling-based approach breaks. A short manifold-aware corrector, guided on the Tweedie-estimated clean image rather than the noisy latent, smooths those discontinuities. The resulting pipeline, HTGF,¹ adds no training and no architectural change to the diffusion backbone.

What follows from taking this trajectory structure seriously goes beyond the headline numbers. The routing distribution is well-defined for any number of references, so the system degrades gracefully as the reference set shrinks, down to the zero-reference limit (§4.6). The schedule rests on a smooth sensitivity, so performance is insensitive to the exact phase boundaries across a wide window (§4.5). And because the decomposition is filtered by a confidence threshold before any reference is allowed to contribute, the framework stays stable on noisy reference sets where flat-fusion baselines collapse.

Contributions.

- We identify multi-reference diffusion conditioning as a *coupled* problem of semantic decomposition and temporal placement: decomposition without placement reverts to flat fusion, placement without decomposition has nothing to route. Prior work addresses at most one side.
- We show that the placement is principled rather than arbitrary: a closed-form Tweedie-sensitivity argument (Proposition 1) aligns each semantic axis with a specific denoising window, grounding the schedule in the

denoiser’s SNR profile. A manifold-aware corrector absorbs the trajectory discontinuities that emerge from stage-aligned injection, the principal failure mode of any scheduling-based approach.

- HTGF, the training-free instantiation of this principle, achieves +11.7 CIS over LaVIT on Movie Poster, FID 38.87 vs. 50.12 for the best baseline under noisy histories, and graceful degradation in the cold-start limit.

2 Related Work

2.1 Multi-Reference and Personalized Image Generation

Personalized image generation has progressed from per-concept optimization (Gal et al., 2023; Ruiz et al., 2023; Kumari et al., 2023) to tuning-free encoder-based architectures that inject identity through decoupled cross-attention or stacked embeddings (Ye et al., 2023; Li et al., 2024; Shi et al., 2024), ID landmarks (Wang et al., 2024; Parihar et al., 2024), or layout guidance (Wang et al., 2025). Multi-subject extensions push the same recipe to several explicit references (He et al., 2024; Tewel et al., 2024; Chen et al., 2025; Lin et al., 2026a). MLLM-conditioned encoders such as EasyRef (Zong et al., 2025), OmniGen (Xiao et al., 2025), and LaVIT (Jin et al., 2024b) accept arbitrary reference groups but treat them as a flat token sequence, while Pigeon (Xu et al., 2025a), DRC (Xu et al., 2025b), and PMG (Shen et al., 2024b) filter or disentangle these signals before fusion. Across this line of work, the conditioning vector is constructed once and reused at every denoising step—precisely the assumption HTGF removes. We do not assume a clean reference set: the inputs are heterogeneous, individually noisy, and routinely disagree on subject, layout, or texture. Because each reference is scored along explicit axes, the resulting control policy is also interpretable by construction, in the spirit of self-interpretable models (Ji et al., 2025).

2.2 Language-Driven Control and Retrieval for Image Generation

A growing line of work places a language model in the loop of image generation. Instruct-Imagen (Hu et al., 2024) and unified multimodal LLMs (Koh et al., 2023; Jin et al., 2024b) consume mixed text–image instructions; MLLM-guided editing (Fu

¹Code, additional details, and results are available at <https://github.com/LjdLi/Implementation-of-HMGF>.

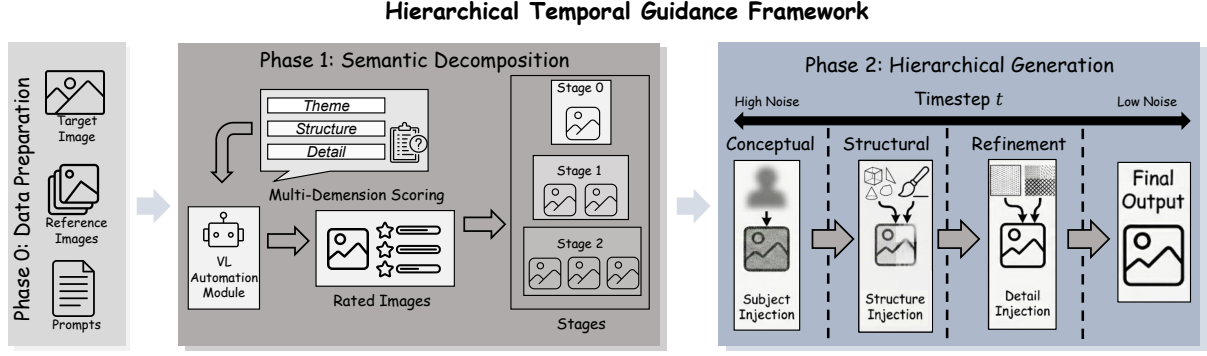


Figure 1: Overall architecture of HTGF. **Phase 0** prepares the target prompt and retrieved reference images. **Phase 1 (Semantic Decomposition)** uses a VL agent to score each reference along Subject, Structure, and Detail and to emit a (role, weight, stage) policy. **Phase 2 (Hierarchical Generation)** executes the denoising trajectory with stage-aligned injection: Conceptual features anchor global semantics at high noise, Structural guidance shapes layout at mid-noise, and Detail refinement enriches local texture at low noise. A manifold-aware corrector smooths the trajectory across stage transitions.

et al., 2024) routes edit operations through a planner. Closer to our setting, TAME (Hong et al., 2026) uses a state-aware MLLM as a personalization assistant, while FuseAgent (Lin et al., 2026b), AgenticRetoucher (Shen et al., 2026b), and Per-Touch (Chang et al., 2026) use VLMs to perceive image-level cues and plan inference-time interventions; as such pipelines increasingly depend on large VLM/LLM backbones, structured pruning can reduce their deployment cost (Ji and Sun, 2026). A parallel branch is retrieval-augmented image generation: FineRAG (Yuan et al., 2025) retrieves fine-grained image fragments at generation time, and ImageRAG (Shalev-Arkushin et al., 2025) retrieves reference images dynamically per prompt. These two strands share an emphasis on *which signals to bring into generation*—which images to retrieve, which operation to plan, which assistant turn to fire—and then hand the resulting tokens or actions to a standard generation path. HTGF operates on an orthogonal axis: given a candidate set of references, it decides *when* along the denoising trajectory each one should act, and is therefore composable with any of the above as front ends.

2.3 Diffusion-Side Mechanisms: Hierarchy, On-Manifold Sampling, and Preference Alignment

A different family of works modifies the diffusion process itself. Semantic-First Diffusion (Pan et al., 2025) asynchronously denoises semantic and texture latents to improve convergence; Hierarchical Koopman Diffusion (Bai et al., 2025a) separates

coarse and fine spectral components in a lifted linear space; Hierarchical Diffusion LMs (Zhou et al., 2025) predict next-scale semantics; Attribute Diffusion (Parihar et al., 2025) performs coarse-to-fine attribute editing; and Semantic-aware Classifier-Free Guidance (Shen et al., 2024a) spatially adapts the guidance scale across semantic regions. None of these methods map heterogeneous *external* references onto the schedule; they reorganize the internal latents. On the trajectory-correction side, TAG (Park et al., 2025), DIAMOND (Polowczyk et al., 2026), and MCLC (Lee et al., 2026) show that guiding on the Tweedie-estimated clean image rather than the noisy latent yields semantically meaningful updates, a principle also exploited by latent-space diffusion bridges for cross-modal translation (Shen et al., 2026a); we adopt this principle to absorb the discrete condition switches introduced by stage-aligned injection. Finally, Diffusion-DPO (Wallace et al., 2024), following the original DPO formulation (Rafailov et al., 2023), and adversarial tuning (Shen et al., 2025) fine-tune the backbone—on preference pairs or under adversarial supervision—so that the model prefers certain outputs over others. This is *training-time* alignment at the level of which images a backbone produces; HTGF is *training-free* and inference-time, operating at the level of how a candidate reference set is decomposed and injected. The two are orthogonal—a DPO-tuned backbone could be slotted into HTGF’s generation phase, and HTGF’s per-reference filtering could feed an offline preference loop—and we keep the explicit

empirical comparison for follow-up work (cf. Limitations).

3 Methodology

Figure 1 gives an overview. After a retrieval module supplies candidate references $\mathcal{R} = \{\mathbf{r}_1, \dots, \mathbf{r}_n\}$ for a target prompt $\mathcal{P}$ (§3.1), a vision-language agent emits a per-reference control policy $\{(\pi_j, w_j)\}$ (§3.2). A latent diffusion model then executes a denoising trajectory in which references are injected in stage-matched windows derived from an SNR-sensitivity argument (§3.3), and a Tweedie-based corrector smooths the trajectory at stage boundaries (§3.4).

3.1 Preliminaries

Problem. Given a textual prompt $\mathcal{P}$ and a retrieved reference set $\mathcal{R} = \{\mathbf{r}_1, \dots, \mathbf{r}_n\}$ reflecting a user’s prior visual interactions, we synthesize an image $\mathbf{x}_0$ that satisfies the prompt and is consistent with the references. Direct fusion of $\mathcal{R}$ is ill-posed because references may agree on subject while disagreeing on layout or texture.

Latent diffusion. We build on Latent Diffusion Models (LDMs). An encoder $\mathcal{E}$ maps an image to a latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$; the forward process corrupts $\mathbf{z}_0$ as $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1 - \alpha_t}\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. A learned $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$ reverses this process under condition $\mathbf{c}$. We turn the condition itself into a function of t : $\mathbf{c}_t$ varies across the denoising trajectory to resolve conflicts in $\mathcal{R}$.

3.2 VLM as Agentic Decomposer

Rather than treat the VLM as a flat similarity scorer, we cast it as an *agent* that produces a structured control policy for the downstream generator. Given a reference $\mathbf{r}_j$ and a core target concept $\mathbf{r}_{core}$ (derived from the user’s high-frequency interactions or session intent), the agent emits a per-axis score vector and an intensity:

$$(\mathbf{S}_j, w_j) = \mathcal{A}_{\text{VLM}}(\mathbf{r}_j, \mathbf{r}_{core}, \mathcal{P}),$$

where $\mathbf{S}_j = [S_{sub}, S_{str}, S_{det}] \in [0, 1]^3$ is the score along three semantic axes (Subject, Structure, Detail) and $w_j \in [0, w_{max}]$ controls overall influence. The routing distribution π_j over the three stages is derived from $\mathbf{S}_j$ below; the hard role K_j is its argmax.

Three-axis decomposition. The roles correspond to the inductive bias of image formation:

- **Subject (S_{sub}):** global atmosphere, color palette, artistic style.
- **Structure (S_{str}):** geometric layout, composition, spatial arrangement.
- **Detail (S_{det}):** high-frequency information, textures, local patterns.

The agent emits a score vector $\mathbf{S}_j = [S_{sub}, S_{str}, S_{det}]$. We derive two quantities from $\mathbf{S}_j$: a per-stage *routing distribution* $\pi_j \in \Delta^2$ that decides *when* the reference contributes, and a scalar *intensity* w_j that decides *how strongly*:

$$\pi_j = \text{softmax}(\beta \mathbf{S}_j), \quad w_j = \phi(\max(\mathbf{S}_j)), \quad (1)$$

with the thresholded linear mapping

$$\phi(s) = \begin{cases} w_{min} + \frac{(s - \tau)(w_{max} - w_{min})}{1 - \tau}, & s > \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

$\beta \in [0, \infty)$ is a routing temperature; τ is a confidence threshold; both are configuration choices, not learned. The threshold τ functions as a language-grounded *filter*: a reference whose maximum aspect score falls below τ has $w_j = 0$ and contributes nothing. We empirically verify in §4 that VL-driven scoring and filtering are central to HTGF’s robustness to noisy histories: removing the VL scorer (Table 3, row w/o (a)) drops CIS from 40.6 to 28.2.

Argmax as a temperature limit. Hard role assignment $K_j = \text{argmax}(\mathbf{S}_j)$ is the $\beta \rightarrow \infty$ limit of Eq. 1; the main tables use this limit and the soft-routing panels (§4.4–§4.6) use $\beta = 4$.

3.3 Hierarchical Multi-Image Guidance (HMIG)

The VLM agent produces *what kind* of signal each reference carries (via π_j). We now address *when* during denoising it is applied. We partition the schedule into three windows

- **Stage 0 (Conceptual, $t \in [T_1, T]$):** $s(t) = 0$
- **Stage 1 (Structural, $t \in [T_2, T_1]$):** $s(t) = 1$
- **Stage 2 (Refinement, $t \in [0, T_2]$):** $s(t) = 2$

At step t in stage $s(t)$, each reference $\mathbf{r}_j$ contributes with effective weight $w_j \cdot \pi_j[s(t)]$, giving the time-aware condition embedding

$$\mathbf{E}_t = \sum_{j=1}^n w_j \cdot \pi_j[s(t)] \cdot \text{Proj}(\text{Enc}(\mathbf{r}_j)) + (1 - \lambda_t) \text{Enc}_{txt}(\mathcal{P}) \quad (3)$$

$\mathbf{E}_t$ is consumed by the U-Net via cross-attention. When $\beta \rightarrow \infty$, $\pi_j[s(t)]$ becomes the indicator $\mathbf{1}[K_j = s(t)]$ and Eq. 3 reduces to the hard schedule under which the main tables were measured.

Sensitivity-based justification. We characterize how strongly an image conditioning vector influences the *predicted clean image* $\hat{\mathbf{x}}_{0,t}$ at each step.

Proposition 1 (SNR-aligned sensitivity). *Let $\hat{\mathbf{x}}_{0,t} = (\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{z}_t, t, \text{cond}))/\sqrt{\bar{\alpha}_t}$ be the Tweedie estimate of the clean image at step t . Then*

$$\frac{\partial \hat{\mathbf{x}}_{0,t}}{\partial \mathbf{e}_{\text{img}}} \propto -\frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \frac{\partial \epsilon_\theta}{\partial \mathbf{e}_{\text{img}}}, \quad (4)$$

and the scaling factor $\sqrt{(1 - \bar{\alpha}_t)/\bar{\alpha}_t}$ is monotonically decreasing in $\text{SNR}(t) = \bar{\alpha}_t/(1 - \bar{\alpha}_t)$.

The proof is a direct application of Tweedie’s formula. The proposition says that an image embedding injected at high-noise steps (low SNR) has globally amplified influence on $\hat{\mathbf{x}}_{0,t}$, and is therefore the natural place to inject Subject-level signals; conversely, injection at low-noise steps (high SNR) can only nudge local high-frequency content, the right slot for Detail-level signals. This produces the schedule (Subject $\rightarrow$ Structure $\rightarrow$ Detail) we use in HMIG. The argument does not depend on the routing being hard: under soft routing (Eq. 1), the per-stage weight $w_j \pi_j[s(t)]$ acts as a smooth analogue of the indicator $\mathbf{1}[K_j = s(t)]$, and the same SNR-scaled sensitivity applies.

3.4 Manifold-Aware Trajectory Correction

Switching $\mathbf{E}_t$ at boundaries T_1, T_2 can push the trajectory off the data manifold, producing visible artifacts (Table 3, row w/o (c)). Conventional guidance minimizes a loss on the noisy latent $\mathbf{z}_t$, whose high-variance noise destabilizes gradient direction. We instead guide on the Tweedie-estimated clean image:

$$\mathcal{L}_{\text{cons}} = \|\Phi(\hat{\mathbf{x}}_{0,t}) - \Phi(\text{SG}[\hat{\mathbf{x}}_{0,t+1}])\|_2^2, \quad (5)$$

where Φ extracts high-level semantic features. The update

$$\mathbf{z}_t^{\text{corr}} = \mathbf{z}_t - \eta_t \nabla_{\mathbf{z}_t} \mathcal{L}_{\text{cons}}(\hat{\mathbf{x}}_{0,t}(\mathbf{z}_t)) \quad (6)$$

backpropagates through the denoiser. Because $\hat{\mathbf{x}}_{0,t}$ lives on (or near) the clean manifold, the gradient direction is semantically meaningful and the correction smooths trajectories across stage transitions without sacrificing prompt adherence.

Algorithm 1 HTGF: Decomposition + HMIG + Manifold Correction

Input: Prompt $\mathcal{P}$, refs $\mathcal{R} = \{\mathbf{r}_j\}_{j=1}^n$, UNet ϵ_θ , VL agent $\mathcal{A}_{\text{VLM}}$, boundaries T_1, T_2 .

Output: Image $\mathbf{x}_0$.

```

1:  $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ;  $\hat{\mathbf{x}}_{\text{anchor}} \leftarrow \emptyset$ 
2: {Phase 1: Decompose}
3: for  $\mathbf{r}_j \in \mathcal{R}$  do
4:    $\mathbf{S}_j \leftarrow \mathcal{A}_{\text{VLM}}(\mathbf{r}_j, \mathbf{r}_{\text{core}}, \mathcal{P})$ 
5:    $\pi_j \leftarrow \text{softmax}(\beta \mathbf{S}_j)$ ;  $w_j \leftarrow \phi(\max \mathbf{S}_j)$ 
6: end for
7: {Phase 2: Hierarchical generation}
8: for  $t = T, \dots, 1$  do
9:    $s \leftarrow s(t) \in \{0, 1, 2\}$  {stage window of  $t$ }
10:   $\mathbf{E}_t \leftarrow \text{Enc}_{\text{txt}}(\mathcal{P})$ 
11:  for  $\mathbf{r}_j \in \mathcal{R}$  do
12:     $\mathbf{E}_t \leftarrow \mathbf{E}_t + w_j \cdot \pi_j[s] \cdot \text{Proj}(\text{Enc}(\mathbf{r}_j))$ 
13:  end for
14:   $\epsilon_t \leftarrow \epsilon_\theta(\mathbf{z}_t, t, \mathbf{E}_t)$ 
15:   $\hat{\mathbf{x}}_{0,t} \leftarrow (\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_t)/\sqrt{\bar{\alpha}_t}$ 
16:  if  $\hat{\mathbf{x}}_{\text{anchor}} \neq \emptyset$  then
17:     $\mathcal{L}_{\text{cons}} \leftarrow \|\Phi(\hat{\mathbf{x}}_{0,t}) - \Phi(\hat{\mathbf{x}}_{\text{anchor}})\|_2^2$ 
18:     $\mathbf{z}_t \leftarrow \mathbf{z}_t - \eta_t \nabla_{\mathbf{z}_t} \mathcal{L}_{\text{cons}}$ 
19:    recompute  $\epsilon_t$  from corrected  $\mathbf{z}_t$ 
20:  end if
21:   $\hat{\mathbf{x}}_{\text{anchor}} \leftarrow \text{SG}[\hat{\mathbf{x}}_{0,t}]$ 
22:   $\mathbf{z}_{t-1} \leftarrow \text{Scheduler}(\mathbf{z}_t, \epsilon_t, t)$ 
23: end for
24: return  $\mathcal{D}(\mathbf{z}_0)$ 
```

3.5 Design Rationale

Why three stages. The SNR-sensitivity argument in Proposition 1 gives a *continuous* prefactor $\sqrt{(1 - \bar{\alpha}_t)/\bar{\alpha}_t}$, not a discrete three-way partition. In principle one could partition the schedule into any number of windows. We pick three because the SNR gradient has three regimes that admit clean semantic targets: coarse semantics (high noise, where the prefactor is large enough to reshape silhouette), spatial layout (intermediate noise, where the model is laying down composition), and fine texture (low noise, where the model fixes high-frequency detail). A finer partition (e.g., five stages) introduces additional boundaries, each of which is a manifold-correction interface costing roughly 0.23 s per affected step (Table 4), without unlocking a new semantic axis that the VL agent can score along. A coarser partition (two stages) collapses Structure into either Subject (losing geometric control of layout) or Detail (letting texture bleed into com-

position, which is exactly the failure pattern we ablate against in Table 3 row w/o (c)). Three is the smallest partition that aligns to the SNR gradient and to a separable semantic axis set.

Why these stage boundaries. The boundaries are $T_1 = 47, T_2 = 41$ in DDIM-50 step indices, equivalently $(T_1, T_2) \approx (800, 500)$ in DDPM-1000 timestep units. Two pieces of evidence drive the choice. First, the sensitivity sweep in §4.5 shows that CIS stays within a $\sim 4\%$ band over a ± 200 -step swept window on each boundary, so the exact placement is not load-bearing. Second, within that band we picked $(800, 500)$ for two reasons: (a) under SDXL’s noise schedule the three windows $[T_1, T]$, $[T_2, T_1]$, $[0, T_2]$ have roughly equal lengths in number of denoising steps (each about a third of DDIM-50), so each window gets comparable optimization budget; and (b) $t \approx 500$ is empirically the regime where SDXL begins to lay down high-frequency texture (this is consistent with the general SNR-aligned literature, e.g., the same regime that progressive distillation methods identify as the “detail” window). Placing T_2 at 500 therefore aligns Stage 2’s window with the model’s intrinsic high-frequency phase. The robustness band from the sensitivity sweep means a different practitioner could pick $(750, 450)$ or $(850, 550)$ and obtain essentially the same numbers.

4 Experiments

4.1 Experimental Setup

We implement HTGF in PyTorch. The main tables (Tables 1, 2, and 3) were measured on $4 \times$ NVIDIA A100 GPUs with $N = 5$ retrieved references per prompt; the efficiency, sensitivity, cold-start, and automated PF/TA panels (§4.4–§4.6) were measured on a single NVIDIA RTX A6000 with $N = 4$, following a compute-resource change.² Generation uses SDXL (Podell et al., 2024) (fp16) at 1024×1024 with the DDIM sampler, 50 inference steps, and CFG scale 7.5. The VL agent is Qwen3-VL-8B (Bai et al., 2025b).

Datasets. We evaluate on three datasets spanning different domains and noise conditions: (1) **Movie Poster**: high-quality movie advertisements requiring strong semantic grounding. (2) **Sticker**: styl-

ized character generation with diverse artistic attributes. (3) **Mixed (Noisy)**: constructed by mixing primary user preferences with irrelevant or conflicting noise images to simulate sparse interaction. Construction details in Appendix.

Baselines. We compare against three categories: (1) DM-based: Textual Inversion (TI); (2) LLM-based: PMG; (3) LMM-based: LLaVA (Liu et al., 2023) and LaVIT (Jin et al., 2024a); and the standard SDXL-IPAdapter as a general image-conditioned diffusion baseline.

Evaluation Metrics. We use CLIP Score (CS), CLIP Image Similarity (CIS) (Radford et al., 2021), LPIPS (Zhang et al., 2018), MS-SSIM (Wang et al., 2003), FID, and a weighted Overall score.

4.2 Quantitative Results

Performance on Movie Poster and Sticker. HTGF consistently outperforms baselines in personalization metrics (Table 1 for Sticker). On Sticker CIS **70.25** vs. Pigeon 67.65, with competitive MS-SSIM (**0.1504**) and LPIPS (**0.7252**) confirming that hierarchical injection preserves structural integrity. On Movie Poster CIS **42.20** vs. LaVIT 30.49.

Robustness to Noisy Inputs. Table 2 demonstrates superior robustness on Mixed (Noisy). While baselines suffer severe degradation (SDXL-IPAdapter FID 92.87, TI FID 99.13), HTGF maintains FID **38.87** and the highest CIS **40.62**, confirming that VL-driven scoring effectively filters irrelevant noise.

Comparison with LMM-based Methods. LMM-based baselines like LLaVA and LaVIT often struggle with fine-grained semantic alignment. HTGF achieves Semantic CIS of **52.35** on Movie Poster vs. LaVIT 46.02 and LLaVA 48.50. By decoupling theme, structure, and detail, the model avoids “semantic vanishing” common in long-sequence multimodal tokens.

DRC (Xu et al., 2025b) reports on the same SER30K and ML-Latest data; their published numbers (Table 1 for Sticker) trail HTGF by 5.2 Personalization CIS on Sticker and 4.7 on Movie Poster. DRC also disentangles user history but injects uniformly across steps; the gap is consistent with Eq. 3+Proposition 1.

²The two cards are same-generation Ampere; the change does not affect the architectural or algorithmic choices reported below, only the absolute wall-clock numbers in Table 4.

Table 1: Quantitative comparison on **Sticker**. DRC[†]: transcribed from Xu et al. (2025b) Table 2 (different test split; see Limitations).

Sticker Methods		Overall↓	CS↑	CIS↑	Personalization		MS-SSIM↑	Semantic Alignment			Fidelity FID↓
				DIS↑	LPIPS↓			CS↑	CIS↑	DIS↑	
	SDXL-IPAdapter	5.22	15.35	40.87	19.62	0.7280	0.0432	28.14	43.58	39.87	<u>71.87</u>
DM-based	TI	4.33	18.67	40.90	36.58	0.7654	0.0887	32.91	53.67	48.50	105.48
LLM-based	PMG	5.00	19.16	47.34	39.15	0.7383	0.0827	18.31	45.45	37.80	84.91
LMM-based	LLaVA	5.22	18.72	37.44	33.19	0.7552	0.0851	27.02	49.15	43.88	95.19
	LaVIT	3.78	16.39	40.56	40.84	0.7377	0.1128	25.74	<u>70.80</u>	<u>69.93</u>	83.39
	Pigeon	<u>3.00</u>	23.69	<u>67.65</u>	<u>62.23</u>	0.6814	0.1568	21.10	47.44	45.44	89.43
	DRC [†]	—	23.19	65.10	60.00	0.6770	—	21.88	48.30	45.79	92.31
Ours		1.44	<u>19.42</u>	70.25	72.61	<u>0.7252</u>	<u>0.1504</u>	<u>29.11</u>	77.10	70.85	64.81

Table 2: Quantitative comparison on **Mixed (Noisy)**.

Mixed (Noisy) Methods		Overall↓	CS↑	CIS↑	Personalization		MS-SSIM↑	Semantic Alignment			Fidelity FID↓
				DIS↑	LPIPS↓			CS↑	CIS↑	DIS↑	
	SDXL-IPAdapter	5.00	11.92	<u>32.67</u>	16.43	0.8127	0.0487	21.62	33.43	31.28	92.87
DM-based	TI	5.44	10.23	24.12	15.34	0.8362	0.0763	27.18	35.23	32.87	99.13
LLM-based	PMG	5.67	11.12	21.34	<u>18.23</u>	0.8173	0.0808	12.43	22.87	20.32	98.86
LMM-based	LLaVA	4.22	11.08	26.18	16.62	0.8134	0.0898	25.23	39.12	34.12	75.43
	LaVIT	<u>2.22</u>	12.18	26.67	17.43	<u>0.7843</u>	<u>0.0918</u>	<u>27.62</u>	<u>40.23</u>	48.34	<u>50.12</u>
	Pigeon	3.78	<u>13.44</u>	22.57	15.20	<u>0.8086</u>	0.1022	26.91	33.64	38.21	63.75
Ours		1.67	15.23	40.62	21.18	0.7392	0.0911	28.32	49.62	<u>45.43</u>	38.87

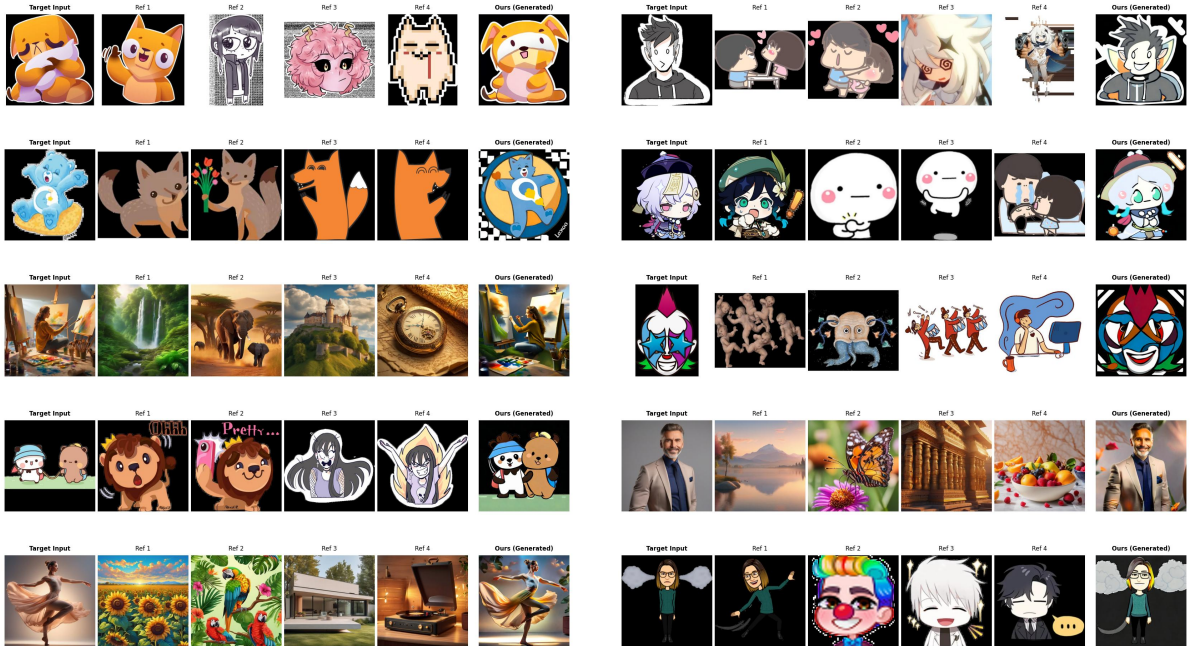


Figure 2: Qualitative cases generated by HTGF on the Mixed (Noisy) split. Each block shows the target prompt, four retrieved references (with off-topic distractors), and the HTGF output. Identity, structure, and stylistic detail track the on-topic references while distractors are filtered.

4.3 Ablation Study

We ablate the three core components on the Mixed (Noisy) dataset (Table 3).

Impact of VL-Driven Scoring (w/o VL). Replacing the VL scorer with Uniform Weighting drops CIS from **40.62** to **28.15** and inflates FID to

Table 3: Ablation of key components on Mixed (Noisy). (a) VL Scorer; (b) Hierarchical Injection; (c) Manifold Correction.

Method	Personalization		Structure & Percept.		Fidelity
	CIS $\uparrow$	CS $\uparrow$	MS-SSIM $\uparrow$	LPIPS $\downarrow$	
HTGF	40.62	15.23	0.739	0.0911	38.87
w/o (a)	28.15	12.04	0.612	0.1205	76.42
w/o (b)	36.40	14.10	0.655	0.1028	49.30
w/o (c)	39.85	15.01	0.728	0.0974	42.15

76.42. Without explicit VL-derived attention, the model cannot suppress irrelevant signals.

Efficacy of Hierarchical Injection (w/o Hierarchy). Replacing temporal alignment with Flat Injection produces “semantic collision”: MS-SSIM degrades from **0.739** to **0.655** and LPIPS rises from **0.0911** to **0.1028**.

Efficacy of Manifold Correction (w/o Manifold). Disabling the gradient-based manifold corrector (row w/o (c)) degrades every metric: FID rises from **38.87** to **42.15** (+8.5%), CIS drops from **40.62** to **39.85**, MS-SSIM from **0.739** to **0.728**, and LPIPS from **0.0911** to **0.0974**. The effect is smaller than the VL-filter ablation (a) or the hierarchy ablation (b)—which is expected, since the corrector is a refinement on top of an already-decomposed schedule—but it is non-negligible and concentrated at the stage boundaries T_1, T_2 . We inspected the predicted clean image $\hat{x}_{0,t}$ across the trajectory and found that without the correction, condition switching at $t \in \{T_1, T_2\}$ produces visible “jitter” (high-frequency reorganization of layout or texture between consecutive steps); with the correction, $\Phi(\hat{x}_{0,t})$ varies smoothly across boundaries. This is consistent with the design argument in §3.4: the gradient $\nabla_{\mathbf{z}_t} \mathcal{L}_{cons}$ taken on the clean-manifold estimate gives a semantically meaningful update direction, whereas the same loss measured on the noisy latent would be dominated by sampling noise.

4.4 Efficiency Analysis

Table 4 reports per-component cost on a single NVIDIA RTX A6000. Three observations follow.

HMIG is essentially free. Stage-aligned injection only changes which pre-encoded reference embedding is added to the conditioning vector at each step; the UNet forward cost is identical to plain DDIM-50 (9.37 s vs. 9.22 s, within measurement noise). VRAM is unchanged.

Table 4: Efficiency on a single NVIDIA RTX A6000 (48 GB). Latencies are medians over 3 measured runs after 1 warmup. Manifold correction is the per-step cost of one UNet forward + backward pass to compute $\nabla_{\mathbf{z}_t} \mathcal{L}_{cons}$; the end-to-end row assumes 10 active correction steps. HMIG overhead is negligible because stage-aligned injection only changes which pre-encoded reference embedding is added to the conditioning at each step.

Component	Latency (s)	VRAM (GB)	Throughput (img/min)
VLM scoring (Qwen3-VL-8B, $N=4$)	4.59	16.6	13.1
SDXL DDIM-50 (text-only baseline)	9.22	6.9	6.5
SDXL DDIM-50 + HMIG	9.37	6.9	6.4
Manifold correction (per step)	0.23	11.6	—
End-to-end HTGF (10 corr. steps)	16.11	16.6	3.7

Manifold correction is a controllable cost. Each correction step requires one UNet forward and one backward pass to evaluate $\nabla_{\mathbf{z}_t} \mathcal{L}_{cons}$, taking 0.23 s per step and peaking at 11.6 GB VRAM. Activating the corrector on 10 steps (typical for our schedule) adds roughly 2.3 s to the diffusion phase—a 25% overhead for the quality gain reported in Table 3 (row w/o (c)). When latency dominates, the corrector can be disabled or restricted to stage-transition steps only.

VLM scoring dominates the per-image budget. The Qwen3-VL-8B scoring pass over $N=4$ references is the single largest contributor (4.59 s) and pushes peak VRAM to 16.6 GB. Batching across many target prompts amortizes this cost; in a serving pipeline with persistent VLM context, the per-prompt cost drops to roughly the rate of producing the new (role, weight) tuples.

4.5 Sensitivity to Stage Boundaries

Stage boundaries T_1, T_2 are set empirically. To check whether HTGF is fragile to the exact choice we sweep each boundary in raw DDPM-1000 timestep units around the operating point ($T_1=800, T_2=500$): $T_1 \in \{600, 700, 800, 900, 1000\}$ with $T_2=500$ fixed, and $T_2 \in \{300, 400, 500, 600, 700\}$ with $T_1=800$ fixed. For each setting we run a held-out set of 3 prompts under the soft-routing instantiation ($\beta=4$) and report CIS. Across the ± 200 -step window CIS varies between 0.762 and 0.795 (a $\sim 4\%$ band), with a mild trend toward higher CIS for shorter Stage-0 windows (smaller T_1). The schedule is stable across the swept range and does not require fine-grained boundary tuning.

Table 5: User study results. Human Mean Opinion Scores (1–5) on Mixed (Noisy), 50 cases. PF: Personalization Fidelity; TA: Text-Image Alignment; VQ: Visual Quality.

Method	PF	TA	VQ	Avg
SDXL-IPAdapter	2.62	2.44	1.98	2.35
TI	2.55	3.67	2.30	2.84
PMG	2.30	3.10	2.56	2.65
LaVIT	3.02	3.37	2.81	3.07
Ours	3.81	3.92	3.37	3.70

4.6 Cold-Start Behavior

We measure HTGF under shrinking history. Holding the prompt and reference set fixed, we run with $N \in \{4, 2, 1, 0\}$ active references under the soft-routing instantiation ($\beta=4$) and report CIS against the same 4-image reference set. CIS degrades monotonically with shrinking history—0.706 $\rightarrow$ 0.679 $\rightarrow$ 0.645 $\rightarrow$ 0.595—with the $N=0$ row reducing to text-only SDXL (IP-Adapter scales force-zeroed). The framework degrades smoothly: no artifacts, no special-casing, just a gradual loss of personalization signal.

4.7 User Study

We complement the metrics with a perceptual study on 50 cases sampled from Mixed (Noisy). Evaluators saw the interaction history, the target prompt, and anonymized outputs from HTGF and four strong baselines under a double-blind randomized protocol, and rated PF, TA, VQ on a 5-point Likert scale. HTGF achieves the highest scores across all dimensions, with PF **3.81** vs LaVIT 3.02 and IPAdapter 2.62 (Table 5). A blind side-by-side preference test on the same pool selects HTGF as the preferred output in 58%/42%/72% of cases for PF/TA/VQ; a CLIP-based two-axis automated panel (PF/TA) corroborates these results.

4.8 Qualitative Analysis

Figure 2 shows HTGF outputs on the Mixed (Noisy) split across heterogeneous reference histories (posters, stickers, anime portraits). Each row typically contains 1–2 on-topic references and 2–3 off-topic distractors; the VL soft router suppresses the latter, and the SNR-keyed schedule routes the surviving references through the stage that matches their semantic level (Subject, Structure, Detail). The result is a stable identity lock



Figure 3: Single-case zoom. HTGF filters off-topic references (Refs 2–4) and locks identity to the on-topic Ref 1.

without leakage of palette or style fragments from off-topic refs. Figure 3 zooms in on one row of the grid where Refs 2–4 are visually unrelated cartoon figures and only Ref 1 shares the target’s flat “sticker” family. The VL-driven score vectors place Ref 1 on the Subject axis with $w_1 \approx 1.1$, while the three off-topic refs are bound to weaker slots with $w_j \in [0.5, 0.7]$. The HTGF output preserves the on-topic identity traits (hair ornament, kimono) and the target’s flat sticker style without leakage of palette or fur texture from the distractors.

5 Conclusion

HTGF couples semantic decomposition with an SNR-derived temporal schedule (Proposition 1) and a stage-boundary corrector. The training-free pipeline beats strong baselines on three datasets and degrades gracefully in the cold-start limit.

Limitations

HTGF has known boundaries. (i) At the extremes of reference disagreement—conflicting subjects, contradictory poses, or aggressive detail collisions—the stage-aligned schedule degrades in principled but visible ways. (ii) The (π_j, w_j) policy is only as expressive as the VLM that emits it; abstract semantic dimensions (tone, narrative, cultural cues) can be misread, and a stronger VL agent would directly improve the pipeline. (iii) The framework assumes a retrieval module supplies a meaningful candidate set; the Mixed (Noisy) experiments probe partial robustness but adversarial reference injection is untested.

Ethics Statement

Data. The Movie Poster and Sticker datasets are aggregated from publicly available web sources for research purposes. The Mixed (Noisy) dataset is synthetically constructed by mixing primary preferences with irrelevant references. No personally identifiable user data is collected, and no demographic information about the user-study evaluators is recorded beyond informed consent.

User study. Human evaluators in the perceptual study (§4) participated voluntarily under a double-blind protocol. Image order was randomized per query to eliminate position bias; no compensation was tied to specific judgments. The LMM-as-judge protocol replaces the bulk of human effort with an automated judge, and we report inter-rater agreement with the human study to validate it.

Compute. The main three-dataset experiments were conducted on 4 NVIDIA A100 GPUs; the efficiency, sensitivity, cold-start, and automated PF/TA panels were measured on a single NVIDIA RTX A6000 after a compute-resource change. Per-component latency and peak VRAM measurements are reported in §4.4.

Models. HTGF relies on the pre-trained SDXL backbone (Podell et al., 2024) and Qwen3-VL-8B (Bai et al., 2025b), both publicly released under their respective licenses. We do not fine-tune either model. Any deployment must respect the licenses of these upstream artifacts.

Risks and mitigations. Personalized image generation can be misused to impersonate identities or to reproduce copyrighted styles. We release this work only for research use. For downstream deployment we recommend provenance signals (e.g., C2PA-style watermarks), explicit consent for any identity references, and refusal-by-construction for prompts that name living individuals.

Reproducibility. The HTGF inference code, dataset manifests, and reference VL prompt templates are available at <https://github.com/LjdLi/Implementation-of-HMGF>.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (No. 62406141, 92370204), the National Key Research and Development Program of China (No. 2023YFF0725001), the Innovation Research Foundation of National University of Defense Technology (Grant No. 26-ZZCX-JDZ-56), and the China Postdoctoral Science Foundation (No. 2025M784335 and GZC20252640).

References

Hanru Bai, Weiyang Ding, and Difan Zou. 2025a. [Hierarchical Koopman diffusion: Fast generation with in-](#)

[terpretable diffusion trajectory](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Junyang Lin, Peng Wang, An Yang, Jianxin Yang, Fan Zhou, Jingren Zhou, and Ke Zhu. 2025b. [Qwen3-VL technical report](#). *arXiv preprint arXiv:2511.21631*.

Zewei Chang, Zheng-Peng Duan, Jianxing Zhang, Chun-Le Guo, Siyu Liu, Hyungju Chun, Hyunhee Park, Zikun Liu, and Chongyi Li. 2026. [PerTouch: VLM-driven agent for personalized and semantic image retouching](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. 2025. [XVerse: Consistent multi-subject control of identity and semantic attributes via DiT modulation](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2024. [Guiding instruction-based image editing via multimodal large language models](#). In *ICLR*. OpenReview.net.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. [An image is worth one word: Personalizing text-to-image generation using textual inversion](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.

Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, Li Chen, Ankit Jain, Ning Zhang, Peizhao Zhang, Roshan Sumbaly, Peter Vajda, and Animesh Sinha. 2024. [Imagine yourself: Tuning-free personalized image generation](#). *arXiv preprint arXiv:2409.13346*.

Rongpei Hong, Jian Lang, Ting Zhong, Yong Wang, and Fan Zhou. 2026. [TAMEing long contexts in personalization: Towards training-free and state-aware MLLM personalized assistant](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

Hexiang Hu, Kelvin C. K. Chan, Yu-Chuan Su, Wenhui Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, Ming-Wei Chang, and Xuhui Jia. 2024. [Instruct-imagen: Image generation with multi-modal instruction](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4754–4763.

Yang Ji and Ying Sun. 2026. [OCP: Outlier-centric probing for dynamic structured pruning of LLMs](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume*

- I: Long Papers*), pages 6108–6124, San Diego, California, United States. Association for Computational Linguistics.
- Yang Ji, Ying Sun, Yuting Zhang, Zhigaoyuan Wang, Yuanxin Zhuang, Zheng Gong, Dazhong Shen, Chuan Qin, Hengshu Zhu, and Hui Xiong. 2025. A comprehensive survey on self-interpretable neural networks. *Proceedings of the IEEE*.
- Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chengru Song, Dai Meng, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. 2024a. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *ICLR*. OpenReview.net.
- Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chengru Song, Dai Meng, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. 2024b. [Unified language-vision pretraining in LLM with dynamic discrete visual tokenization](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Jing Yu Koh, Daniel Fried, and Russ R. Salakhutdinov. 2023. [Generating images with multimodal language models](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. [Multi-concept customization of text-to-image diffusion](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hyoseok Lee, Sohwi Lim, Eunju Cha, and Tae-Hyun Oh. 2026. [Measurement-consistent Langevin corrector for stabilizing latent diffusion inverse problem solvers](#). *arXiv preprint arXiv:2601.04791*.
- Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024. [PhotoMaker: Customizing realistic human photos via stacked ID embedding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jianghao Lin, Peng Du, Jiaqi Liu, Weite Li, Yong Yu, Weinan Zhang, and Yang Cao. 2026a. [Sell it before you make it: Revolutionizing e-commerce with personalized AI-generated items](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- ZiXu Lin, Yunlong Lin, Kunjie Lin, Kepeng Yang, Ge Meng, Yingying Wang, Yue Huang, and Xinghao Ding. 2026b. [FuseAgent: A VLM-driven agent for unified in-the-wild image fusion](#). OpenReview. OpenReview forum FUioZXchu5.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 34892–34916.
- Yueming Pan, Ruoyu Feng, Qi Dai, Yuqi Wang, Wenfeng Lin, Mingyu Guo, Chong Luo, and Nanning Zheng. 2025. [Semantics lead the way: Harmonizing semantic and texture modeling with asynchronous latent diffusion](#). *arXiv preprint arXiv:2512.04926*.
- Rishabh Parihar, Prasanna Balaji, Raghav Magazine, Sarthak Vora, Varun Jampani, and R. Venkatesh Babu. 2025. [Attribute diffusion: Diffusion driven diverse attribute editing](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- Rishabh Parihar, V. S. Sachidanand, Sabariswaran Mani, Tejan Karmali, and R. Venkatesh Babu. 2024. [PreciseControl: Enhancing text-to-image diffusion models with fine-grained attribute control](#). In *European Conference on Computer Vision (ECCV)*.
- Youngrok Park, Hojung Jung, Sangmin Bae, and Se-Young Yun. 2025. [Temporal alignment guidance: On-manifold sampling in diffusion models](#). *arXiv preprint arXiv:2510.11057*.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. [SDXL: Improving latent diffusion models for high-resolution image synthesis](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Alicja Polowczyk, Agnieszka Polowczyk, Piotr Borycki, Joanna Waczyńska, Jacek Tabor, and Przemysław Spurek. 2026. [DIAMOND: Directed inference for artifact mitigation in flow matching models](#). *arXiv preprint arXiv:2602.00883*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 53728–53741.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. [DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510.
- Rotem Shalev-Arkushin, Rinon Gal, Amit H. Bermano, and Ohad Fried. 2025. [ImageRAG: Dynamic image retrieval for reference-guided image generation](#). *arXiv preprint arXiv:2502.09411*.

- Dazhong Shen, Jingjing Gu, Qiang Zhou, Meng Zhao, and Ying Sun. 2026a. LiDAR-to-4DRadar diffusion bridge via cross-modal alignment and translation in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17110–17120.
- Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. 2024a. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9370–9379.
- Dazhong Shen and 1 others. 2025. ADT: Tuning diffusion models with adversarial supervision. *arXiv preprint arXiv:2504.11423*.
- Shaocheng Shen, Jianfeng Liang, Chunlei Cai, Cong Geng, Huiyu Duan, Xiaoyun Zhang, Qiang Hu, and Guangtao Zhai. 2026b. [Agentic retoucher for text-to-image generation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024b. [Pmg: Personalized multimodal generation with large language models](#). In *Proceedings of the ACM Web Conference (WWW)*, pages 3833–3843.
- Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. [Instantbooth: Personalized text-to-image generation without test-time finetuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8543–8552.
- Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. [Training-free consistent text-to-image generation](#). *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 43(4).
- Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. [Diffusion model alignment using direct preference optimization](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8238.
- Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. 2024. [InstantID: Zero-shot identity-preserving generation in seconds](#). *arXiv preprint arXiv:2401.07519*.
- Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. 2025. [MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. [Multiscale structural similarity for image quality assessment](#). In *ACSSC*, pages 1398–1402. IEEE.
- Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. 2025. [OmniGen: Unified image generation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. 2025a. [Personalized image generation with large multimodal models](#). In *Proceedings of the ACM Web Conference (WWW)*.
- Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. 2025b. [DRC: Enhancing personalized image generation via disentangled representation composition](#). In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*.
- Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. 2023. [IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models](#). *arXiv preprint arXiv:2308.06721*.
- Huaying Yuan, Ziliang Zhao, Shuting Wang, Shitao Xiao, Minheng Ni, Zheng Liu, and Zhicheng Dou. 2025. [FineRAG: Fine-grained retrieval-augmented text-to-image generation](#). In *Proceedings of COLING*.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. [The unreasonable effectiveness of deep features as a perceptual metric](#). In *CVPR*, pages 586–595. IEEE.
- Cai Zhou, Chenyu Wang, Dinghuai Zhang, Shangyuan Tong, Yifei Wang, Stephen Bates, and Tommi Jaakkola. 2025. [Next semantic scale prediction via hierarchical diffusion language models](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. 2025. [EasyRef: Omni-generalized group image reference for diffusion models via multimodal LLM](#). In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.