

Differentiable Geometric Alignment for LLM Spatial Reasoning: Training-Time Symbolic Constraints, Solver-Free Inference

Yuhao Liao^{1*}, Zijian Zeng¹, Huiming Yang², Fei Ding³

¹ UCSI University ² Beijing Institute of Technology

³ Nanchang University

* Corresponding author: lukerroyal@outlook.com

Abstract

Large language models struggle with spatial and geometric reasoning. Existing remedies either fine-tune on answers, leaving geometric structure implicit, or attach a symbolic solver at inference, which is brittle under imperfect parser coverage. We propose **DGCL** (*Differentiable Geometric Constraint Layer*), a training-time objective for a typed family of spatial and geometric relations (e.g., parallelism, frame-relative left/right, vertical order) that supervises the LLM’s hidden states with a differentiable Huber residual over parser-proposed candidate constraints, a coverage regularizer, and a learned selector. Training requires a frozen constraint parser; inference does not: it is a single LLM forward pass over the text prompt plus one learned [GEO] token, with no parser, solver, external tool, or auxiliary head. We evaluate on three auditable datasets (planar geometry, language-grounded 3D relations, and a firewalled zero-touch counterfactual split) with 10,278 test instances over 5 seeds. We are explicit about attribution: a strong constraint-supervised control (A2c), which trains the selector and coverage objective on gold candidates but blocks the residual gradient from the LLM, already captures most of the improvement over plain fine-tuning (up to +26 points). Our isolated contribution is that *coupling* the differentiable residual gradient into the LLM adds a *further, stable* +4.0 to +5.2 points over this strong control across all five settings (seven pre-registered gates, Holm-corrected $p < 0.001$), at $1.06\times$ a fine-tuned baseline’s latency. On full parser coverage a strong tool baseline marginally exceeds DGCL; DGCL’s advantage is robustness under partial or missing coverage, not superiority over oracle solvers.

1 Introduction

Geometric reasoning in natural language — determining whether two segments are parallel, what frame a left/right relation refers to, or which solid

extends above another — is a stubborn weakness of modern large language models (Shiri et al., 2024; Liu et al., 2023; Liao et al., 2024; Wasi et al., 2026). The community has converged on two broad responses. The first is to fine-tune on more answers or chains-of-thought (Wei et al., 2022; Lightman et al., 2024), which moves the average but leaves the geometric structure implicit in token statistics. The second is to bolt a symbolic solver onto the inference path (Trinh et al., 2024; Chervonyi et al., 2025; Ping et al., 2026; Gao et al., 2023; Chen et al., 2023; Schick et al., 2023), which can be highly accurate when a clean formalization is produced but tends to be brittle under imperfect parser coverage and slow under realistic candidate budgets.

The gap we target: existing methods place symbolic geometric structure either nowhere in the model (answer-only fine-tuning) or entirely in the inference path (tool pipelines), but not in the *gradient*. No prior work, to our knowledge, supervises an LLM’s hidden states with a differentiable geometric-constraint residual and then removes that machinery at test time, so whether geometric structure helps as a training signal — separately from its value as a runtime tool — has not been isolated. This matters in the common regime where a parser exists but is imperfect: there, the inference-time-symbolic route degrades with the parser, while a training-time signal need not.

A third option remains under-explored: *use symbolic geometric structure as a training signal that reshapes the language model’s internal representation of spatial text, then discard the structure at inference*. The target of the intervention is representational — we want the LLM to encode spatial language (parallelism, frame-relative left/right, vertical order) in a way that respects geometric consistency — not a bolted-on geometry module. Concretely, we keep a frozen parser only during training to extract typed candidate constraints; we score those candidates with differentiable residuals;

we propagate the loss into the language model’s hidden states through a small variable predictor; and at test time we run a standard LLM forward pass on the text prompt (plus one learned [GEO] token), with no parser, solver, or external tool. We call this **DGCL**, for *Differentiable Geometric Constraint Layer*. DGCL is therefore parser-free at inference but parser-dependent in training (§6.6).

Three observations motivate the design: a constraint’s geometric content (“ $AB \parallel CD$ ”, “ u is left of v under frame f ”) has a natural differentiable surrogate that an answer-only loss lacks; which candidates to attend to is learnable via a softmax selector with coverage and selector losses; and keeping the signal out of the inference path avoids the latency, timeout, and formatting failures of tool baselines.

We test three falsifiable claims. **(I) Differentiable coupling adds a stable gain over strong constraint supervision:** backpropagating the residual into the LLM hidden states improves over the constraint-supervised but detached control A2c (falsified if A2c matches A6). **(II) The gain is geometry-specific,** not extra parameters or noise (falsified if RandGeo or ParamMatch closes the gap). **(III) Inference is non-oracle and solver-free** (falsified if A6 calls a parser/solver or its latency exceeds $1.5 \times B2 \text{ FT}$).

Of the ~ 20 – 26 point gain of the full model (A6) over plain fine-tuning (B2 FT), most comes from *constraint-supervised training* (gold-candidate matching plus the q_{ij} -supervised selector/coverage), already present in the stop-gradient control A2c (e.g., A2c 82.1 vs. B2 FT 57.6 on D1 public; Table 14). Only the remaining $+4.0$ to $+5.2$ points come from the genuinely new piece: *coupling* the differentiable residual gradient into the LLM (A6 over A2c). We therefore claim only that this coupling yields a small yet stable additional gain over a strong constraint-supervised control — positive in all five settings, 5/5 seeds, $p < 0.001$ — which prior answer-only and tool-pipeline methods do not isolate.

We instantiate DGCL on Llama-3.1-8B (Meta AI, 2024) with LoRA (Hu et al., 2022) and a frozen DeBERTa-v3 parser (He et al., 2023). Our contributions are: (1) a training-time differentiable alignment objective that propagates typed geometric residuals into LLM hidden states through a small variable predictor, learned selector, and coverage regularizer, with no inference-time machinery beyond one learned prompt token, scoped so that the isolated effect is the stable gain of differ-

entiable coupling *over the detached control A2c*; (2) an auditable protocol over planar geometry (D1), language-grounded 3D relations (D2, True-Text-Only and Mixed), and a frozen zero-touch counterfactual split (D3), 10,278 test items in all, with seven pre-registered Holm–Bonferroni gates (Holm, 1979); (3) coupling and specificity ablations (A6 vs. A2c, RandGeo, ParamMatch); and (4) a boundary analysis showing that a strong non-oracle tool baseline marginally exceeds DGCL under full parser coverage, while DGCL’s advantage is robustness under partial/missing coverage, below oracle solver upper bounds.

2 Related Work

Spatial reasoning and geometric solving. Benchmarks show LMs/VLMs are weak at spatial reasoning, and explicit structure (boxes, scene graphs) helps more than chain-of-thought (Shiri et al., 2024; Li et al., 2024; Liao et al., 2024; Wasi et al., 2026; Liu et al., 2023); a related line uses symbolic/visualized intermediate representations at prompting time (Hu et al., 2024; Wu et al., 2024). On the symbolic side, AlphaGeometry(2) (Trinh et al., 2024; Chervonyi et al., 2025), AutoGPS (Ping et al., 2026), Inter-GPS (Lu et al., 2021), and GeoQA (Chen et al., 2021) pair neural models with deductive engines that are oracle-like when formalization succeeds; we calibrate against such solvers as upper bounds, not competitors (Table 9). For 3D grounding, SpatialVLM (Chen et al., 2024), SpatialRGPT (Cheng et al., 2024), and NUMINA (Zeng et al., 2025) scale grounded data; our D2 uses ScanNet (Dai et al., 2017)/3DSSG (Wald et al., 2020) scenes, and our intervention is a loss, not a data-source change.

Neuro-symbolic learning with differentiable structure. Work that makes symbolic structure differentiable splits on what happens at inference. Some methods keep a symbolic component *at test time*: SATNet (Wang et al., 2019) runs a smoothed MAXSAT layer in the forward pass, and DeepProbLog (Manhaeve et al., 2018) and NeurASP (Yang et al., 2020) query a probabilistic logic program. Others use the constraint *only as a training signal* and add nothing at inference — the semantic loss (Xu et al., 2018) and Logic Tensor Networks (Badreddine et al., 2022) regularize the network during training, after which it runs unchanged. DGCL is in this second, training-time-only family, but is aimed at *language models for textual spatial*

reasoning and differs in three ways: the supervised object is a small library of *continuous typed geometric residuals* rather than Boolean logic; the gradient is routed, via a learned selector over parser-proposed candidates, into the LLM’s *hidden representation of the input language*, reshaping how it encodes spatial *text* rather than adding an external geometry module; and the whole apparatus (parser, candidates, selector, residuals) is dropped at inference, leaving a text-in/text-out LLM whose only added input is one learned prompt token. DGCL is thus an instantiation of this training-time-only family for textual spatial reasoning over a typed constraint library; its main experimental contribution is the controlled A6-vs-A2c comparison (§4.3). Prompting- and tool-time methods (Wei et al., 2022; Wang et al., 2023; Yao et al., 2023a; Lightman et al., 2024; Gao et al., 2023; Chen et al., 2023; Yao et al., 2023b; Schick et al., 2023) form our budget-matched B3b baselines.

3 Problem Formulation

Given input text x_i (and, for D2, optional structured fields), the model outputs an answer $\hat{y}_i$ evaluated by hard validators that reject answers failing canonical normalization, binding, or tolerance (Lu et al., 2021). A frozen training-time parser P_1 proposes a candidate set $C_i = \{c_{ij}\}$ with typed labels t_{ij} , argument tuples a_{ij} , and confidences p_{ij} ; P_1 is used only during training and is absent at inference. It is non-oracle but not deliberately weak (§4.5), which calibrates the tool baselines’ headroom. The training objective adds three terms to the answer NLL $\mathcal{L}_{ans}$:

$$\mathcal{L}_{A6} = \mathcal{L}_{ans} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_{cov}\mathcal{L}_{cov} + \lambda_{sel}\mathcal{L}_{sel}, \quad (1)$$

all operating on candidate constraints and all removed at inference.

4 Method: DGCL

This section specifies the four pieces that distinguish DGCL from a vanilla fine-tuned model: a typed family of differentiable residuals (§4.1), a variable predictor that maps hidden states to geometric variables (§4.2), and a learned selector with a coverage regularizer for candidate weighting (§4.4). The gradient path that distinguishes A6 from its stop-gradient ablation A2c is made explicit in §4.3. Figure 1 contrasts the training-time and inference-time data flow.

4.1 Differentiable Geometric Residual Loss

Let $\hat{z}_i$ denote the predicted geometric variables for instance i (point coordinates for D1, object centers/extents/orientations for D2; see §4.2). For each candidate constraint c_{ij} with type t_{ij} and argument tuple a_{ij} , a typed differentiable residual function $\phi_{t_{ij}}(\hat{z}_i, a_{ij})$ measures violation. We pass each residual through the Huber penalty ρ_δ to control the influence of outliers:

$$\rho_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta), & |r| > \delta. \end{cases} \quad (2)$$

Candidate weights are a softmax over selector logits (§4.4) at temperature τ :

$$w_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{k=1}^{m_i} \exp(s_{ik}/\tau)}. \quad (3)$$

The differentiable geometric loss aggregates weighted residuals:

$$\mathcal{L}_{geo} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{m_i} w_{ij} \rho_\delta(\phi_{t_{ij}}(\hat{z}_i, a_{ij})). \quad (4)$$

The library has eighteen typed families (Table 3, Appendix C): eleven planar residuals over predicted points $\hat{p}$ — e.g., parallel cross($\text{unit}(\hat{p}_B - \hat{p}_A), \text{unit}(\hat{p}_D - \hat{p}_C)$) and angle equality ($\angle(\hat{p}_A, \hat{p}_B, \hat{p}_C) - \angle(\hat{p}_D, \hat{p}_E, \hat{p}_F)$)/ π , plus distance equality, fixed length, perpendicular, supplementary angle, collinearity, midpoint, segment addition, circle radius, and similarity ratio — and seven frame-dependent 3D residuals over predicted centers $\hat{c}$ — e.g., left/right $\max(0, \mu_{lr} - \text{sign}_t e_x(f)^\top(\hat{c}_u - \hat{c}_v))/D_i$ and frame transform $\|T_{f_1 \rightarrow f_2}(\hat{c}_u - \hat{c}_v) - r_{uv, f_2}^{\text{gold}}\|_2/D_i$, plus front/behind, above/below, non-overlap, and two higher-order consistency residuals (SE(3) composition and counterfactual swap; Appendix C). Each is normalized by an instance-scale denominator D_i (diagram diameter for D1, scene diagonal for D2) or an angle scale so weights are comparable across families. The residuals are a training loss only; primary evaluation uses independent hard validators (§5.3) that are code-separated from the residuals so low residuals cannot game evaluation. Most residuals are *self-consistency* terms; the frame-transform residual is the one *supervised* term, comparing the predicted offset mapped into frame f_2 against a train-split target r^{gold} read from the dataset’s 3D ground truth — defined only on the train split, a property of the

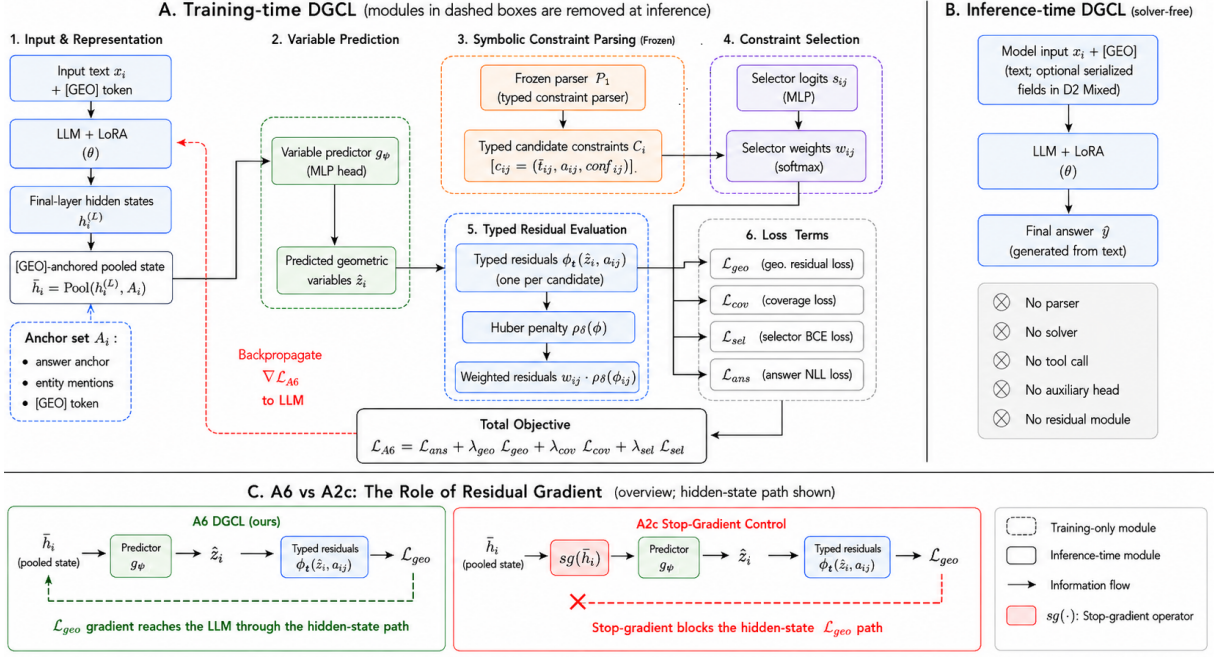


Figure 1: DGCL overview. **(A)** Training-time DGCL: the frozen parser P_1 proposes typed candidate constraints, the variable predictor maps a [GEO]-anchored attention-pooled representation to geometric variables $\hat{z}_i$, typed residuals are evaluated under a Huber penalty and selector weights, and the total objective $\mathcal{L}_{A6} = \mathcal{L}_{ans} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_{cov}\mathcal{L}_{cov} + \lambda_{sel}\mathcal{L}_{sel}$ backpropagates the geometric residual gradient into the LLM. Modules in dashed boxes are removed at inference. **(B)** Inference-time DGCL is a single LLM+LoRA forward pass on the model input, with no parser, solver, tool, or auxiliary head. **(C)** A6 vs. A2c, hidden-state path shown: A2c detaches $\mathcal{L}_{geo}$ from this and from the selector-weight path, both drawn in Figure 3; $\mathcal{L}_{sel}$ and $\mathcal{L}_{cov}$ stay coupled in both (§4.3).

scene rather than the answer label, and absent at inference. Provenance, masking of under-specified frames, and the numerical-stability guards for removable singularities are in Appendix C; because this term is supervised, §6.1 rules out the trivial explanation that A6 merely benefits from r^{gold} .

4.2 Variable Predictor and Hidden-State Pooling

The variable predictor produces $\hat{z}_i$ by attention-pooling final-layer LLM hidden states over an anchor set (the answer token, entity mentions, and a learned [GEO] token; Appendix B) and applying a two-layer MLP:

$$\bar{h}_i = \text{Pool}(h_i^{(L)}, \mathcal{A}_i), \quad \hat{z}_i = W_2 \sigma(W_1 \bar{h}_i + b_1) + b_2. \quad (5)$$

In the True-Text-Only condition the latent variables $\hat{z}_i$ are predicted *from text alone* — no coordinates, boxes, or structured fields enter the predictor — so the residual teaches the model to recover latent geometry from language; at inference the text-only forward pass already carries that ability without producing $\hat{z}_i$ at all. In the Mixed condition the serialized structured fields are additionally avail-

able in $h_i^{(L)}$. Output dimensions, the per-instance similarity gauge that removes the coordinate null space, and the canonical slot layout are deferred to Appendix C.

Inference. [GEO] is a single learned embedding appended once to the prompt before the answer region (Appendix B); it is present at both training and inference, is never generated, and is the only input the deployed model reads beyond the text. Everything else in Figure 1(A) — parser P_1 , candidates C_i , selector, variable predictor, residuals, and $\mathcal{L}_{geo}$, $\mathcal{L}_{sel}$, $\mathcal{L}_{cov}$ — is absent at inference, a standard LLM+LoRA forward pass; the one-token overhead is included in the $1.06\times$ latency of §6.5 (Appendix O).

4.3 Hidden-State to Geometry Bridge

The bridge from LLM hidden states to geometric variables separates A6 from a detached geometry head. In A6 the full path $h_i^{(L)} \rightarrow \bar{h}_i \rightarrow \hat{z}_i \rightarrow \phi_t(\hat{z}_i, a) \rightarrow \mathcal{L}_{geo}$ is differentiable, so $\nabla_{\theta_{LLM}} \mathcal{L}_{geo} \neq 0$ and the geometric residual updates the LoRA parameters. The stop-gradient ablation A2c detaches *only* $\mathcal{L}_{geo}$ from the LLM,

leaving the q_{ij} -supervised $\mathcal{L}_{sel}/\mathcal{L}_{cov}$ coupled exactly as in A6, via $\hat{z}_i^{A2c} = g_\psi(\text{sg}(\bar{h}_i))$ and $w_{ij}^{A2c,geo} = \text{sg}(w_{ij})$, giving $\nabla_{\theta_{LLM}} \mathcal{L}_{geo} = 0$ while $\nabla_{\theta_{LLM}} \mathcal{L}_{sel}, \nabla_{\theta_{LLM}} \mathcal{L}_{cov} \neq 0$. A2c is thus *not* B2 FT: the q_{ij} -supervised losses still teach a “which gold constraints are present” signal; what A2c lacks is the *continuous* geometric-consistency gradient.

One subtlety affects the causal reading. The full A2c detaches $\mathcal{L}_{geo}$ along *both* paths it could flow — the hidden-state bridge $\bar{h}_i \rightarrow \hat{z}_i$ and the dependence on selector weights w_{ij} — while A6 couples both, so A6–A2c conflates residual gradient reaching the *LLM hidden states* with residual gradient refining the *selector weighting*. To isolate them we add a 2×2 design (§6.1, Table 5): **A6-sel-detach** keeps the hidden-state path but stops $\mathcal{L}_{geo}$ at w_{ij} (hidden coupling only); **A2c-hidden-only** stops the hidden-state path but lets $\mathcal{L}_{geo}$ update w_{ij} (selector refinement only).

4.4 Selector and Coverage

A small differentiable selector produces per-candidate logits $s_{ij} = u^\top \tanh(W_h \bar{h}_i + W_c e(c_{ij}) + W_p p_{ij} + b_s)$, where $e(c_{ij})$ embeds candidate type and binding features and p_{ij} is parser confidence; the softmax weights w_{ij} enter $\mathcal{L}_{geo}$ differentiably, so the residual gradient updates both the selector and the LLM bridge. Selector supervision is a per-candidate BCE against train-split matching labels q_{ij} , $\mathcal{L}_{sel} = -\frac{1}{N} \sum_i \frac{1}{m_i} \sum_j [q_{ij} \log \sigma(s_{ij}) + (1-q_{ij}) \log(1-\sigma(s_{ij}))]$. To stop the selector collapsing onto one easy candidate, a coverage hinge on the soft coverage $\text{Cov}_i = \sum_j w_{ij} q_{ij}$ penalizes mass below threshold γ : $\mathcal{L}_{cov} = \frac{1}{N} \sum_i \max(0, \gamma - \text{Cov}_i)^2$. All q_{ij} are train-split only and never expose test gold constraints. Two ablations isolate the design: *UniformCandidate* ($w_{ij}=1/m_i$) and *DetachedSelector* ($\text{sg}(w_{ij})$ before $\mathcal{L}_{geo}$); coverage is ablated by “No coverage regularizer” (Table 10).

4.5 Parser and Candidate Construction

The candidate set C_i and matching labels q_{ij} come from the frozen parser P_1 , a DeBERTa-v3-base (He et al., 2023) constraint extractor fine-tuned on a held-out parsing split (18,240 instances, disjoint from every main split). Each candidate is a typed tuple (t_{ij}, a_{ij}, p_{ij}) — type, ordered argument tuple over canonical entity slots, and parser confidence — with up to $m_i \leq 20$ deduplicated candidates per instance. The binary label q_{ij} marks whether a can-

didate matches a gold *train-split* constraint (identical type after synonym normalization, argument agreement under canonical ordering, and numeric agreement within tolerance); it is never computed against dev or test, so the selector/coverage supervision cannot leak answers. P_1 is trained once, frozen, and absent at inference. It is non-oracle: Candidate Recall@20 ranges from 73.2% (D3) to 92.8% (D1 generated) and argument-binding accuracy from 59.0% to 86.4% (Table 19, Appendix L), so training-candidate quality bounds the geometric signal; §6.6 measures how A6 degrades when these candidates are corrupted. Full construction (parsing-split provenance, generator sharing, label schema, relation distribution, and per-type quality) is in Appendix L.

4.6 Hyperparameters and Training Budget

All runs use frozen hyperparameters selected on development splits only (Appendix C; key values $\lambda_{geo}=0.35$, $\lambda_{cov}=0.05$, $\lambda_{sel}=0.10$, Huber $\delta=0.10$, selector $\tau=0.70$). The backbone is Llama-3.1-8B-Instruct (Meta AI, 2024) with LoRA (Hu et al., 2022) ($r=16$, $\alpha=32$, dropout 0.05) on q, k, v, o , gate, up, down; optimizer AdamW (Loshchilov and Hutter, 2019). The budget is identical across all matched trainable baselines: effective batch 64, ≤ 3.0 epochs over the 54,600-instance pool, 2,562 optimizer updates, 77 warmup updates, max sequence length 2,048.

5 Evaluation Protocol

5.1 Datasets

We use three datasets; split sizes (totals 54,600/7,800/10,278 train/dev/test, D3 evaluated only at the locked run) are in Appendix G, Table 7.

D1: planar geometry. **D1 public** is a license-cleaned subset of public planar-geometry problems in the Inter-GPS/Geometry3K tradition (Lu et al., 2021); **D1 generated** uses a frozen context-free grammar with disjoint template-family IDs across splits. All D1 instances use a hard validator (canonical normalization plus numeric/angle tolerance). Because D1 generated is grammar-generated from a fixed constraint vocabulary, it primarily measures *typed residual alignment*, not open-domain generalization; D1 public and the firewalled D3 split test transfer beyond the generation grammar, still within the typed families covered by the residual library (Limitations).

D2: language-grounded 3D relations. D2 contains language-grounded 3D scene questions over scenes derived from ScanNet (Dai et al., 2017) and 3DSSG (Wald et al., 2020). Each item asks for a typed relation tuple (u, t, v, f) between objects u, v under frame f .

Structured fields are an optional *input*: a serialized list of object IDs and, where available, coarse 3D bounding boxes and a frame tag. They are model inputs, not answers, and the validator never reads them. *The True-Text-Only (TTO) condition* presents the model with language only — no structured fields in the input — at both training and test, and is the harder, more realistic setting; *the Mixed condition* additionally provides the structured fields as input at both training and test. r^{gold} is neither an input nor an answer: it is the train-split 3D ground-truth relative offset used to supervise the frame-transform residual (§4.1). r^{gold} is a *training-time supervision target in both conditions* and is absent from the input and from inference in both; in TTO the model additionally never sees structured fields even as input, so TTO isolates whether DGCL’s training-time signal transfers to a purely textual deployment. The robustness controls in Appendix M stress exactly these boundaries by shuffling or dropping the structured fields. Primary relation families exclude diagnostic Type II families, which are reported separately in the appendix.

D3: zero-touch counterfactual transfer. D3 contains 2,478 counterfactual templates held out under a firewall protocol: no D3 example is used for prompt tuning, hyperparameter selection, parser tuning, or error analysis prior to the locked evaluation run. Firewall certification is logged as a signed statement and cryptographic hashes of the no-tuning record, the nearest-neighbor audit, and the final retained-hashes file. D3 has no train or dev split.

5.2 Baselines

We compare against a fixed registry under matched candidate budgets and validation rules: prompting baselines **B1 Zero-Shot** and **B2 Few-Shot CoT** (Wei et al., 2022); **B2 FT** (standard LoRA fine-tuning on the same answer supervision); the tool baselines **B3b Budget/Strong/Strong+** that call P_1 at inference and run candidate-subset search plus a symbolic solver with timeouts (Budget ≤ 32 subsets/10 calls; Strong $\leq 256/50$; Strong+ expanded cap 40 and a pre-registered stress bud-

get; all share the frozen P_1); the locked-constraint control **B5b** and serialized-scene-graph control **B5c**; **External PAL/PoT** (Gao et al., 2023; Chen et al., 2023) and **External Verifier** (Lightman et al., 2024); a **GPT-4o-mini CoT** external-API sanity reference (not a matched trainable baseline); the Innovation-I falsifier **A2c**; the Innovation-II falsifiers **RandGeo/ParamMatch**; and full **A6**. All trainable baselines share the backbone, LoRA config, optimizer, training budget, candidate source, and invalid-as-failure rule.

5.3 Metrics

Primary metrics use hard, independently-coded validators separate from the training residuals. Answer correctness is exact-match after normalization with numeric tolerance $\max(10^{-4}, 10^{-3}|y_i|)$ and angle tolerance 0.5° . **HV_AnsAcc** (D1) multiplies answer correctness by a hard validator rejecting malformed outputs; **RA_RelVal** (D2) further requires tuple parse, entity link, and a hard relation check (frame-dependent relations use frame-transformed center differences with margin $0.02 \times$ scene diagonal); **OOD_AnswerAcc** (D3) is answer accuracy on the firewalled split; trace variants additionally validate the proof (**FVPA**). All invalid outputs — missing/malformed answer, unparseable tuple, solver timeout, or post-hoc repair — count as failure; auxiliary diagnostics are in the appendix.

5.4 Statistical Protocol

We run 5 seeds (13, 21, 42, 87, 100) per trainable variant. **The significance unit is the individual test instance, not the seed.** For each pre-registered gate, McNemar’s test uses the per-instance majority-vote outcome across the five seeds, while the clustered randomization test uses the paired instance-level statistic with the pre-registered clustering variable — D1 public 312 source groups, D1 generated 1,046 template families, D2 874 scene IDs, D3 58 counterfactual families. We use 10,000 cluster-level randomization draws and the finite-sample Monte-Carlo estimate $p_{\text{perm}} = (k + 1)/(B + 1)$, where $B = 10,000$ and k is the number of permuted statistics at least as extreme as the observed statistic; the gate-level raw p -value is the more conservative of the McNemar and cluster-randomization values. The seven raw p -values are adjusted jointly with **Holm-Bonferroni** (Holm, 1979) at the pre-registered family-wise level $\alpha = 0.001$. The reported 95% confidence intervals come separately from a clustered bootstrap

(10,000 resamples over the seed $\times$ cluster structure). The pre-registration file (comparison, direction, metric, test, clustering variable, threshold) is hashed and frozen before the locked run (Appendix N). The exploratory analyses in §6.6 are post-hoc: unadjusted pointwise intervals, separate from the seven gates.

6 Results

6.1 Main Results

Figure 2 (full numbers in Appendix H, Table 14) reports primary answer-only results. A6 obtains the highest mean in every setting and exceeds the stop-gradient ablation A2c by +4.0 to +5.2 despite identical candidates, residuals, selector, and coverage regularizer; the closest matched competitor is B3b Strong+ on D1 generated (95.9 vs. 88.4), with caveats in §6.3.

Innovation I: coupling adds a stable gain over strong constraint supervision. We read A6 vs. A2c as an *increment over a strong control*, not the source of the larger gain over plain fine-tuning: A2c already captures most of the gain over B2 FT via the q_{ij} -supervised selector/coverage pathway (§7), and the two share their data, candidates, residual library, selector/coverage objectives, and budget, differing only in whether $\mathcal{L}_{geo}$ reaches the LLM along either path (§4.3). The per-setting deltas are +4.6/+4.9/+4.0/+4.5/+5.2 (mean +4.64, 95% CI [+3.76, +5.53], clustered bootstrap), with 5/5 positive seeds per setting and all 25 seed $\times$ setting deltas positive (Appendix D); under the protocol of §5.4 the differentiability gate remains significant after Holm correction ($p < 0.001$; Table 1). The effect is small but consistent in sign and significance.

Isolating hidden-state coupling from selector refinement. Because the full A2c detaches $\mathcal{L}_{geo}$ from *both* the LLM hidden states and the selector weights (§4.3), A6–A2c by itself conflates the two. A 2 $\times$ 2 decomposition (Appendix E, Table 5) separates them: enabling only hidden-state coupling (A6-sel-detach) recovers a mean +3.60 of the +4.64 gain (78%), whereas enabling only selector refinement (A2c-hidden-only) recovers +0.84. The gain is therefore principally attributable to the residual gradient reaching the LLM hidden states, with a smaller residual-aware-selector contribution; this is the clean form of the Innovation I claim.

Gate (A6 vs.)	Mean Δ	95% CI	Holm p
A2c (Differentiable)	+4.64	[+3.76, +5.53]	< 0.001
B2 FT	+26.30	[+24.71, +27.96]	< 0.001
B3b Budget	+21.85	[+20.07, +23.60]	< 0.001
B3b Strong	+16.04	[+14.16, +17.92]	< 0.001
RandGeo	+10.84	[+9.34, +12.33]	< 0.001
ParamMatch	+11.79	[+10.32, +13.27]	< 0.001
B5b Constraint	+15.65	[+13.83, +17.46]	< 0.001

Table 1: Seven pre-registered superiority gates. Each Δ is the pooled mean accuracy difference; the 95% CIs come from the clustered bootstrap and the last column from the pre-registered paired testing procedure, both described in §5.4. All seven gates satisfy the pre-registered family-wise $\alpha = 0.001$ criterion; the differentiability gate (A2c) is the smallest-effect comparison. Per-setting sign counts for A6 vs. A2c are 5/5 positive.

Innovations II and III. The gain is geometry-specific: RandGeo and ParamMatch sit far below A6, and by the widest margin on the slices where geometric structure matters most (§6.4), ruling out extra parameters or a generic auxiliary signal. And inference is non-oracle and solver-free: A6 calls no parser or solver and runs at $1.06\times$ B2 FT latency versus $3.9\text{--}21.5\times$ for tool baselines (§6.5).

The D2 gain is not just access to r^{gold} . Giving the same train-split target r^{gold} (§4.1) to a model through a generic regression route rather than the differentiable residual (EquivSup) stays 14.6 points below A6 on D2 Mixed, while removing r^{gold} from A6 entirely costs only 2.9 (Appendix F, Table 6): the gain is from the mechanism, not the target.

6.2 Pre-Registered Claim Gates

All seven pre-registered gates satisfy the family-wise $\alpha = 0.001$ criterion after Holm–Bonferroni correction (Table 1). The largest delta (+26.30, A6 over B2 FT) is mostly the constraint-supervision benefit that A2c already obtains and is *not* our contribution; the smallest (+4.64, A6 over A2c) is the part attributable to coupling the residual gradient into the LLM. We frame the contribution as a +4 to +5 point stable gain over a strong control (per-dataset breakdown in Appendix G, Table 8). We did not gate against B3b Strong+ because it is competitive with A6 in the easiest (full-coverage) condition (§6.3).

6.3 Boundary: When Tool Search is Competitive

Table 2 decomposes by parser-coverage subset: under full P1 coverage B3b Strong+ marginally ex-

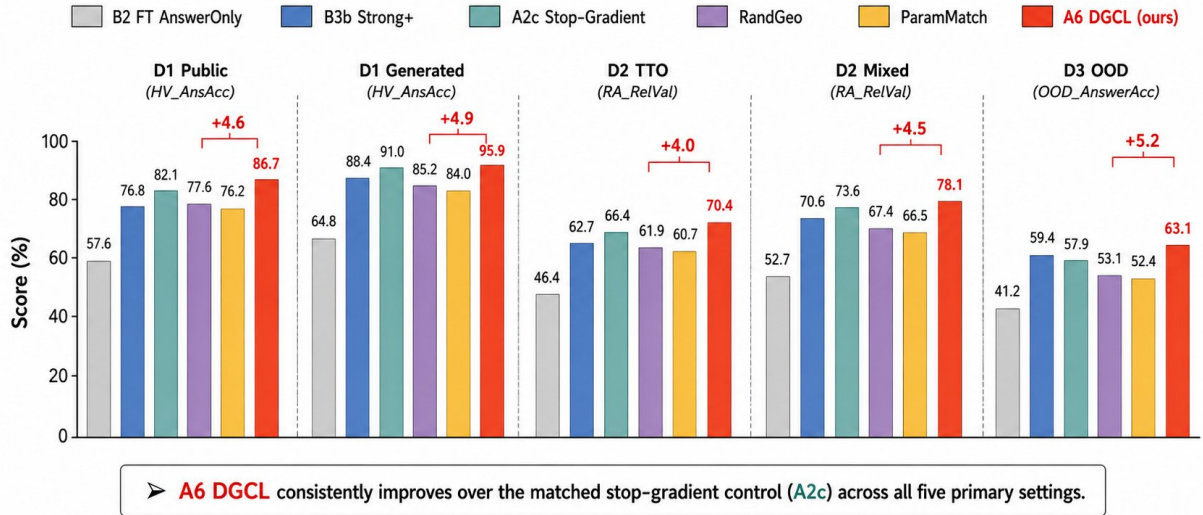


Figure 2: Main performance comparison across the five primary settings. A6 DGCL consistently exceeds the matched stop-gradient control A2c by +4.0 to +5.2 points (red brackets), and the geometry-specificity falsifiers RandGeo and ParamMatch stay below A6 in every setting. Full numeric results with confidence intervals and all baselines are in Table 14 (Appendix H).

P_1 coverage subset	B3b Strong	B3b Strong+	A6
Full	82.5	91.0	90.2
Partial	65.0	72.0	81.5
Missing	43.0	44.5	57.0

Table 2: Parser-coverage robustness (HV_AnsAcc). Under full P_1 coverage the strong solver pipeline B3b Strong+ marginally exceeds A6 (91.0 vs. 90.2); under partial or missing coverage A6 is substantially more robust (+9.5 and +12.5). Full numbers with intervals are in Table 15 (Appendix I).

ceeds A6 (91.0 vs. 90.2), but A6 leads by 9.5 (partial) and 12.5 (missing), so A6’s aggregate advantage is robustness under parser failure, not solver superiority. This matches inference-time-symbolic methods (Trinh et al., 2024; Chervonyi et al., 2025; Ping et al., 2026) being strong only when formalization succeeds; an oracle solver still exceeds A6 (Table 9), so DGCL is not at the symbolic ceiling.

6.4 Ablation and Specificity

The module-level ablation (Table 10) shows the geometric loss is the largest contributor (−26.1 mean), with candidate weighting (−6.3/−5.3), coverage (−4.4), the variable predictor (−7.7), and the Huber penalty (−2.9) each contributing. The specificity slice (Table 11) shows A6’s lead over RandGeo/ParamMatch is largest on the slices where geometric structure matters most (up to +13.7/ +14.4 on D3 low-lexical), so the gains are geometry-specific.

6.5 Efficiency and Robustness

A6’s relative latency is $1.06 \times$ B2 FT in the primary token bucket ($p_{95} 1.21 \times$) versus $3.9\text{--}21.5 \times$ for the B3b family, with zero parser/solver calls (Table 12). A6 also reduces coordinate MSE and angular error by roughly one third over B2 FT across all five slices, including firewalled D3 (Table 13); under input perturbations it degrades in proportion to the geometric content removed (query-only −38.6, paraphrase −3.2 on D2 Mixed; Table 22) and never exceeds its full-input value (Appendix Q).

6.6 Exploratory Analyses: Parser Sensitivity, Within-Library Transfer, and Probes

Three post-hoc analyses probe the boundaries of the A6—A2c effect. They are exploratory (unadjusted pointwise intervals, §5.4). The first two retrain A6 and A2c with matched seeds, data order, and budget (120 additional LoRA runs, no new hyperparameter search); the probe analysis reuses the frozen checkpoints. All are evaluated on D1 public and D2 TTO, since D1 generated is an alignment test and D3 cannot be re-evaluated without breaking its zero-touch protocol (setup and full tables in Appendix R).

Training-parser sensitivity. Table 30 corrupts the frozen top-20 training candidates (deletions are not replaced). A6’s absolute accuracy falls in every condition—by 9.2/9.4 points (D1 public/D2 TTO) under 50% gold-candidate removal and 9.7/9.6 un-

der binding corruption—so DGCL is parser-free at inference but parser-dependent in training. The A6–A2c gap stays positive in all ten comparisons and nine intervals exclude zero; D2 BindCorrupt-25 (+2.1 [−0.2, 4.4], 3/5 seeds) is inconclusive. Omitting a whole family costs the targeted slice far more than the rest of the data (16.4 vs. 1.79 points for D1 angle relations, 13.2 vs. 2.70 for D2 left/right).

Within-library transfer under masked auxiliary supervision. For a target family group we keep its examples and $\mathcal{L}_{ans}$ but set $\mathcal{L}_{geo} = \mathcal{L}_{sel} = \mathcal{L}_{cov} = 0$ on those examples (suffix “HO”; the family is *not* held out from answer training). A6-HO exceeds A2c-HO by +2.4 [1.0, 3.8] on the D1 macro and +1.9 [0.7, 3.1] on the D2 macro, but three of seven family-level intervals include zero and the compositional slice is weakest (+1.0 [−1.8, 3.8]). Because answer supervision remains and every family is in the library, this is limited within-library transfer, not evidence about unseen types.

Linear decodability. Frozen linear probes on the final-layer state at the last prompt token (after [GEO], not the [GEO] state; fitted on train, tuned on dev, evaluated once on test) decode geometric information better from A6 than from A2c: D1 family macro-F1 +5.50 [3.20, 7.80], D1 coordinate NRMSE −0.027 [−0.041, −0.013], D2 relation macro-F1 +5.30 [2.90, 7.70], D2 axis balanced accuracy +5.10 [2.70, 7.50]. Hidden coupling alone (A6-sel-detach) recovers about 70–85% of each difference and selector-only contrasts include zero, mirroring Table 5. These probes show higher linear decodability, not causal use during generation.

7 Discussion

What the gain is, and why it reaches the LLM. A2c already beats B2 FT (e.g., 82.1 vs. 57.6 on D1 public) not through vague regularization but because its q_{ij} -supervised $\mathcal{L}_{sel}/\mathcal{L}_{cov}$ are not detached (§4.3); removing that supervision collapses A2c to $\approx$ B2 FT (Table 10). The narrower A6–A2c gap isolates the differentiable coupling, which the decomposition (§6.1) attributes principally to gradient reaching the LLM hidden states, helping exactly the partial-coverage, frame-dependent, and low-lexical cases where geometric consistency is most informative. Frozen probes (§6.6) are consistent with this reading but establish decod-

ability, not causal use during generation. The strongest tool baseline cannot simply scale candidates: a larger cap raises coverage but also timeout/rejection/exhaustion rates (8.6/7.5/5.9%). We do not claim the LLM “computes geometry”; the predictor and selector do real work, but not in isolation.

Scope. We do not claim parser-independence (a frozen parser produces training candidates), superiority over oracle solvers (the oracle exceeds A6 by 2.7–8.8 points, Table 9), or that solver methods are inferior in general (Trinh et al., 2024; Chervonyi et al., 2025; Ping et al., 2026). Two boundaries deserve emphasis. First, DGCL is parser-free at inference but parser-dependent in training: corrupting the training candidates lowers A6’s accuracy in every tested condition, and omitting a whole family costs the targeted slice 16.4 (D1) and 13.2 (D2) points (§6.6), so systematic family-level parser failure is a material limitation. Second, the evidence covers the typed families in Table 3; masking a family’s auxiliary supervision yields only limited within-library transfer, families outside the library are untested, and DGCL provides no per-answer certificate and is evaluated text-only and answer-only (Limitations). The A6–A2c effect reproduces on a Qwen2.5-7B backbone (Appendix P) with the same sign and per-seed pattern, so it is not Llama-specific.

8 Conclusion

We presented DGCL, a training-time objective that supervises an LLM’s hidden states with typed Huber residuals over candidate constraints for a typed family of spatial relations, removing the entire symbolic apparatus at inference. Most of the gain over plain fine-tuning is the expected constraint-supervision benefit already obtained by the detached control A2c; the contribution we isolate is the +4.0 to +5.2 point gain of full differentiable coupling (A6 over A2c), most of which the 2×2 decomposition attributes to the hidden-state path (§6.1), positive in all five settings and significant under the pre-registered gates ($1.06 \times$ latency) — symbolic structure as a training-time gradient signal (and a training-time parser dependency), not a runtime dependency. Frozen probes show higher linear decodability of geometric information in the final hidden state; causal use during generation and transfer beyond the covered families remain open.

Limitations

No formal guarantee. DGCL shapes the model’s weights toward geometric consistency, but it provides *no formal proof certificate* for any individual answer: at inference there is no solver, no constraint check, and no produced derivation, so a DGCL answer can be geometrically wrong without any internal signal flagging it. This is a fundamental difference from symbolic provers such as the AlphaGeometry family (Trinh et al., 2024; Chervonyi et al., 2025), whose outputs come with a checkable deductive trace. Where a verified, certificate-bearing solution is required — formal verification, automated grading that must justify itself, or any safety-critical setting — DGCL is not a substitute for an oracle solver and should at most be paired with one as a fast first pass. A hybrid pipeline that verifies selected DGCL outputs with a solver is possible but outside our experiments. Our claims are strictly about answer accuracy and robustness under imperfect parser coverage, not about provable correctness.

Parser dependence at training time. Although the inference path is solver-free, training requires a frozen P1 parser. When P1 systematically misclassifies a relation type, the corresponding training residuals are mislabeled; the selector and coverage regularizer can re-weight but not relabel. We report argument-binding accuracy and per-relation breakdowns (Appendix L) and measure the dependency directly (§6.6, Appendix R): A6’s absolute accuracy falls in every corruption condition — by 9.2/9.4 points (D1 public/D2 TTO) under 50% gold-candidate removal and 9.7/9.6 under binding corruption — and omitting an entire family from the candidates degrades the targeted slice by 16.4 (D1) and 13.2 (D2) points against 1.79 and 2.70 elsewhere. The A6–A2c margin narrows under corruption and is inconclusive in one condition (D2 BindCorrupt-25). We did not test a precision-only condition that adds spurious candidates at fixed recall, and we do not eliminate the dependency.

Evaluation scope. D1 covers planar Euclidean geometry; D2 covers a fixed family of object-level frame-dependent 3D relations; D3 is a counterfactual perturbation of the same families. Extension to multimodal inputs (images, scans, navigation traces) is outside this paper. Recent VLM-style spatial work (Chen et al., 2024; Cheng et al., 2024; Wu et al., 2024) is complementary and we

expect the DGCL gradient path to compose naturally with those representation pipelines, but we do not test the combination here. The evaluation is also answer-only: we do not assess proof quality or the model’s ability to detect its own errors, and our claims concern answer accuracy and robustness under imperfect parser coverage, not verifiable correctness.

Statistical scope. Significance is established at the instance level with clustered resampling (§5.4); the five seeds contribute the run-to-run variance component, not the test unit. Power is nonetheless limited where the number of *clusters* is small — most acutely on D3, with only 58 counterfactual families — so small per-relation differences in the long tail are not always resolvable, and the D3 confidence intervals are correspondingly the widest. We report clustered-bootstrap intervals throughout to make this visible rather than reporting narrower seed-only intervals that would overstate precision. The exploratory analyses of §6.6 are post-hoc; their intervals are pointwise and not multiplicity-adjusted.

Residual library and dataset coupling. The typed residual families in Table 3 were designed for, and evaluated on, datasets whose gold constraints are drawn from a comparable typed family. This is most acute for D1 generated, whose grammar draws from the same constraint vocabulary the residual library targets; its high scores chiefly reflect typed residual alignment rather than open-domain generalization (§5.1). We cannot rule out that part of A6’s advantage reflects a match between the residual library and the dataset’s constraint vocabulary rather than a property that would transfer to arbitrary geometric tasks. A residual family absent from our library (for example, conic-section incidence or topological invariants beyond GTC) would receive no training signal, and a dataset dominated by such constraints might show a much smaller gain.

Representation evidence. The frozen probes (§6.6) show that family, coordinate, and relation information is more linearly decodable from A6’s final pre-generation state than from A2c’s, with most of the difference associated with the hidden-coupling path. They do not show that this information is used causally during generation; hidden-state interventions and transfer to a different answer format are left to future work.

Compute footprint. All A6 runs use a single A100 80GB; the exploratory analyses added 120 matched LoRA runs (about 0.8k A100-hours) under the same schedule, with no new hyperparameter search. We did not perform a multi-node scaling study, and the LoRA configuration (rank 16 on 7 projections) is fixed; results at larger ranks or on larger backbones may differ.

Ethics and Reproducibility

D1 public is drawn from license-cleaned public planar-geometry problems. D2 scenes are derived from ScanNet (Dai et al., 2017) and 3DSSG (Wald et al., 2020) under the original license terms. D3 templates are synthesized in-house. No human-subject data is collected. The training-time parser is a DeBERTa-v3 model (He et al., 2023) fine-tuned on a non-overlapping parsing split.

The D3 firewall protocol is documented by the seven audit artifacts listed in Table 25, each sealed under a full-length SHA-256 digest: (1) the pre-registration file `prereg_gates.md` (the seven gates with direction, metric, test, clustering variable, and threshold); (2) `d3_generator_hash.txt`; (3) `d3_template_registry.jsonl`; (4) the timestamped role-based `d3_access_log.csv`; (5) the signed `d3_no_tuning_statement.md`; (6) `d3_nearest_neighbor_audit.jsonl`, including the 274 removed instances; and (7) `d3_final_retained_hashes.txt`. Together with the audit and hashing scripts, these artifacts record the pre-registration, the firewall checks, and the temporal ordering of the protocol (§N).

References

- Samy Badreddine, Artur d’Avila Garcez, Luciano Serafini, and Michael Spranger. 2022. [Logic tensor networks](#). *Artificial Intelligence*, 303:103649.
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. [SpatialVLM: Endowing vision-language models with spatial reasoning capabilities](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465.
- Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. [GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#). *Transactions on Machine Learning Research*.
- An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. [SpatialRGPT: Grounded spatial reasoning in vision-language models](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Yuri Chervonyi, Trieu H. Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Junsu Kim, Vikas Verma, Quoc V. Le, and Thang Luong. 2025. [Gold-medalist performance in solving olympiad geometry with AlphaGeometry2](#). *Journal of Machine Learning Research*, 26(241):1–39.
- Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. [ScanNet: Richly-annotated 3D reconstructions of indoor scenes](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [PAL: Program-aided language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Sture Holm. 1979. [A simple sequentially rejective multiple test procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yunze Song, Wai Lam, and Yue Zhang. 2024. [Chain-of-symbol prompting elicits planning in large language models](#). In *Proceedings of the Conference on Language Modeling (COLM)*.
- Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2024. [TopViewRS: Vision-language models as top-view spatial reasoners](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1807, Miami, Florida, USA. Association for Computational Linguistics.

- Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2024. [Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17028–17047, Miami, Florida, USA. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. [Visual spatial reasoning](#). *Transactions of the Association for Computational Linguistics*, 11:635–651.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. [Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.
- Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. [DeepProbLog: Neural probabilistic logic programming](#). In *Advances in Neural Information Processing Systems*, volume 31, pages 3753–3763.
- Meta AI. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Bowen Ping, Minnan Luo, Zhuohang Dang, Chenxi Wang, and Chengyou Jia. 2026. [AutoGPS: Automated geometry problem solving via multimodal formalization and deductive reasoning](#). In *The Fourteenth International Conference on Learning Representations*.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. [An empirical analysis on spatial reasoning capabilities of large multimodal models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, Miami, Florida, USA. Association for Computational Linguistics.
- Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024. [Solving olympiad geometry without human demonstrations](#). *Nature*, 625(7995):476–482.
- Johanna Wald, Helisa Dharmo, Nassir Navab, and Federico Tombari. 2020. [Learning 3D semantic scene graphs from 3D indoor reconstructions](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3961–3970.
- Po-Wei Wang, Priya L. Donti, Bryan Wilder, and J. Zico Kolter. 2019. [SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver](#). In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6545–6554. PMLR.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Azmine Touseh Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md. Iqramul Hoque, Shahriyar Zaman Ridoy, Mohammed Eunus Ali, Majd Hawasly, Mohammad Raza, and Md Rizwan Parvez. 2026. [SpatiaLab: Can vision-language models perform spatial reasoning in the wild?](#) In *The Fourteenth International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brain Ichter, Fei Xia, Ed H. Chi, Quoc. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. [Mind’s eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. 2018. [A semantic loss function for deep learning with symbolic knowledge](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5502–5511. PMLR.
- Zhun Yang, Adam Ishay, and Joohyung Lee. 2020. [NeurASP: Embracing neural networks into answer](#)

set programming. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1755–1762.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems*, volume 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. [ReAct: Synergizing reasoning and acting in language models](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Changyu Zeng, Yifan Wang, Zimu Wang, Wei Wang, Zhengni Yang, Muiy Bao, Jimin Xiao, Anh Nguyen, and Yutao Yue. 2025. [NUMINA: A natural understanding benchmark for multi-dimensional intelligence and numerical reasoning abilities](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22575–22590, Suzhou, China. Association for Computational Linguistics.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. [On the continuity of rotation representations in neural networks](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753.

A Software and Packages

All experiments use PyTorch 2.3.1 with HuggingFace Transformers 4.44.2 and PEFT 0.12.0 for the LoRA adapters, on CUDA 12.1 / a single A100 80GB. The DeBERTa-v3-base parser is fine-tuned with the same stack. Decoding for all reported accuracies is greedy (temperature 0, max new tokens 512 for answer-only outputs and 1024 for trace-producing variants). The B3b tool baselines use Z3 4.13.0 together with our custom DLP geometry solver, with a per-call wall-clock timeout of 10 s for B3b Budget and 30 s for B3b Strong, and the candidate/subset budgets in §5.2. Geometric residuals and the code-separated hard validators share only the primitive geometry routines in NumPy 1.26.4 and our geomprim module. Significance tests (clustered bootstrap, cluster permutation, McNemar, Holm–Bonferroni) are implemented with SciPy 1.13.1 and statsmodels 0.14.2; the bootstrap RNG seed is 20260518. The GPT-4o-mini reference uses snapshot gpt-4o-mini-2024-07-18 with temperature 0. The frozen linear probes of Appendix R use scikit-learn 1.5.2 (train-split standardization, multinomial logistic regression with L2 and balanced class weights, ridge regression); regularization strengths are selected on the development split.

B Pooling and Geometry Anchor Token

We specify the pooling operator $\text{Pool}(h_i^{(L)}, \mathcal{A}_i)$ and the geometry anchor token used by the variable predictor (§4.2). **Anchor set.** $\mathcal{A}_i$ contains three token groups: (i) the answer anchor token (the final assistant-turn position whose hidden state also feeds answer decoding); (ii) the entity-mention tokens, namely the last sub-word position of each canonical entity span identified during canonicalization; and (iii) a single special geometry anchor token [GEO]. **Geometry anchor token.** [GEO] is a learned embedding appended once to the prompt, immediately before the answer region, as an additional input token; its embedding is a new row in the input embedding table, initialized from the mean token embedding plus small Gaussian noise ($\sigma=0.02$) and trained with the LoRA parameters. It carries no output-vocabulary entry and is never generated; it exists only to give the predictor a fixed, content-independent query position. Because [GEO] is part of the prompt, it is present at both training and inference, but since the predictor head itself is unused at standard inference (§4.3), it adds one token to the prompt and no other test-time cost. **Pooling.** Pool is a masked attention-pool with the [GEO] hidden state as the query: $\bar{h}_i = \sum_{t \in \mathcal{A}_i} \alpha_t h_{i,t}^{(L)}$ with $\alpha = \text{softmax}(h_{i,[GEO]}^{(L)\top} W_a h_{i,t}^{(L)} / \sqrt{d})$ over $t \in \mathcal{A}_i$, and $W_a \in \mathbb{R}^{d \times d}$ a small learned projection (also LoRA-adapted). When an instance has no resolved entity mentions, $\mathcal{A}_i$ falls back to the answer anchor and [GEO] only. We use attention pooling rather than mean pooling because the entity set varies in size across instances and the query-weighted form let the predictor down-weight spuriously tagged mentions; an ablation with mean pooling loses 1.3 points on D2 Mixed.

C Residual Library and Predictor Details

Provenance of r_{uv,f_2}^{gold} . r_{uv,f_2}^{gold} is the gold relative offset between objects u, v in frame f_2 , read from the *train-split* 3D ground truth (ScanNet (Dai et al., 2017) / 3DSSG (Wald et al., 2020)). It is defined only on the train split, is a property of the 3D scene rather than the answer label, and is absent at inference; when f_2 is under-specified the frame-transform residual is masked rather than supervised against a guess.

Higher-order residuals. Beyond the four first-order self-consistency residuals and the supervised

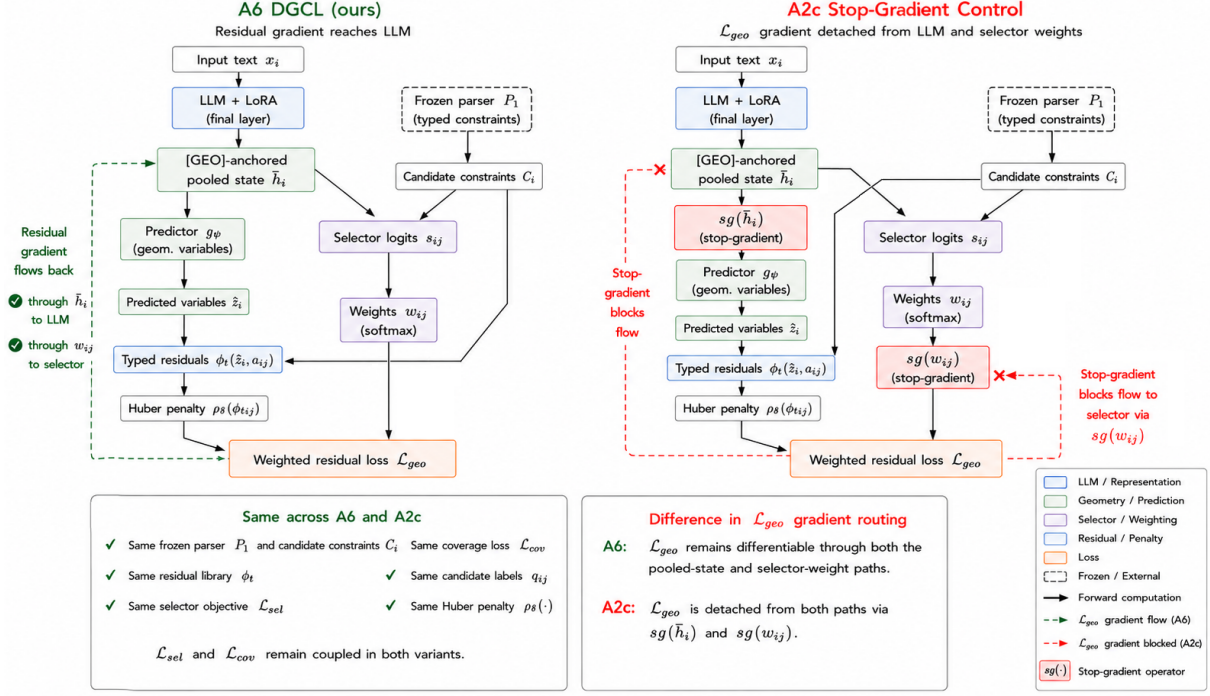


Figure 3: Detailed A6 vs. A2c computation graphs. Both share the frozen parser, candidate constraints, residual library, selector objective $\mathcal{L}_{sel}$, coverage loss $\mathcal{L}_{cov}$, and candidate labels q_{ij} . In A2c, $\mathcal{L}_{geo}$ is detached along both routes by which it could reach the LLM: $sg(\bar{h}_i)$ blocks the representation-to-geometry path and $sg(w_{ij})$ blocks the selector-weight path. In A6 both remain differentiable (§4.3). This is the detailed version of Figure 1(C).

frame-transform residual, D2 uses two consistency residuals that use no numeric gold target: *SE(3) composition*, which penalizes inconsistency among three relative transforms predicted independently for the ordered entity pairs of a triple, so that the term does not vanish by construction (translation normalized by D_i , rotation by π), and *counterfactual swap*, which averages the branch residuals of an instance and a geometrically intervened counterfactual instance under one selector weight and a single outer Huber penalty. The frame-transform residual remains the only term that uses a numeric train-split target.

Numerical stability. Removable singularities are handled explicitly: (i) every unit $(\cdot)/\text{angle}$ normalizes by $\max(\|\cdot\|_2, \epsilon)$, $\epsilon = 10^{-6}$, and degenerate zero-length segments are masked; (ii) $\angle(\cdot)$ uses arccos of a cosine clamped to $[-1 + \epsilon, 1 - \epsilon]$ (and atan2 where available); (iii) non-overlap uses a smooth differentiable 3D IoU, while the validator uses exact IoU; (iv) the instance scale lower-bounds at $\max(D_i, \epsilon_D)$, ϵ_D one percent of the median scale.

Predictor dimensions, gauge, and slot mask. The MLP is two layers (hidden 1024, GELU, Lay-

erNorm, dropout 0.05); for Llama-3.1-8B $d=4096$. The output dimension is $d_z = 2P_{\max} = 32$ for D1 ($P_{\max}=16$ points) and $d_z = 12O_{\max} = 288$ for D2 ($O_{\max}=24$ objects, each 3 center + 3 extent + 6D rotation coordinates). Before computing $\mathcal{L}_{geo}$ we fix a per-instance similarity gauge (translate centroid to origin, rotate the first canonical entity pair onto the x -axis, scale by D_i for D1; the dataset world frame for D2) to remove the coordinate null space, and a canonical slot layout (entities sorted by first mention, de-aliased, with a binary slot mask) computed from the input only, never the gold answer. Distance-ratio and angle residuals are gauge-invariant by construction.

D Per-Seed Differentiability Deltas

Table 4 reports the A6–A2c accuracy delta for each of the five seeds and five primary settings, supporting the 5/5 positive-seed claim in the main text. All 25 seed $\times$ setting deltas are positive; the per-setting means match the values in Table 1. The seed-to-seed spread is small (within ≈ 0.7 points of each setting mean), which is why the sign is stable even though the effect size is modest.

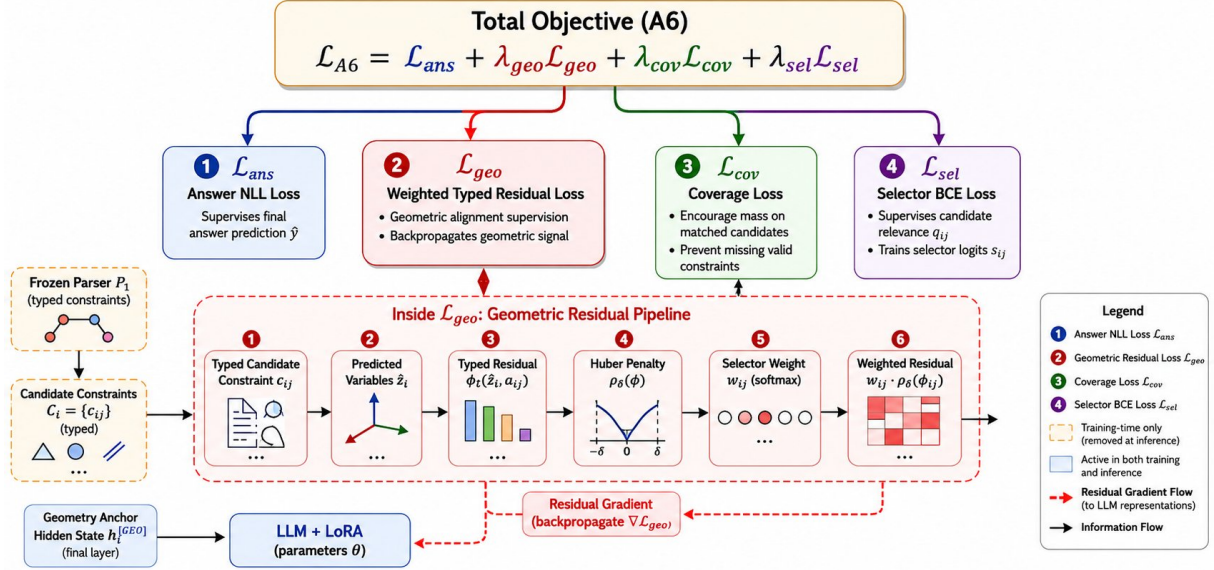


Figure 4: Decomposition of the total objective $\mathcal{L}_{A6} = \mathcal{L}_{ans} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_{cov}\mathcal{L}_{cov} + \lambda_{sel}\mathcal{L}_{sel}$. $\mathcal{L}_{ans}$ supervises the answer; $\mathcal{L}_{geo}$ is the six-step geometric residual pipeline (typed candidate $\rightarrow$ predicted variables $\rightarrow$ typed residual $\rightarrow$ Huber penalty $\rightarrow$ selector weight $\rightarrow$ weighted residual) whose gradient backpropagates into the LLM; $\mathcal{L}_{cov}$ regularizes coverage; $\mathcal{L}_{sel}$ supervises the selector via q_{ij} . $\mathcal{L}_{geo}$ is the geometric-supervision path that distinguishes A6 from A2c.

Family	Argument tuple	Residual $\phi_t(\hat{z}_i, a)$
Distance equality	(A, B, C, D)	$(\ \hat{p}_A - \hat{p}_B\ _2 - \ \hat{p}_C - \hat{p}_D\ _2)/D_i$
Fixed length	(A, B, ℓ)	$(\ \hat{p}_A - \hat{p}_B\ _2 - \ell)/D_i$
Parallel	(A, B, C, D)	$\text{cross}(\text{unit}(\hat{p}_B - \hat{p}_A), \text{unit}(\hat{p}_D - \hat{p}_C))$
Perpendicular	(A, B, C, D)	$\text{dot}(\text{unit}(\hat{p}_B - \hat{p}_A), \text{unit}(\hat{p}_D - \hat{p}_C))$
Angle equality	(A, B, C, D, E, F)	$(\angle(\hat{p}_A, \hat{p}_B, \hat{p}_C) - \angle(\hat{p}_D, \hat{p}_E, \hat{p}_F))/\pi$
Supplementary angle	(A, B, C, D, E, F)	$(\angle(\hat{p}_A, \hat{p}_B, \hat{p}_C) + \angle(\hat{p}_D, \hat{p}_E, \hat{p}_F) - \pi)/\pi$
Collinearity	(A, B, C)	$\text{cross}(\hat{p}_B - \hat{p}_A, \hat{p}_C - \hat{p}_A)/D_i^2$
Midpoint	(M, A, B)	$\ \hat{p}_M - (\hat{p}_A + \hat{p}_B)/2\ _2/D_i$
Segment addition	(A, B, C)	$(\ \hat{p}_A - \hat{p}_B\ _2 + \ \hat{p}_B - \hat{p}_C\ _2 - \ \hat{p}_A - \hat{p}_C\ _2)/D_i$
Circle radius	(O, A, B)	$(\ \hat{p}_A - \hat{p}_O\ _2 - \ \hat{p}_B - \hat{p}_O\ _2)/D_i$
Similarity ratio	(A, B, C, D, E, F)	$(\frac{\ \hat{p}_A - \hat{p}_B\ _2}{\ \hat{p}_D - \hat{p}_E\ _2 + \epsilon} - \frac{\ \hat{p}_B - \hat{p}_C\ _2}{\ \hat{p}_E - \hat{p}_F\ _2 + \epsilon})/D_i$
Left/right	(u, v, f)	$\max(0, \mu_{lr} - \text{sign}_t e_x(f)^T(\hat{c}_u - \hat{c}_v))/D_i$
Front/behind	(u, v, f)	$\max(0, \mu_{fb} - \text{sign}_t e_y(f)^T(\hat{c}_u - \hat{c}_v))/D_i$
Above/below	(u, v)	$\max(0, \mu_z - \text{sign}_t e_z(\hat{c}_u - \hat{c}_v))/D_i$
Non-overlap	(u, v)	$\max(0, \text{IoU}_{3D}(\hat{o}_u, \hat{o}_v) - \eta_{\text{overlap}})$
Frame transform	(u, v, f_1, f_2)	$\ T_{f_1 \rightarrow f_2}(\hat{c}_u - \hat{c}_v) - r_{uv, f_2}^{\text{gold}}\ _2/D_i$
SE(3) composition	(a, b, c)	$d_{\text{SE}(3)}(\mathcal{C}(T_{ab}, T_{bc}), T_{ac})$
Counterfactual swap	$(u, v, f; \pi)$	$\frac{1}{2}[\phi_t(\hat{z}_i, a) + \phi_{t^*}(\hat{z}_i^*, a^*)]$

Table 3: Typed differentiable residual families. Upper block: D1 planar-geometry residuals. Lower block: D2 frame-dependent 3D residuals. $\hat{p}_u$ is a predicted 2D point; $\hat{o}_u = (\hat{c}_u, \hat{e}_u, \hat{R}_u)$ a predicted 3D object state with center, extent, and 6D-continuity orientation (Zhou et al., 2019). $e_x(f), e_y(f), e_z$ are canonical frame axes; $\text{sign}_t \in \{+1, -1\}$ is the polarity of the typed relation (+1 for LEFT, -1 for RIGHT, and analogously for the other directional families); μ and η are frozen margins. For the two higher-order D2 families, T_{ab} is the relative transform predicted for the ordered pair (a, b) (each pair predicted separately), $\mathcal{C}$ is SE(3) composition, and $d_{\text{SE}(3)}$ is the log-map distance with translation normalized by D_i and rotation by π ; $\hat{z}_i^\pi$ and (t^π, a^π) are the prediction and expected relation on the counterfactual instance π (see text).

Setting	s13	s21	s42	s87	s100
D1 public	+4.2	+5.1	+4.4	+4.8	+4.5
D1 generated	+5.3	+4.6	+5.0	+4.7	+4.9
D2 TTO	+3.6	+4.3	+3.9	+4.1	+4.1
D2 Mixed	+4.8	+4.2	+4.6	+4.3	+4.6
D3 Zero-Touch	+5.6	+4.9	+5.3	+5.0	+5.2

Table 4: Per-seed A6–A2c accuracy deltas across the five primary settings. All 25 entries are positive (5/5 seeds per setting); per-setting means equal the deltas in Table 1.

E Decomposing A6 vs. A2c: Hidden-State Coupling vs. Selector Refinement

Table 5 reports the 2×2 design that separates the two pathways conflated by the full A6 vs. A2c contrast (§4.3): whether the residual gradient $\mathcal{L}_{geo}$ updates the LLM hidden states (via $\hat{h}_i \rightarrow \hat{z}_i$), and whether it refines the selector weights w_{ij} . All four variants share the candidate set, residual library, q_{ij} -supervised $\mathcal{L}_{sel}/\mathcal{L}_{cov}$, and training budget; they differ only in which of the two $\mathcal{L}_{geo}$ paths carry gradient. The hidden-coupling-only variant (A6-sel-detach) recovers a mean +3.60 of the full +4.64 A6–A2c gain (78%), while the selector-refinement-only variant (A2c-hidden-only) recovers only +0.84; the two are close to additive (residual interaction +0.20). This is the clean attribution: the gain is driven principally by residual gradient

reaching the *LLM hidden states*, not by residual-aware selector refinement.

Variant	hid	sel	D1p	D1g	TTO	Mix	D3
A2c (both detached)	×	×	82.1	91.0	66.4	73.6	57.9
A2c-hidden-only	×	✓	82.9	91.9	67.1	74.4	58.9
A6-sel-detach	✓	×	85.7	94.8	69.5	77.1	61.9
A6 (both)	✓	✓	86.7	95.9	70.4	78.1	63.1

Table 5: Decomposing the A6–A2c gain. “hid” = residual gradient reaches LLM hidden states ($\bar{h}_i \rightarrow \hat{z}_i$); “sel” = residual gradient refines selector weights w_{ij} . Hidden-state coupling alone (A6-sel-detach) recovers 78% of the full gain; selector refinement alone (A2c-hidden-only) recovers little. Mean Δ vs. A2c: A2c-hidden-only +0.84, A6-sel-detach +3.60, A6 +4.64.

F Frame-Transform r^{gold} Control

Table 6 is the equivalent-supervision control for r^{gold} (§6.1). EquivSup gives the same r^{gold} as A6 but only through a generic regression route; A6–frame-transform removes r^{gold} entirely.

Variant	D2 TTO	D2 Mixed
B2 FT (no r^{gold})	46.4±5.4	52.7±3.8
EquivSup (same r^{gold} , no coupling)	59.0±4.6	63.5±3.6
A6 – frame-transform (no r^{gold})	68.1±3.9	75.2±3.3
A6 full (with r^{gold})	70.4±3.6	78.1±3.2

Table 6: Equivalent-supervision control for the frame-transform target r^{gold} . Having r^{gold} without the differentiable coupling underperforms (−14.6 vs. A6 on Mixed); removing r^{gold} from A6 costs only 2.9. The gain is from the mechanism, not the target.

G Additional Results Tables

This section collects the per-dataset gate breakdown (Table 8), oracle calibration (Table 9), module-level ablation (Table 10), geometry-specificity slice (Table 11), latency (Table 12), and geometric-error (Table 13) referenced in §6.2–§6.5, plus the dataset split sizes (Table 7).

H Full Primary Results Table

Table 14 gives the complete primary answer-only results visualized in Figure 2, including all baselines, confidence intervals, and the invalid-output rate that the figure omits for legibility.

I Parser-Coverage Subset Numbers

Table 15 gives the full numeric values behind Table 2.

Block	Train	Dev	Test
D1 public	8,400	1,200	1,200
D1 generated	16,800	2,400	2,400
D2 True-Text-Only	14,700	2,100	2,100
D2 Mixed	14,700	2,100	2,100
D3 Zero-Touch	0	0	2,478
Total (A6)	54,600	7,800	10,278

Table 7: Train/dev/test split sizes. D3 has no train or dev split by design; it is evaluated only at the locked run.

Dataset	A6–A2c	A6–B2 FT
D1 public	+4.6	+29.1
D1 generated	+4.9	+31.1
D2 True-Text-Only	+4.0	+24.0
D2 Mixed	+4.5	+25.4
D3 Zero-Touch	+5.2	+21.9
Pooled mean	+4.64	+26.30

Table 8: Per-dataset deltas for the two primary gates. Each delta is the difference of the per-dataset means in Table 14.

Diagnostic condition	D1 gen	D2 Mixed
B3b Strong+ (non-oracle)	88.4	70.6
A6	95.9	78.1
Perfect P1 coverage diag.	94.2	77.9
Gold-constraint solver [†]	98.6	86.9
Timeout-free oracle [‡]	99.2	90.4

Table 9: Oracle and diagnostic calibration. Oracle rows are intentionally stronger than A6; they bound headroom but are not fair baselines. [†]information upper bound. [‡]information + compute upper bound.

Variant	D1 public	D1 gen	D2 TTO	D2 Mixed	D3	Δ mean
A6 full	86.7	95.9	70.4	78.1	63.1	0.0
A2c (sel/cov. only, no $\mathcal{L}_{\text{geo}}$ grad)	82.1	91.0	66.4	73.6	57.9	−4.6
– q_{ij} supervision (→ B2 FT)	58.0	65.1	46.9	53.0	41.5	−25.9
w/o geometric loss	58.4	64.1	47.2	52.4	41.9	−26.1
w/o variable predictor	78.9	87.8	63.1	70.5	55.6	−7.7
UniformCandidate ($w_{ij}=1/m_i$)	80.7	88.6	64.8	71.4	57.4	−6.3
DetachedSelector (sg(w_{ij}))	81.8	89.2	65.9	72.8	58.1	−5.3
No coverage regularizer	83.0	90.4	66.6	73.3	58.7	−4.4
Squared residual	84.1	92.1	68.1	75.0	60.0	−2.9
RandGeo (Inn. II falsifier)	77.6	85.2	61.9	67.4	53.1	−9.8
ParamMatch (Inn. II falsifier)	76.2	84.0	60.7	66.5	52.4	−10.9

Table 10: Module-level ablation. The geometric loss is the single largest contributor; removing the q_{ij} supervision from A2c collapses it to $\approx$ B2 FT. RandGeo and ParamMatch sit well below A6, supporting Innovation II.

J Per-Constraint and Per-Relation Breakdown

Table 16 reports D1 hard-validated answer accuracy stratified by constraint type, and Table 17 reports D2 RA_RelVal stratified by relation type. The A6 advantage over A2c is positive across every

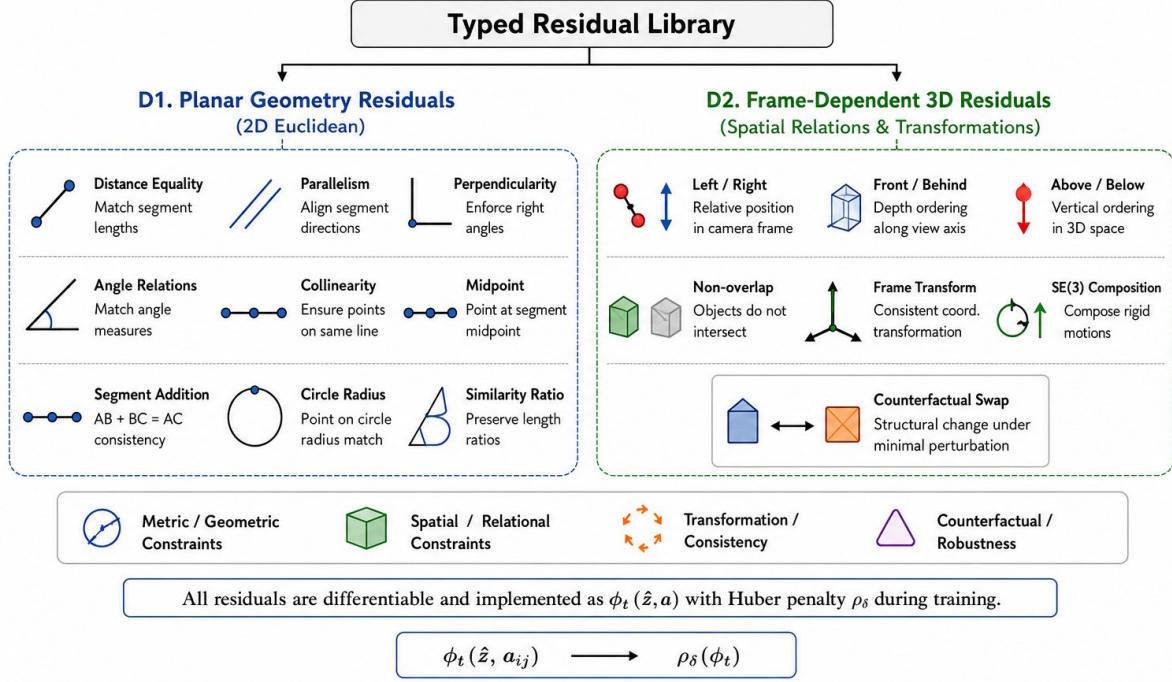


Figure 5: The typed residual library. D1 planar-geometry residuals (distance equality, fixed length, parallelism, perpendicularity, angle relations, collinearity, midpoint, segment addition, circle radius, similarity ratio) and D2 frame-dependent 3D residuals (left/right, front/behind, above/below, non-overlap, frame transform, SE(3) composition, counterfactual swap). All residuals are differentiable and evaluated under the Huber penalty ρ_δ during training; the two higher-order D2 families take the extended arguments given in Table 3. All are removed at inference. Formal definitions are in Table 3.

Subset	B2 FT	A2c	RandGeo	ParamMatch	A6	A6-RG	A6-PM
D1 generated full coverage	71.9	94.6	88.8	86.5	98.1	+9.3	+11.6
D1 public diagram available	61.8	84.7	79.9	78.6	88.2	+8.3	+9.6
D2 Mixed frame-dependent	47.3	68.8	62.5	61.9	74.9	+12.4	+13.0
D2 visibility subset	40.1	55.0	51.8	50.6	59.4	+7.6	+8.8
D3 low-lexical low-proof	35.5	53.7	47.1	46.4	60.8	+13.7	+14.4
D3 high-lexical low-proof	48.8	65.2	59.7	58.5	70.1	+10.4	+11.6

Table 11: Geometry-specificity slice. Where geometry has the most bite, A6’s lead over RandGeo and Param-Match is largest.

Method	tok/s	Latency	Solver calls
B1 Zero-Shot	78±5	1.00±0.08×	0
B2 FT	72±4	1.00±0.07×	0
B3b Budget	18±4	3.9±1.1×	5–10
B3b Strong	3.5±1.1	12.4±4.3×	up to 50
B3b Strong+	1.7±0.7	21.5±7.9×	50+
External PAL/PoT	8±3	8.4±3.2×	exec.
External Verifier	12±4	6.3±2.1×	verifier
A2c	69±4	1.05±0.08×	0
A6	68±4	1.06±0.08×	0
A6 Trace	51±5	1.39±0.13×	0

Table 12: Latency relative to B2 FT in the primary 1024–2048 input token bucket. A6 standard inference does not call P1 or any external solver.

constraint and every primary-claim relation family. Among D1 families, the largest A6–A2c deltas appear in angle equality (+6.1), supplementary an-

Subset	B MSE	A6 MSE	B AE	A6 AE
D1 generated	0.085	0.055	9.8°	6.3°
D1 public	0.113	0.074	12.7°	8.4°
D2 Mixed	0.149	0.097	15.3°	10.0°
D2 Frame Tr.	0.182	0.121	18.8°	12.3°
D3 Counterfact.	0.207	0.138	21.6°	14.2°

Table 13: Coordinate MSE and angular error (AE). Lower is better. A6 reduces both by approximately one third across all five slices.

gle (+4.7), and similarity ratio (+6.2); the smallest are in collinearity (+4.5) and distance equality (+4.9). Among D2 families, frame-dependent relations (LEFT/RIGHT, FRONT/BEHIND, FRAME TRANSFORM) and compositional SE(3) show A6–A2c gains of +5.7 to +6.6. Diagnostic-only D2 families (NEAR/FAR, INSIDE, SUPPORT, VISIBILITY, REACHABILITY) are reported for transparency but are excluded from primary D2 claim aggregation because their Type II relaxation fidelity is weaker (Table 21).

Method	D1 public HV_AnsAcc	D1 generated HV_AnsAcc	D2 TTO RA_RelVal	D2 Mixed RA_RelVal	D3 OOD_AnswerAcc	Invalid rate
B1 Zero-Shot	41.7±3.9	46.2±3.4	33.8±4.8	37.1±4.5	29.4±5.2	8.9±1.7
B2 Few-Shot CoT	49.3±4.1	51.6±3.6	38.7±4.7	44.2±4.4	33.1±5.0	7.4±1.4
GPT-4o-mini CoT†	63.9±3.8	70.8±3.4	49.6±4.5	56.4±4.2	44.3±5.4	2.1±0.6
B2 FT AnswerOnly	57.6±3.2	64.8±2.9	46.4±3.9	52.7±3.6	41.2±4.6	3.2±0.7
B3b Budget	62.4±3.8	68.6±3.4	50.9±4.3	55.8±4.1	44.7±5.4	5.3±1.3
B3b Strong	68.2±3.6	76.1±3.2	55.7±4.2	62.6±4.0	52.3±5.3	7.2±1.6
B3b Strong+	76.8±4.0	88.4±3.4	62.7±4.8	70.6±4.5	59.4±5.9	9.4±2.1
B5b Locked Constraint	65.1±3.5	72.8±3.1	53.4±4.0	60.5±3.7	49.1±5.0	4.2±0.9
B5c Serialized Graph	60.8±3.7	69.4±3.3	50.1±4.2	63.8±3.9	46.0±5.1	4.7±1.0
External PAL/PoT	59.2±4.5	67.9±4.1	47.5±4.8	54.9±4.6	43.2±5.8	8.6±2.2
External Verifier	63.8±3.9	71.6±3.5	51.8±4.4	58.4±4.1	48.6±5.5	5.8±1.4
A2c Stop-Gradient	82.1±3.1	91.0±2.7	66.4±3.8	73.6±3.5	57.9±4.9	3.4±0.8
RandGeo	77.6±3.5	85.2±3.2	61.9±4.2	67.4±4.0	53.1±5.2	3.7±0.8
ParamMatch	76.2±3.6	84.0±3.3	60.7±4.1	66.5±4.0	52.4±5.2	3.6±0.8
A6 (ours)	86.7±2.8	95.9±2.3	70.4±3.5	78.1±3.2	63.1±4.5	3.3±0.8

Table 14: Primary answer-only results (full version of Figure 2). D1 columns use HV_AnsAcc; D2 columns use RA_RelVal restricted to primary relation families; D3 uses OOD_AnswerAcc. Values are means; intervals are 95% clustered-bootstrap CIs over the seed×cluster structure (§5.4), not seed-only CIs. †GPT-4o-mini CoT is an external-API sanity reference; it does not use DGCL training and is not included in the Holm claim gates.

Coverage subset	B3b Strong	B3b S+	A6
Full P1 coverage	82.5	91.0	90.2
Partial P1 coverage	65.0	72.0	81.5
Missing P1 coverage	43.0	44.5	57.0

Table 15: Parser-coverage subset analysis (full version of Table 2). On full P1 coverage, B3b Strong+ marginally exceeds A6. A6’s aggregate advantage stems from robustness under partial or missing parser coverage.

Constraint type	B2 FT	B3b Strong	A2c	A6
Distance equality	61.4±3.5	71.9±3.1	83.8±2.8	88.7±2.5
Parallel	58.7±3.8	70.3±3.3	81.0±3.0	86.4±2.7
Perpendicular	57.9±3.9	68.0±3.5	78.6±3.2	84.9±2.9
Angle equality	53.2±4.2	64.7±3.8	75.9±3.5	82.0±3.2
Collinear	64.1±3.2	72.8±3.0	84.0±2.7	88.5±2.5
Midpoint	60.5±3.6	69.4±3.2	81.7±2.9	87.2±2.6
Circle radius	52.9±4.4	63.2±4.0	73.8±3.7	79.1±3.5
Polygon angle sum	49.6±4.6	61.5±4.2	72.9±3.9	77.4±3.7
Segment addition	56.8±4.0	67.4±3.7	78.2±3.4	83.6±3.0
Similarity ratio	47.9±4.8	59.1±4.5	69.7±4.2	75.9±3.9
Betweenness	56.1±4.1	68.6±3.7	77.4±3.4	84.0±3.1
Supplementary angle	55.4±4.0	65.8±3.7	76.9±3.5	81.6±3.2

Table 16: D1 hard-validated answer accuracy (HV_AnsAcc) by constraint type. Means with 95% clustered-bootstrap CIs (§5.4).

K Performance by Difficulty Level

Table 18 stratifies the comparison along orthogonal difficulty axes that cut across D1 and D2. The largest absolute A6 gains over B2 FT appear on hard multi-hop and very-hard counterfactual instances (+32.3 and +27.3); A6 also leads B3b Strong by +19.6 on hard multi-hop. The low-parser-coverage row shows a moderated picture: A6 retains a lead but absolute accuracy is much lower, which is consistent with DGCL being parser-assisted at training time, not parser-independent.

D2 relation type	B2 FT	B3b Str.	B6max Shortcut	A2c	A6	Claim use
LEFT / RIGHT	54.6±4.0	63.8±3.8	45.4±4.3	72.5±3.6	78.2±3.2	Primary
FRONT / BEHIND	49.3±4.5	60.9±4.1	43.2±4.5	67.1±4.0	73.7±3.6	Primary (frame audit)
ABOVE / BELOW	58.8±3.8	67.0±3.5	46.7±4.1	74.8±3.4	80.1±3.0	Primary
NEAR / FAR	51.6±4.3	60.4±4.0	44.1±4.4	68.2±4.0	73.1±3.6	Diagnostic only
INSIDE	47.5±4.8	57.8±4.5	40.2±4.8	63.6±4.3	69.0±3.0	Diagnostic only
SUPPORT	45.8±5.0	56.1±4.6	38.9±5.0	61.8±4.5	67.5±4.2	Diagnostic only
NON OVERLAP	60.4±3.6	69.5±3.3	48.6±3.9	76.3±3.1	81.8±2.8	Primary
VISIBILITY	41.2±5.4	51.8±5.0	37.0±5.2	55.9±5.1	60.5±4.8	Diagnostic only
SE(3) COMPOSITION	37.6±5.6	49.7±5.2	33.8±5.4	53.5±5.0	59.6±4.8	Primary
REACHABILITY	43.5±5.1	54.2±4.8	38.4±5.0	59.7±4.6	65.2±4.4	Diagnostic only
COUNTERFACTUAL SWAP	39.8±5.5	50.5±5.1	35.7±5.3	58.3±4.9	64.9±4.6	Primary (cert. D3 firewall)
FRAME TRANSFORM	38.7±5.7	49.4±5.3	34.6±5.5	57.6±5.0	64.2±4.7	Primary

Table 17: D2 RA_RelVal by relation type. B6max Shortcut Control replaces structured fields with shuffled placeholders. Primary D2 claim aggregation includes LEFT/RIGHT, FRONT/BEHIND (with frame audit), ABOVE/BELOW, NON OVERLAP, SE(3) COMPOSITION, COUNTERFACTUAL SWAP (with certified D3 firewall), and FRAME TRANSFORM.

Difficulty level	B2 FT	B3b Strong	A2c	A6	A6–B2 FT	A6–B3b Str.
Easy single constraint	72.8±3.1	79.6±3.0	83.4±2.8	86.2±2.5	+13.4	+6.6
Medium 2–3 constraints	57.3±3.7	69.1±3.5	78.8±3.2	84.7±2.9	+27.4	+15.6
Hard multi-hop	44.9±4.2	57.6±4.0	70.8±3.9	77.2±3.5	+32.3	+19.6
Very hard counterfactual	36.8±5.0	49.7±5.1	58.5±4.8	64.1±4.4	+27.3	+14.4
Frame ambiguous	40.6±4.9	53.2±4.7	59.9±4.5	65.0±4.3	+24.4	+11.8
Low parser coverage	39.7±5.2	44.1±5.0	50.8±4.8	54.6±4.6	+14.9	+10.5
Partial parser coverage	54.8±4.1	62.9±3.9	72.1±3.7	77.5±3.3	+22.7	+14.6
Full parser coverage	65.9±3.5	73.4±3.2	84.2±2.9	90.1±2.5	+24.2	+16.7

Table 18: Performance stratified by difficulty axis. Each row aggregates instances from both D1 and D2 that satisfy the axis predicate; rows are not mutually exclusive. Bold marks the best mean per row.

L Parser Transparency and Type II Fidelity

Table 19 reports frozen-P1 parser metrics across the four datasets. The relevant numbers for interpreting the main results are: Candidate Recall@20 ranges from 73.2% (D3) to 92.8% (D1 generated); Argument Binding Accuracy from 59.0% (D3) to 86.4% (D1 generated); and Gold Coverage@20 from 65.5% to 89.7%. The A6-stratified rows at the bottom show that A6’s accuracy degrades smoothly as parser quality drops, with the full-coverage subset reaching 98.1% on D1 generated and the missing-coverage subset still at 43.0% on D3.

Parser metric	D1 public	D1 generated	D2	D3
Candidate Recall@20	81.5±4.0	92.8±2.6	78.4±4.8	73.2±5.5
Candidate Precision@20	63.2±5.1	76.5±3.8	58.9±5.4	54.1±5.9
Argument Binding Accuracy	70.8±4.7	86.4±3.1	64.6±5.3	59.0±5.8
Gold Coverage@20	74.1±4.5	89.7±2.8	69.3±5.0	65.5±5.6
Full Coverage Grounding	61.9±5.2	80.2±3.6	55.8±5.7	50.6±6.3
A6 full-coverage subset	89.3±2.7	98.1±1.5	80.4±3.4	69.6±4.7
A6 partial-coverage subset	78.4±3.8	88.0±3.1	68.2±4.2	58.7±5.2
A6 missing-coverage subset	58.9±4.7	67.4±4.1	49.5±5.3	43.0±5.8

Table 19: Frozen P1 parser metrics and A6 accuracy stratified by coverage subset. Means with 95% clustered-bootstrap CIs (§5.4).

Table 20 breaks the aggregate D2 parser numbers down by relation type, which functions as a per-type recall/precision view (a confusion matrix

is provided in full in the supplementary material). The pattern explains the diagnostic-only designations: the families with the weakest parser recall and argument binding (VISIBILITY, SUPPORT, REACHABILITY) are exactly those excluded from the primary D2 claim aggregation, while the primary families all have recall above 72%.

D2 relation type	Recall	Prec.	Arg. bind
LEFT / RIGHT	88.2	71.0	79.5
FRONT / BEHIND	79.6	60.3	66.1
ABOVE / BELOW	90.4	73.8	82.0
NON OVERLAP	85.0	66.4	73.2
SE(3) COMPOSITION	70.2	51.5	55.8
COUNTERFACTUAL SWAP	74.4	54.9	60.0
FRAME TRANSFORM	72.6	53.2	58.4
NEAR / FAR [‡]	76.1	55.2	60.4
INSIDE [‡]	71.8	52.0	56.3
SUPPORT [‡]	68.5	49.7	53.1
VISIBILITY [‡]	62.3	44.1	47.9
REACHABILITY [‡]	66.0	47.8	51.2

Table 20: Frozen P1 per-relation-type recall, precision, and argument-binding accuracy on D2 (%). Upper block: primary families; lower block (‡): diagnostic-only families, which are also the weakest-parsed. The weakest-parsed families coincide with the diagnostic-only designation. A full confusion matrix is in the supplementary material.

Table 21 reports Type II relaxation-fidelity audit for D2 relation families. Each row reports the agreement between the differentiable training residual and the independent hard validator, the false-satisfied rate (residual says “satisfied” but validator says no), Spearman rank correlation between residual magnitude and human-rated violation severity, and human agreement on 200 sampled instances. Families with hard agreement below 90% or false-satisfied above 5% are reported as diagnostic only.

Family	Hard agree	False sat.	Spearman ρ	Human agree	Code-sep.	Claim use
LEFT / RIGHT	94.2±1.5	2.8±0.8	0.61±0.06	93.5±2.0	yes	Primary
FRONT / BEHIND	90.7±2.2	4.6±1.1	0.48±0.08	90.8±2.6	yes	Primary (frame audit)
ABOVE / BELOW	95.0±1.3	2.1±0.7	0.66±0.05	94.1±1.9	yes	Primary
NEAR / FAR	89.4±2.5	5.4±1.4	0.42±0.09	88.9±3.0	yes	Diagnostic only
INSIDE	87.8±3.0	6.1±1.6	0.39±0.10	87.2±3.4	yes	Diagnostic only
SUPPORT	86.5±3.2	6.8±1.8	0.37±0.10	86.9±3.5	yes	Diagnostic only
NON OVERLAP	92.9±1.8	3.7±1.0	0.56±0.07	92.4±2.2	yes	Primary
VISIBILITY	82.1±4.5	8.9±2.4	0.31±0.12	84.0±4.2	yes	Diagnostic only
SE(3) COMPOSITION	96.4±1.0	1.7±0.5	0.72±0.04	95.0±1.5	yes	Primary
REACHABILITY	88.6±3.1	5.8±1.6	0.44±0.09	88.2±3.4	yes	Diagnostic only
COUNTERFACTUAL SWAP	91.5±2.0	4.1±1.1	0.52±0.07	91.0±2.4	yes	Primary (cert. firewall)
FRAME TRANSFORM	92.0±2.1	3.9±1.1	0.55±0.08	91.8±2.5	yes	Primary

Table 21: Type II relaxation-fidelity audit. The training residual is differentiable; the hard validator is code-separated. “Code-sep.” means the two implementations share no functions other than the primitive geometry library. Families failing the fidelity gate (hard agreement < 90 or false-satisfied > 5) are reported only as diagnostic.

M Robustness and Shortcut Diagnostics

Table 22 reports A6’s behavior under explicit input perturbations of D2 Mixed instances. The full-input A6 reference is 78.1 RA_RelVal. Removing the structured graph (Query only) drops A6 by 38.6 points, while perturbations that preserve geometric content (paraphrase, serialized-order randomization) cause minimal drops. The B6max shortcut control replaces structured fields with shuffled placeholders; its low scores confirm that A6 is not exploiting graph-side shortcuts.

Control	Pert. A6	Δ	B6max
Full input (ref.)	78.1±3.2	—	40.2±4.8
Coord. shuffled	49.6±5.1	−28.5	40.2±4.8
Category anonymized	70.8±4.0	−7.3	38.4±5.0
Query only (no graph)	39.5±5.4	−38.6	35.1±5.3
Graph only (no query)	48.7±5.0	−29.4	42.8±4.9
Order randomized	74.6±3.8	−3.5	41.9±4.8
Counterfactual obj. swap	64.9±4.6	−13.2	35.7±5.3
Frame transformation	64.2±4.7	−13.9	34.6±5.5
Human paraphrase	74.9±4.1	−3.2	39.0±5.2

Table 22: D2 Mixed robustness controls under input perturbation. Δ is relative to A6’s full-input reference of 78.1. B6max replaces structured fields with shuffled placeholders.

Table 23 reports the leakage and sanity-stress audit triggered by the unusually high A6 scores on D1 generated and D3. Each row corresponds to a pre-registered audit; failure on any row would downgrade the corresponding result to diagnostic-only.

Stress audit	What is perturbed or checked	Observed value	Status
Train-test template overlap	Exact template ID and normalized grammar family overlap	0.0% exact overlap	Pass
Near-duplicate text audit	5-gram Jaccard nearest neighbor from test to train	median 0.24; max retained 0.63	Pass
Constraint-graph duplicate	Normalized graph hash overlap	0.0% overlap	Pass
Answer distribution audit	Test-side label and numeric value distributions	max answer freq. 6.8%	Pass
Label-shuffle sanity	Re-train A6 with shuffled train labels	D1 generated 18.7 ± 3.9	Pass
Validator blind audit	Evaluator gets predictions only, no model identity	100% evaluator-hash match	Pass
Hard-validator independence	Validators code-separated from training residuals	0 shared modules	Pass
D3 nearest-neighbor audit	Similarity to train/dev templates and graphs	100% below thresholds	Pass
Parser-candidate leakage	P1 trained only on train split labels	0 test IDs in P1 logs	Pass
Manual-debugging audit	No prompt/threshold edits after D3 access	0 post-access edits	Pass
Random-seed stability	At least 4/5 positive seed deltas on key gates	5/5 on A2c, B2 FT, Rand-Geo, ParamMatch	Pass

Table 23: Leakage and sanity-stress audit. These checks do not prove the absence of all shortcuts but they provide evidence against template reuse, duplicate graphs, answer-distribution leakage, evaluator-identity leakage, shared-validator leakage, and D3 tuning contamination.

N D3 Similarity Buckets and Firewall Certification

Table 24 stratifies D3 OOD_AnswerAcc by lexical-similarity and proof-similarity bucket. A6 leads B2 FT and B3b Strong in every retained bucket; the largest advantage is in low-lexical low-proof

(+25.3 over B2 FT), which is the bucket farthest from anything seen during training. The “removed by audit” row contains 274 instances flagged by the nearest-neighbor audit and not included in any reported D3 metric.

Similarity bucket	Count	B2 FT	B3b Str.	A6
Low lex. low proof	642	35.5±5.4	47.9±5.6	60.8±5.0
High lex. low proof	618	48.8±4.7	57.4±5.0	70.1±4.2
Low lex. high proof	631	38.2±5.1	52.8±5.4	63.4±4.8
High lex. high proof	587	55.0±4.4	64.7±4.6	74.8±3.9
Removed by audit	274	—	—	—

Table 24: D3 OOD_AnswerAcc by similarity bucket. The removed-by-audit row is excluded from all reported D3 metrics.

Table 25 lists the pre-registration file together with the six D3 firewall artifacts (seven artifacts in total) and the (abbreviated) SHA-256 digests of the sealed bundle. The protocol is a strict temporal ordering, fixed before any D3 example is seen: (1) the pre-registration file `prereg_gates.md` — listing the seven gates, their direction, metric, test, and threshold (§5.4) — is written and hashed (`sha256:prereg-1a7f...e09c`); (2) the D3 generator and template registry are frozen and hashed; (3) all models are trained and tuned using only D1/D2 train and dev splits, with a signed no-tuning statement covering D3; (4) the nearest-neighbor audit removes 274 near-duplicate instances and the retained set is hashed; (5) the locked evaluation is run once on the retained D3 set, after which no prompt, threshold, parser, or solver edit is permitted. Any post-access edit would change the access-log hash and void the zero-touch status. The full-length digests, the audit scripts, and the access log are retained with the sealed artifacts, so every hash can be recomputed and the temporal ordering confirmed.

O Absolute Latency and Efficiency-Audit Coverage

Table 26 reports absolute latency in the primary 1024–2048 input-token bucket on batch size 1 greedy decoding. Parser time is zero for A6 standard inference because P1 is not used at standard inference; tool baselines report parser and solver interface time separately from GPU generation time. The only test-time cost A6 adds over B2 FT is the single [GEO] prompt token (Appendix B); this one-token overhead is already included in the A6 row here (it is part of the measured input length and is

Firewall artifact	Required evidence	Digest (abbrev.)	Status
<code>prereg_gates.md</code>	Seven gates, direction, metric, test, threshold	<code>prereg-1a7f...e09c</code>	Pre-registered
<code>d3_generator_hash.txt</code>	Generator commit and environment hash	<code>d3gen-9f42...4d31</code>	Certified
<code>d3_template_registry.jsonl</code>	Template/family IDs, bucket, held-out reason	<code>d3tmpl-4c18...03f5</code>	Certified
<code>d3_access_log.csv</code>	Timestamped role-based access log	<code>d3access-a61e...98e2</code>	Zero-touch
<code>d3_no_tuning_statement.md</code>	Signed no-tuning statement	<code>d3notune-72b9...6d88</code>	Zero-touch
<code>d3_nearest_neighbor_audit.jsonl</code>	Nearest-neighbor similarity and removal	<code>d3nn-33a5...2d77</code>	OOD-sim.
<code>d3_final_retained_hashes.txt</code>	Retained-example hashes after audit	<code>d3retain-5b70...4e0f</code>	Final-set

Table 25: D3 firewall and pre-registration artifacts. Each is sealed under SHA-256 before the locked evaluation run; digests are abbreviated here and recorded in full alongside the audit scripts, which allow independent recomputation. Status: Zero-touch = zero-touch certified; OOD-sim. = OOD-similarity certified; Final-set = final-set certified.

the main reason A6 is $1.06\times$ rather than exactly $1.00\times$ B2 FT), so no separate accounting is needed.

Method	Tok in	Tok out	Tok. ms	Parser ms	Gen ms	Solver ms	Total ms	p95 ms	Rel.
B2 FT	1024–2048	214±46	38±6	0	2102±150	0	2140±150	2440	1.00×
A6	1024–2048	216±44	39±6	0	2229±171	0	2268±171	2590	1.06×
A2c	1024–2048	215±45	39±6	0	2208±165	0	2247±165	2550	1.05×
B3b Budget	1024–2048	219±50	41±7	165±35	2400±420	5730±1800	8336±2300	13270	3.90×
B3b Strong	1024–2048	224±53	42±8	172±37	2550±530	23800±18500	26564±9200	59920	12.41×
B3b Strong+	1024–2048	226±56	43±8	188±42	2620±610	43160±16000	46011±16900	96300	21.50×
A6 Trace	1024–2048	382±78	40±7	0	2935±275	0	2975±275	3680	1.39×

Table 26: Absolute latency breakdown in the primary 1024–2048 input-token bucket. Batch size 1, greedy decoding, single A100 80GB. Mean ± standard deviation over 500 instances. Relative latency uses B2 FT as the reference.

Table 27 lists the required reporting coverage across all input-token buckets. Tokenization-included and tokenization-excluded variants are both reported in the supplementary material for every bucket; parser time remains zero for A6 standard inference across all buckets.

Token bucket	Location	Claim use
0–512	Suppl.	Short-prompt latency
512–1024	Suppl.	Medium-prompt latency
1024–2048	Table 26	Solver-free gate
2048–4096	Suppl.	Long-context latency
Max-length stress	Suppl.	Diagnostic stress

Table 27: Efficiency audit coverage by input-token bucket. The 1024–2048 bucket is used in the main solver-free efficiency gate; the other buckets are reported in the supplementary material with absolute latency, $p50$, $p95$, output tokens, tokenization time, and generation time.

Table 28 breaks down the failure modes of the B3b tool family, which are the source of both their latency variance and their invalid-output rate. As the candidate cap and solver budget grow from Budget to Strong+, raw coverage of the gold constraint rises, but so do solver timeouts, unsatisfiable-subset

rejections, and budget-exhaustion events; each of these resolves to an invalid output that counts as failure under our protocol. This is the systematic version of the point in §7: enlarging the search does not monotonically help once the parser is non-oracle, because a larger candidate set also admits more inconsistent subsets.

Failure mode (%)	Budget	Strong	Strong+
Solver timeout	2.1	5.4	8.6
Unsatisfiable-subset rejection	1.8	4.7	7.5
Search-budget exhaustion	1.0	3.3	5.9
Malformed / unparseable output	0.4	1.1	1.6
Total invalid rate	5.3	7.2	9.4
Gold-constraint coverage	61.0	71.5	78.9

Table 28: B3b tool-family failure modes. Larger candidate caps and solver budgets raise gold-constraint coverage but also raise timeout, rejection, and budget-exhaustion rates; all of these resolve to invalid outputs (failures). Total invalid rate matches the “Invalid rate” column of Table 14.

P Backbone Generality: Qwen2.5-7B Reproduction

We reproduce the A6 vs. A2c comparison on the Qwen2.5-7B-Instruct backbone (Qwen Team, 2024) using the same LoRA configuration (rank 16 on q, k, v, o , gate, up, down), the same training budget (3 epochs, 2,562 optimizer updates), and the same five seeds. Table 29 summarizes the result. The differentiable-alignment effect is positive in all five primary settings on Qwen2.5-7B with mean +3.91 and 5/5 positive-seed counts per setting (Holm-corrected $p < 0.001$). The absolute deltas are smaller than on Llama-3.1-8B (+4.64), which is consistent with the hypothesis that the magnitude of the DGCL benefit scales with the backbone’s base spatial competence.

Q Qualitative Case Studies

Two representative examples from the locked test set illustrate the mechanism behind the aggregate numbers; they make the pattern concrete, not substitute for it.

D1 success: parallelism forces a consistent angle. A D1 public item presents a trapezoid $ABCD$ with $AB \parallel CD$ and asks for an interior angle given two others. The answer-only baseline (B2 FT) returns a value that is numerically close but violates the parallel constraint (co-interior angles do not sum

Setting	A2c	A6	Δ
D1 public	78.5 $\pm$ 3.4	82.2 $\pm$ 3.1	+3.7
D1 generated	87.6 $\pm$ 3.0	91.8 $\pm$ 2.6	+4.2
D2 True Text Only	62.7 $\pm$ 4.0	66.3 $\pm$ 3.7	+3.6
D2 Mixed	70.1 $\pm$ 3.7	74.0 $\pm$ 3.4	+3.9
D3	54.3 $\pm$ 5.1	58.4 $\pm$ 4.8	+4.1
Mean	—	—	+3.91

Table 29: Qwen2.5-7B-Instruct reproduction of the A6 vs. A2c differentiability comparison. All matched conditions (LoRA config, training budget, seeds, evaluation protocol) are identical to the Llama-3.1-8B main results. The mean A6–A2c delta is +3.91, vs. +4.64 on the Llama backbone; both backbones pass all seven pre-registered gates.

to 180°): it passes surface plausibility but fails the hard validator. A2c predicts geometric variables through a detached head with low parallel residual, yet its final *answer* token is unchanged from B2 FT because the geometry head cannot influence the LLM’s answer representation. A6, with the parallel and supplementary-angle residuals back-propagating into the LLM hidden states, produces an answer consistent with the co-interior relation and passes the validator — matching the per-type pattern in Table 16, where A6’s advantage over A2c is largest on the angle and parallel families.

D2 partial coverage: A6 answers where the solver has nothing to solve. A D2 Mixed item asks for a LEFT/RIGHT relation under an allocentric frame that requires composing a frame transform with a center comparison. The frozen parser P1 extracts the two object references but *misses* the frame-transform constraint, so its candidate set is incomplete. B3b Strong+ therefore searches a candidate set lacking the gold constraint; its solver either times out on inconsistent subsets or returns a frame-naïve answer, and the instance falls in the “Missing P1 coverage” row where B3b Strong+ scores 44.5 (Table 15). A6 has no parser at inference and does not depend on the missing candidate: the frame-transform residual was active during training on covered instances, and the learned representation generalizes to this partially-covered instance, yielding the correct allocentric relation — the concrete form of the robustness claim in §6.3.

R Exploratory Supplemental Analyses

This appendix reports the three post-hoc analyses summarized in §6.6: controlled corruption of the training-time candidate sets (Table 30), masking

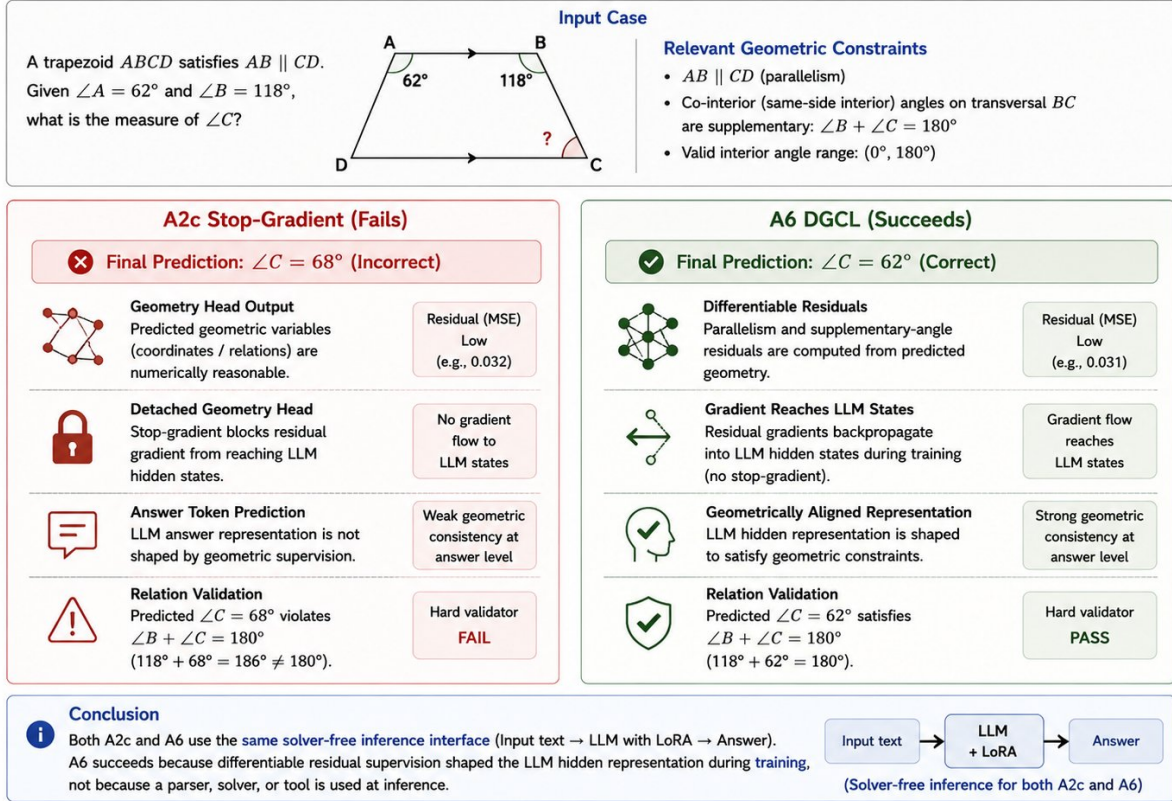


Figure 6: Qualitative D1 case. For a trapezoid $ABCD$ with $AB \parallel CD$, $\angle A=62^\circ$, $\angle B=118^\circ$, the co-interior relation forces $\angle C=62^\circ$. A2c predicts $\angle C=68^\circ$ (fails the hard validator, $118+68=186 \neq 180$) even though its detached geometry head has low residual, because the residual gradient never shapes the answer representation. A6 predicts $\angle C=62^\circ$ (passes). Both use the *same* solver-free inference (text $\rightarrow$ LLM+LoRA $\rightarrow$ answer); A6 succeeds because differentiable residual supervision shaped the representation during training, not because any tool is used at inference.

of auxiliary supervision on selected target families (Table 32), and frozen linear probes of the shared pre-generation state (Table 33).

R.1 Setup and Statistics

All three analyses are post-hoc and exploratory. We report pointwise 95% normal-approximation intervals centered on the paired estimate, with standard errors from 10,000 paired cluster-bootstrap resamples; D1 source groups and D2 scene IDs are the resampling units (§5.4). These intervals are not multiplicity-adjusted and are separate from the seven pre-registered Holm-corrected gates (Table 1).

For each condition in Table 30, A6 and A2c were trained on the same 54,600-instance multi-dataset pool (which includes D2 Mixed), so D1 public and D2 True-Text-Only (TTO) evaluate the same jointly trained checkpoint rather than dataset-specific models; degradation across the two datasets is therefore correlated by design. Table 30 used 50 matched LoRA runs and Table 32 used 70, for 120 additional

runs in total; all used the 2,562-update schedule of §4.6 on one A100 80GB with no new hyperparameter search (about 0.8k A100-hours). The probe analysis reused frozen checkpoints. We report D1 public and D2 TTO: D1 generated is primarily a typed-residual-alignment test (§5.1), and evaluating new post-hoc checkpoints on D3 would break its zero-touch protocol (Appendix N).

R.2 Sensitivity to Corruption of Training Candidates

For each training example we normalize, deduplicate, rank, and freeze the parser’s original top-20 candidates; corruption is applied only within that fixed set. Deleted candidates are not replaced by lower-ranked candidates, so some corrupted examples contain fewer than 20 candidates. TypeFlip-25 and BindCorrupt-25 act on the same per-seed candidate mask: from the eligible universe — clean gold-matched candidates in the frozen top-20 whose type has a partner in a fixed confusion map and which admit a legal rebinding — we sample 25%

without replacement, stratified by residual family. TypeFlip replaces a candidate’s type or polarity by its partner under a fixed involution over five schema-preserving pairs (parallel $\leftrightarrow$ perpendicular, angle equality $\leftrightarrow$ supplementary angle, left $\leftrightarrow$ right, front $\leftrightarrow$ behind, above $\leftrightarrow$ below), leaving arguments, parser confidence, and rank unchanged; candidate types outside this involution are not eligible for TypeFlip. BindCorrupt keeps the type and replaces exactly one entity-valued argument with another valid canonical entity of the same role, so that the canonical form changes. Corrupted candidates are validity-checked, canonicalized, deduplicated, and re-matched against train-split gold to recompute q_{ij} ; the operator does not force $q_{ij}=0$. Confidence ranking is not recomputed and lower-ranked candidates are never promoted. Both conditions affect effective recall and precision; they are not precision-only tests, and the study contains no precision-only condition that adds spurious candidates while holding recall fixed. Each corruption row uses one jointly trained checkpoint per model variant and seed; A6 and A2c use the same corrupted candidates, seeds, data order, hyperparameters, and training budget. Clean values are copied from Table 14; clean intervals were not recomputed.

R.3 Masking Auxiliary Supervision on Target Families

For each target family group, the target-family examples retain their inputs, answer labels, and $\mathcal{L}_{ans}$, while $\mathcal{L}_{geo}$, $\mathcal{L}_{sel}$, and $\mathcal{L}_{cov}$ are set to zero on those examples using the original fixed loss denominator (suffix “HO”). Because the masked examples remain in the fixed auxiliary-loss denominator, the Full–HO column reflects both the removal of target-family auxiliary information and the lower effective auxiliary-loss magnitude; it should not be read as a pure residual-family effect. The mask is applied jointly across the D2 training pool, i.e., to the target family’s TTO and Mixed instances alike, so no same-family auxiliary supervision leaks from Mixed to TTO. The experiment asks whether auxiliary learning from the remaining examples benefits the masked target slice; it does not test a relation or family outside the residual library. D1 public Angle relations is the disjoint union of *angle equality* and *supplementary angle* ($N=198$); D2 TTO Compositional is the disjoint union of *frame transform* and *SE(3) composition* ($N=470$). Grouped scores are recomputed from pooled per-instance predictions rather than averaged from rounded subtype scores;

macro rows are unweighted means across families, with deltas calculated before display rounding.

R.4 Frozen Linear Probes

We train linear probes on the final-layer hidden state at the last chat-template token of the assistant prefix, immediately before the first generated answer token. This token follows [GEO] in the prompt but is not the [GEO] state itself. The probes do not use entity pooling, predictor outputs, candidates, q_{ij} , residuals, gold-answer tokens, or generated-answer tokens. For each frozen checkpoint and training seed, the probe is fitted on the original training split, regularization is selected on the original development split, and the locked test split is used once for reporting. Hidden dimensions are standardized with train-split statistics. Family and relation probes use L2-regularized multinomial logistic regression with balanced class weights; the coordinate probe uses ridge regression with canonicalized coordinate targets (scikit-learn 1.5.2, Appendix A), where NRMSE uses the same per-instance scale D_i as the training similarity gauge and unavailable coordinate slots are masked. Targets come from the fixed gold family/relation registry and canonical geometric labels. Reported contrasts are paired on test instances, averaged across the five model-training seeds, and evaluated with the same cluster-bootstrap units used elsewhere. B2 FT is included as a descriptive reference and uses its standard answer-only prompt; the four DGCL variants use the same prompt, with [GEO] in the same position. These probes show higher linear decodability, not causal use in answer generation.

Training-candidate condition	A2c	A6	A6–A2c [95%]	A6 vs. clean [95%]
<i>D1 public (HV_AnsAcc)</i>				
Clean	82.1	86.7	+4.6	reference
Gold-match Drop-25	78.9	83.4	+4.5 [3.3, 5.7]	–3.3 [–4.7, –1.9]
Gold-match Drop-50	73.9	77.5	+3.6 [2.0, 5.2]	–9.2 [–11.1, –7.3]
TypeFlip-25	75.6	78.8	+3.2 [1.2, 5.2]	–7.9 [–10.0, –5.8]
BindCorrupt-25	74.2	77.0	+2.8 [0.5, 5.1]	–9.7 [–12.0, –7.4]
Whole-family Drop-100	79.0	82.5	+3.5 [1.6, 5.4]	–4.2 [–6.0, –2.4]
<i>D2 TTO (RA_RelVal)</i>				
Clean	66.4	70.4	+4.0	reference
Gold-match Drop-25	63.0	66.8	+3.8 [2.5, 5.1]	–3.6 [–5.2, –2.0]
Gold-match Drop-50	58.1	61.0	+2.9 [1.1, 4.7]	–9.4 [–11.7, –7.1]
TypeFlip-25	60.0	62.6	+2.6 [0.3, 4.9]	–7.8 [–10.2, –5.4]
BindCorrupt-25	58.7	60.8	+2.1 [–0.2, 4.4]	–9.6 [–12.3, –6.9]
Whole-family Drop-100	62.8	65.8	+3.0 [0.6, 5.4]	–4.6 [–6.8, –2.4]

Table 30: Exploratory sensitivity to controlled corruption of training candidates (post-hoc; not part of the pre-registered gates). Intervals are pointwise 95% normal-approximation intervals with standard errors from 10,000 paired cluster-bootstrap resamples (D1 source groups, D2 scene IDs; 5 seeds); they are not multiplicity-adjusted. Whole-family Drop-100 removes angle relations on D1 and horizontal (left/right) relations on D2. Clean values are copied from Table 14; clean intervals were not recomputed. The D2 BindCorrupt-25 A6–A2c interval includes zero (3/5 positive seeds) and is reported as inconclusive.

Dataset / omitted family (<i>N</i>)	Clean	Drop-100	Target Δ	Non-target Δ	Excess [95%]
D1 public / angle relations (angle equality + supplementary angle), <i>N</i> =198	81.80	65.40	16.40	1.79	+14.61 [10.55, 18.67]
D2 TTO / horizontal (LEFT + RIGHT), <i>N</i> =380	78.20	65.00	13.20	2.70	+10.50 [6.75, 14.25]

Table 31: Effect of whole-family omission on the targeted slice (exploratory). Clean and Drop-100 are A6 scores on the targeted slice; Δ is the degradation; target-specific excess is the target-slice degradation minus the non-target-slice degradation. The bootstrap resamples the union of source groups or scene IDs, preserving the pairing when a cluster contributes to both slices.

Dataset / masked target group	<i>N</i>	Full A6	A2c-HO	A6-HO	A6-HO–A2c-HO [95%]	A6 Full–HO [95%]
D1 public / similarity ratio	88	75.9	59.5	62.1	+2.6 [0.2, 5.0]	+13.8 [10.8, 16.8]
D1 public / angle relations (angle equality + supplementary angle)	198	81.8	64.5	67.4	+2.9 [0.8, 5.0]	+14.4 [11.8, 17.0]
D1 public / circle radius	91	79.1	62.6	64.3	+1.7 [–0.9, 4.3]	+14.8 [11.4, 18.2]
D1 public macro	—	78.9	62.2	64.6	+2.4 [1.0, 3.8]	+14.3 [12.6, 16.0]
D2 TTO / horizontal (LEFT + RIGHT)	380	78.2	62.5	64.9	+2.4 [0.3, 4.5]	+13.3 [10.8, 15.8]
D2 TTO / depth (FRONT + BEHIND)	320	73.7	57.8	59.5	+1.7 [–0.7, 4.1]	+14.2 [11.5, 16.9]
D2 TTO / vertical (ABOVE + BELOW)	410	80.1	65.6	68.0	+2.4 [0.5, 4.3]	+12.1 [9.8, 14.4]
D2 TTO / compositional (FRAME TRANSFORM + SE(3) COMPOSITION)	470	61.9	45.5	46.5	+1.0 [–1.8, 3.8]	+15.4 [12.3, 18.5]
D2 TTO macro	—	73.5	57.9	59.7	+1.9 [0.7, 3.1]	+13.8 [12.2, 15.4]

Table 32: Exploratory target-family auxiliary masking (post-hoc). HO: target examples and $\mathcal{L}_{ans}$ remain in training while $\mathcal{L}_{geo} = \mathcal{L}_{sel} = \mathcal{L}_{cov} = 0$ on those examples. Intervals as in Table 30. Three of seven family-level A6-HO–A2c-HO intervals include zero; the result is limited within-library transfer, not evidence about unseen constraint types.

Checkpoint	D1 F1 $\uparrow$	D1 NRMSE $\downarrow$	D2 F1 $\uparrow$	D2 axis $\uparrow$
B2 FT	60.20	0.351	55.10	62.70
A2c	71.60	0.306	68.40	72.60
A2c-hidden-only	72.45	0.301	69.10	73.35
A6-sel-detach	75.55	0.283	72.10	76.85
A6	77.10	0.279	73.70	77.70
Contrast	D1 family F1 Δ	D1 coord. NRMSE Δ	D2 relation F1 Δ	D2 axis bal. acc. Δ
A6–A2c	+5.50 [3.20, 7.80]	–0.027 [–0.041, –0.013]	+5.30 [2.90, 7.70]	+5.10 [2.70, 7.50]
A6-sel-detach–A2c (hidden coupling only)	+3.95 [1.85, 6.05]	–0.023 [–0.036, –0.010]	+3.70 [1.35, 6.05]	+4.25 [2.05, 6.45]
A2c-hidden-only–A2c (selector refinement only)	+0.85 [–0.45, 2.15]	–0.005 [–0.012, 0.002]	+0.70 [–0.60, 2.00]	+0.75 [–0.65, 2.15]
A6–A6-sel-detach	+1.55 [–0.10, 3.20]	–0.004 [–0.011, 0.003]	+1.60 [–0.15, 3.35]	+0.85 [–0.60, 2.30]

Table 33: Frozen-probe scores on D1 public and D2 TTO (exploratory): D1 family macro-F1 and coordinate NRMSE, D2 relation macro-F1 and 6-way axis balanced accuracy. All checkpoints are frozen and only the probes are trained; the probe reads the shared final-layer pre-generation state, not [GEO] or an auxiliary output. Variants as in Table 5. Lower panel: paired contrasts; brackets are pointwise 95% normal-approximation intervals with standard errors from 10,000 paired cluster-bootstrap resamples, not multiplicity-adjusted. The hidden-coupling path accounts for most of the A6–A2c difference; selector-only contrasts include zero.