STRUCTALIGN: LEARNING WHAT TO PRESERVE AND WHAT TO COMPLEMENT FOR SKETCH-TEXT 3D RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieving textured 3D objects from freehand sketches and text requires combining complementary yet unequally reliable cues. Sketches provide strong geometric constraints through contours, proportions, and part relationships, whereas text supplies category and appearance information but may be incomplete, ambiguous, or inconsistent with the sketch. Existing fusion methods often lack explicit protection of sketch structure, allowing textual semantics to distort geometry-sensitive representations. We present StructAlign, which treats the sketch representation as a geometric anchor and uses multimodal large language model features to complement missing semantics. It constructs a foreground-aware sketch representation and selectively injects locally relevant semantic features through cross-modal attention and gated residual updates. A foreground Structural KL constraint limits changes in the relative importance of key stroke regions. To align semantic completion with downstream retrieval, we further introduce a shape–texture graded ranking reward and retrieval-drift-calibrated GRPO with completion-specific adaptive clipping. We also construct ST-Tex3D v2, containing 15,600 textured 3D instances with aligned sketches, retrieval-oriented descriptions and texture variants. Experiments demonstrate the effectiveness of explicitly modeling asymmetric modality reliability for fine-grained multimodal retrieval.

1 INTRODUCTION

Retrieving textured 3D assets from sketches and text provides an intuitive interface for content creation and interactive design. The two modalities offer complementary cues: sketches describe geometry through contours, proportions, and part relationships, whereas text specifies category and appearance attributes such as material, color, and texture. Combining them is therefore important for retrieving 3D instances that are correct in both shape and appearance. However, sketch and text differ substantially in information density and representation granularity. Freehand sketches are spatially sparse, with useful cues concentrated in a small number of foreground strokes and structural regions, while text features encode richer and more distributed semantics. Existing multimodal retrieval methods typically combine the two modalities through score fusion, feature concatenation, weighted summation, global attention, or shared embedding alignment Schlachter et al. (2022a); Chowdhury et al. (2023a); Koley et al. (2024a); Lee et al. (2025). Although effective, these methods generally do not explicitly account for this modality mismatch. Direct global fusion may dilute sparse geometry-sensitive sketch responses, while global aggregation cannot determine which semantic information is relevant to individual sketch regions. As a result, the joint representation may not sufficiently preserve local contours, part boundaries, and proportions while incorporating appearance information. This can lead to retrieved models with mismatched shapes or textures, as illustrated in Fig. 1.

To address this problem, we propose **StructAlign**, a structure-preserving semantic completion framework for sketch–text textured 3D retrieval. StructAlign treats the sketch representation as a geometric anchor and uses MLLM semantics to complement missing or ambiguous category, part, and appearance information. It first constructs a foreground-aware sketch representation that retains informative stroke tokens while aggregating background context. Locally relevant MLLM features are then selectively injected into foreground tokens through cross-modal attention and gated residual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

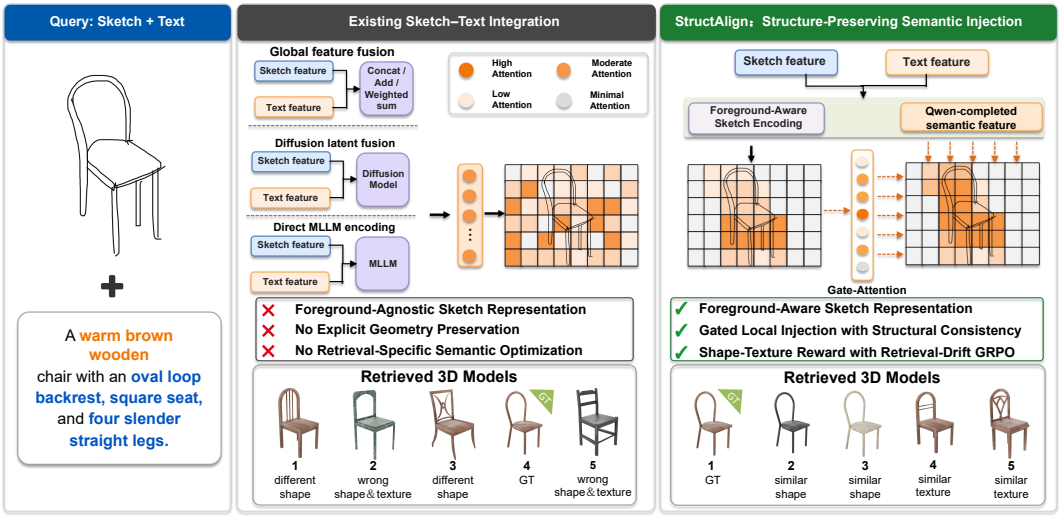

Figure 1: Existing sketch–text fusion methods may allow textual semantics to override reliable sketch geometry, resulting in shape or texture retrieval errors. StructAlign uses foreground-aware encoding and gated local semantic injection to preserve sketch-derived geometric structure while complementing missing category, part, and texture information.

updates. A foreground structural consistency constraint further limits changes in the relative importance of key stroke regions. To align semantic completion with downstream retrieval, we introduce a shape–texture graded ranking reward that considers exact-target ranking, geometric discrimination, and texture discrimination. The resulting dense feedback is used to optimize the completion policy through group-relative policy optimization. We further stabilize training at two levels: foreground Structural KL constrains representation drift caused by semantic injection, while retrieval-drift calibration measures changes in the induced candidate distribution and assigns completion-specific GRPO clipping ranges. These constraints preserve sketch structure in feature space and stabilize retrieval behavior in task space. We also construct ST-Tex3D v2, containing 15,600 textured 3D instances with aligned sketches and text descriptions. Experiments on chair and lamp categories show that StructAlign substantially outperforms single-modal, conventional multimodal fusion, and MLLM-based retrieval baselines. The main contributions are summarized as follows:

- We propose a structure-preserving semantic injection mechanism that selectively complements foreground sketch tokens while preserving sketch-derived geometric structure.
- We introduce a shape–texture graded ranking reward that provides dense retrieval supervision for optimizing MLLM-generated semantic completions.
- We develop a dual-level stability strategy that jointly constrains structural drift through foreground Structural KL and retrieval drift through adaptive GRPO clipping, preserving both sketch structure and downstream retrieval behavior.

2 Related Work

Sketch and Text for 3D Retrieval. Sketch-based 3D retrieval has progressed from shared embedding and cross-domain alignment to fine-grained matching and generalization under limited supervision Wang et al. (2015); Chen & Fang (2018); Qi et al. (2021); Chowdhury et al. (2023a); Wang & Zhou (2023), but remains largely geometry-oriented. Recent methods further reduce paired supervision through image-based pivoting or prototype alignment for unseen categories. Text-based methods have evolved from global and sequential text–shape alignment to part–word matching and trimodal contrastive learning Chen et al. (2018); Han et al. (2019); Tang et al. (2023); Ruan et al. (2024). Although weighted CLIP features have enabled zero-shot retrieval with sketches, text, and reference images Schlachter et al. (2022b), existing methods typically fuse sketches and text sym-

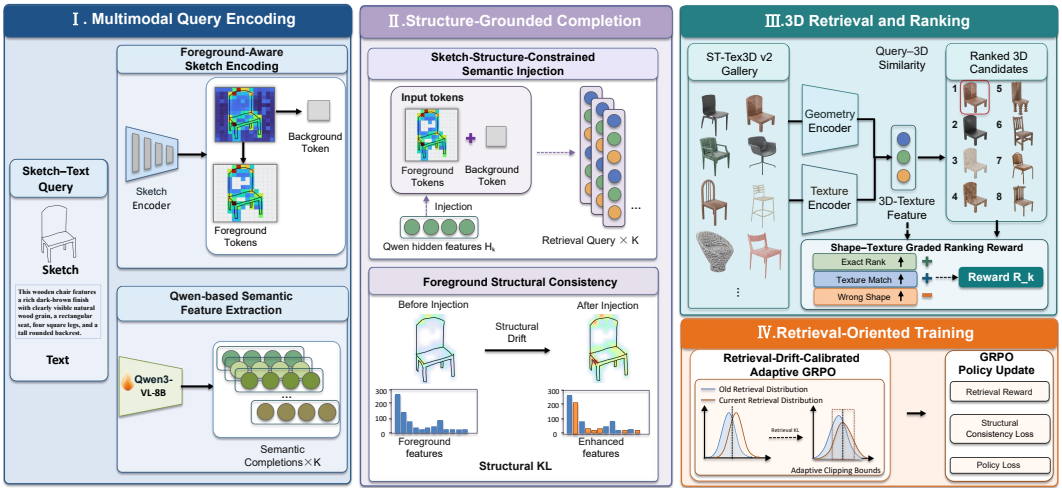

Figure 2: Overview of StructAlign. The framework performs foreground-aware sketch encoding, selectively injects Qwen3-VL semantic features through gated local attention under a structural consistency constraint, and matches the enhanced query against geometry–texture features of the 3D gallery. The completion policy is optimized using a shape–texture ranking reward and retrieval-drift-calibrated GRPO.

metrically, without preserving sketch-derived structure or finely aligning sketch-specified geometry with text-specified appearance.

Sketch-Text Fusion. Sketch–text fusion is commonly performed through global feature aggregation, structured token interaction, or generative conditioning. Global methods combine independently encoded features Sangkloy et al. (2022); Schlachter et al. (2022b), while structured approaches use attention or pseudo-word tokens for finer composition Chowdhury et al. (2023b); Wu et al. (2023); Koley et al. (2024b). SceneTrilogy disentangles shared and modality-specific representations, whereas Sketch2Word maps sketches into compositional pseudo-word embeddings. In 3D generation, language-aligned priors complement sketches for part-aware shape synthesis and colored mesh generation Lee et al. (2025); Zang et al. (2025); diffusion-based methods instead inject sketch structure through control branches or adapters under text guidance MaungMaung et al. (2023); Zhang et al. (2023); Mou et al. (2024); Zuo et al. (2024). However, these methods rarely distinguish the roles of sketches in specifying geometry and text in describing material and texture for textured 3D retrieval.

MLLMs for Sketch Understanding and Reasoning. MLLM-based sketch research mainly focuses on visual reasoning, structured understanding, and retrieval. Sketches have been used as intermediate visual states to support complex reasoning Hu et al. (2024). Other studies employ MLLMs to convert workflow and scientific sketches into structured semantic or executable representations Bechard et al. (2026); Latif et al. (2026). More directly related to retrieval, recent work either trains sketch-specific language models or introduces MLLM-generated semantics to enhance fine-grained sketch representations Gupta et al. (2026); Li et al. (2026a). However, existing methods rarely combine independent user text with retrieval feedback to model geometric and appearance constraints.

3 Method

3.1 Task Definition and Method Overview

As shown in Fig. 2, given a freehand sketch S and a textual description T, we aim to retrieve a target object G^+ from a repository of textured 3D models $\mathcal{G} = \{G_j\}_{j=1}^{N}$. The target object should be geometrically consistent with S and satisfy the appearance description in T, including color, material, and texture attributes. We feed the sketch–text pair into Qwen3-VL-8B and incorporate its multimodal features through a **sketch-structure-constrained semantic injection mechanism**, rather than directly using these features to replace the global query representation produced by the

sketch encoder. Given a query $x = (S, T)$, Qwen samples K candidate semantic completions $\{y_k\}_{k=1}^{K}$. Local cross-modal attention and a gating module then selectively inject sketch-relevant semantic information into the sketch representation. Throughout this process, the original sketch features serve as a geometric anchor, while Qwen is used only to supplement missing or ambiguous information about object categories, parts, and textures. Leveraging the ranking feedback provided by the retrieval results, we construct a task-level reward and adopt GRPO to optimize the semantic completion policy. We further introduce two KL constraints: **Structural KL** limits the structural deviation caused by semantic injection, thereby preserving the sparse geometric organization of the sketch; **Retrieval KL** constrains the discrepancy between the retrieval distributions induced by the old and updated policies and accordingly adjusts the GRPO clipping range in an adaptive manner.

3.2 Foreground-Aware Sketch Encoding

Freehand sketches usually contain large blank background regions, while informative geometric cues are mainly concentrated in local areas such as contours, stroke intersections, and part boundaries. Retaining all patch tokens allows the numerically dominant background tokens to dilute the importance differences among foreground regions, thereby interfering with subsequent semantic injection and structural consistency measurement. Directly discarding low-response tokens, however, may result in contextual information loss. Therefore, we evaluate each patch token according to its attention response from the global token, retain the high-response tokens as the foreground support set, and aggregate the remaining low-response tokens into a single feature that is concatenated with the retained foreground tokens:

$$\mathcal{F} = \mathrm{Concat}\left(\mathrm{TopK}\left(\{z_i\}_{i=1}^{M}; \{r_i\}_{i=1}^{M}, \lceil \rho M \rceil\right), z_s^{\mathrm{bg}}\right). \tag{1}$$

Here, r_i denotes the importance score of the i-th patch token, and ρ is the token retention ratio, z_s^{bg} is the background token.

3.3 Sketch-Structure-preserving Semantic Injection

Given a sketch S and its text description T, Qwen samples K candidate semantic completions and extracts their corresponding hidden representations:

$$y_k \sim \pi_\theta(\cdot \mid S, T),\ H_k = \{h_{k,m}\}_{m=1}^{L_k},\ k = 1, \ldots, K, \tag{2}$$

where π_θ denotes the Qwen generation policy, y_k is the k-th semantic completion, H_k is its corresponding hidden representation, and L_k is the sequence length.

For each sketch token z_i in the foreground support set $\mathcal{F}$, we retrieve locally relevant semantics from H_k and estimate the injection weight:

$$u_i^{(k)} = \mathrm{Attn}(z_i, H_k, H_k), \alpha_i^{(k)} = \sigma\Big(\phi\Big([z_i; u_i^{(k)}]\Big)\Big), \tag{3}$$

where $u_i^{(k)}$ is the retrieved semantic feature, ϕ is a learnable gating network, $[\cdot;\cdot]$ denotes feature concatenation, and $\alpha_i^{(k)} \in (0, 1)$ controls the injection strength.

The semantic feature is then incorporated through a gated residual connection:

$$\tilde{z}_i^{(k)} = z_i + \alpha_i^{(k)} P_q\Big(u_i^{(k)}\Big), \qquad i \in \mathcal{F}, \tag{4}$$

where P_q projects the Qwen feature into the sketch feature space, $\tilde{z}_i^{(k)}$ is the enhanced sketch token.

3.4 Geometry–Texture Representation of 3D Models

Each textured 3D model G_j consists of geometric data X_j and V rendered texture views $\{I_{j,v}\}_{v=1}^{V}$. We extract its geometry and texture features as

$$f_{\mathrm{geo},j} = E_{\mathrm{geo}}(X_j),\ f_{\mathrm{tex},j} = \mathrm{Pool}\left(\{E_{\mathrm{tex}}(I_{j,v})\}_{v=1}^{V}\right), \tag{5}$$

where E_{geo} and E_{tex} denote the geometry and texture encoders, respectively, and $\mathrm{Pool}(\cdot)$ aggregates the features extracted from multiple rendered views.

The two features are concatenated and projected to obtain the final representation of the 3D model:

$$g_j = \mathrm{Norm}\left(P_{\mathrm{3D}}\left([f_{\mathrm{geo},j}; f_{\mathrm{tex},j}]\right)\right), \tag{6}$$

where $[\cdot;\cdot]$ denotes feature concatenation, P_{3D} is a learnable projection, g_j is a representation of G_j.

3.5 Shape–Texture Ranking Reward

Different completions may produce retrieval queries of varying quality. A binary top-1 reward provides sparse feedback and cannot distinguish near-miss rankings or reveal whether retrieval is biased toward shape or texture. We therefore construct a shape–texture ranking reward using controlled conflicting candidates.

For each query $x = (S, T)$, the candidate group $\mathcal{B}_x$ consists of the exact target G^+, same-shape different-texture candidates $\mathcal{N}_x^{\mathrm{tx}}$, and same-texture different-shape candidates $\mathcal{N}_x^{\mathrm{sh}}$. For the query representation q_k produced by the k-th semantic completion, we define $s_{k,j}^{\mathrm{joint}} = s(q_k, g_j)$ and $s_{k,j}^{\mathrm{tex}} = s(q_k, f_{\mathrm{tex},j})$ as the joint and texture-specific cosine similarities. We define the exact-target rank and three score margins as

$$\begin{aligned} r_k &= 1 + \sum_{G_j \in \mathcal{N}_x^{\mathrm{all}}} \mathbf{1}\left[s_{k,j}^{\mathrm{joint}} > s_{k,+}^{\mathrm{joint}}\right], \qquad \Delta_k^{\mathrm{joint}} = s_{k,+}^{\mathrm{joint}} - \max_{G_j \in \mathcal{N}_x^{\mathrm{all}}} s_{k,j}^{\mathrm{joint}}, \\ \Delta_k^{\mathrm{tx}} &= s_{k,+}^{\mathrm{tex}} - \max_{G_j \in \mathcal{N}_x^{\mathrm{tx}}} s_{k,j}^{\mathrm{tex}}, \qquad \Delta_k^{\mathrm{sh}} = s_{k,+}^{\mathrm{joint}} - \max_{G_j \in \mathcal{N}_x^{\mathrm{sh}}} s_{k,j}^{\mathrm{joint}}. \end{aligned} \tag{7}$$

where $s_{k,+}^{\mathrm{joint}}$ and $s_{k,+}^{\mathrm{tex}}$ denote the joint and texture similarities to G^+, respectively. The three margins compare the target score with the maximum score among all negatives, same-shape different-texture negatives, and same-texture different-shape negatives, respectively. The retrieval reward is

$$\begin{aligned} R_k^{\mathrm{ret}} = \lambda_{\mathrm{rank}} r_k^{-1} + \lambda_{\mathrm{margin}} \sigma\left(\frac{\Delta_k^{\mathrm{joint}}}{\tau_{\mathrm{m}}}\right) + \lambda_{\mathrm{tex}} \frac{\exp(s_{k,+}^{\mathrm{tex}}/\tau_{\mathrm{tx}})}{\sum_{G_j \in \{G^+\} \cup \mathcal{N}_x^{\mathrm{tx}}} \exp(s_{k,j}^{\mathrm{tex}}/\tau_{\mathrm{tx}})} \\ - \lambda_{\mathrm{txp}} \left[m_{\mathrm{tx}} - \Delta_k^{\mathrm{tx}}\right]_+ - \lambda_{\mathrm{shp}} \left[m_{\mathrm{sh}} - \Delta_k^{\mathrm{sh}}\right]_+ . \end{aligned} \tag{8}$$

where $[a]_+ = \max(a, 0)$, the λ terms weight the reward components, τ_{m} and τ_{tx} are temperatures, and m_{tx} and m_{sh} are the required separation margins. The reward jointly promotes exact-target ranking, score separation, and texture discrimination while penalizing insufficient separation from conflicting candidates.

3.6 Foreground Structural Consistency Loss

Semantic injection can complement the category and appearance information missing from a sketch. However, excessive semantic injection may overwrite the original sketch representation and alter the relative importance of key stroke regions, thereby weakening geometric cues such as contours, proportions, and part relationships. To preserve the geometric information, we introduce a foreground structural consistency loss that constrains the importance distributions of foreground sketch tokens before and after semantic injection. Since freehand sketches contain large blank regions, computing structural changes over all patch tokens may cause the numerically dominant background tokens to dilute the differences among informative strokes. Therefore, we use $\mathcal{F}$ obtained by the foreground-aware sketch encoder and measure structural drift only over this set. Let r_i and $\tilde{r}_i^{(k)}$ denote the importance scores of the i-th foreground token before and after injecting the k-th semantic completion, respectively. The corresponding normalized importance distributions are defined as

$$p_i^{\mathrm{str}} = \frac{\exp(r_i/\tau_s)}{\sum_{m \in \mathcal{F}} \exp(r_m/\tau_s)}, \quad \tilde{p}_{k,i}^{\mathrm{str}} = \frac{\exp(\tilde{r}_i^{(k)}/\tau_s)}{\sum_{m \in \mathcal{F}} \exp(\tilde{r}_m^{(k)}/\tau_s)}. \tag{9}$$

where τ_s is the temperature parameter and $i \in \mathcal{F}$ indexes the foreground sketch tokens.

The structural drift caused by the k-th semantic completion is defined as

$$D_k^{\mathrm{str}} = D_{\mathrm{KL}}\left(\mathrm{sg}\left[p^{\mathrm{str}}\right] \,\|\, \tilde{p}_k^{\mathrm{str}}\right), \tag{10}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation, which treats the original sketch importance distribution as a fixed geometric reference. This divergence measures whether semantic injection excessively alters the relative importance of key foreground regions. Since effective semantic enhancement may require limited local adjustments, we penalize only structural drift that exceeds a predefined tolerance:

$$\mathcal{L}_{\mathrm{str}} = \frac{1}{K}\sum_{k=1}^{K}\left[D_k^{\mathrm{str}} - \delta_{\mathrm{str}}\right]_+, \tag{11}$$

where δ_{str} denotes the tolerated structural variation and $[x]_+ = \max(x, 0)$.

3.7 Retrieval-Drift-Calibrated Adaptive GRPO

Standard GRPO applies a fixed clipping range to all semantic completions. However, changes in token probabilities do not reflect their effects on 3D retrieval behavior, as similar policy updates may cause different degrees of ranking variation. We therefore measure the retrieval drift induced by the old and current policies in the 3D candidate space and use it to adaptively adjust the clipping range for each completion.

For a semantic completion y_k sampled from the old policy, we fix its discrete token sequence and extract its conditional hidden representations using the old and current policies:

$$H_k^{\mathrm{old}} = F_{\theta_{\mathrm{old}}}(c_x, y_k), \quad H_k^{\mathrm{cur}} = F_{\theta}(c_x, y_k), \tag{12}$$

where $c_x = (S, T)$ denotes the sketch–text query context. The two hidden representations are passed through the same frozen semantic injection and query construction modules, producing q_k^{old} and q_k^{cur}, respectively.

Given the same candidate set $\mathcal{B}_x$, the retrieval distribution induced by each policy is defined as

$$p_{k,j}^{v} = \frac{\exp(s(q_k^v, g_j)/\tau_{\mathrm{ret}})}{\sum_{G_m \in \mathcal{B}_x}\exp(s(q_k^v, g_m)/\tau_{\mathrm{ret}})}, \quad v \in \{\mathrm{old}, \mathrm{cur}\}, \tag{13}$$

where $s(\cdot,\cdot)$ denotes the similarity between a query and a 3D model, τ_{ret} is the retrieval temperature.

The retrieval drift associated with the k-th semantic completion is defined as

$$D_k^{\mathrm{ret}} = D_{\mathrm{KL}}\left(p_k^{\mathrm{old}} \,\|\, p_k^{\mathrm{cur}}\right). \tag{14}$$

Since the semantic completion, candidate set, and downstream retrieval modules remain fixed, D_k^{ret} measures the change in retrieval preference caused by updating the Qwen policy. We map the retrieval drift to a completion-specific clipping range:

$$\epsilon_k = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})\exp\left(-\gamma\ \mathrm{sg}\left[D_k^{\mathrm{ret}}\right]\right), \tag{15}$$

where γ controls the sensitivity of the clipping range to retrieval drift, and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation. A large retrieval drift produces a smaller ϵ_k and restricts the policy update, whereas stable retrieval behavior allows a wider clipping range. The policy ratio for the t-th token of the k-th completion is

$$\rho_{k,t}(\theta) = \frac{\pi_\theta\left(y_{k,t} \mid c_x, y_{k,<t}\right)}{\pi_{\theta_{\mathrm{old}}}\left(y_{k,t} \mid c_x, y_{k,<t}\right)}. \tag{16}$$

Based on the retrieval reward R_k^{ret}, the group-normalized advantage is computed as

$$A_k = \frac{R_k^{\mathrm{ret}} - \frac{1}{K}\sum_{n=1}^{K} R_n^{\mathrm{ret}}}{\sqrt{\frac{1}{K}\sum_{n=1}^{K}\left(R_n^{\mathrm{ret}} - \frac{1}{K}\sum_{m=1}^{K} R_m^{\mathrm{ret}}\right)^2 + \varepsilon_A}}, \tag{17}$$

where ε_A is a small constant used for numerical stability.

The policy ratio is clipped using the completion-specific range

$$\bar{\rho}_{k,t}(\theta) = \mathrm{clip}\left(\rho_{k,t}(\theta), 1-\epsilon_k, 1+\epsilon_k\right), \tag{18}$$

where $\bar{\rho}_{k,t}(\theta)$ denotes the clipped policy ratio, $\rho_{k,t}(\theta)$ is the original policy ratio, and ϵ_k is the adaptive clipping range for the k-th semantic completion.

$$\mathcal{L}_{\mathrm{GRPO}} = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|y_k|}\sum_{t=1}^{|y_k|}\min(\rho_{k,t}(\theta)A_k,\ \bar{\rho}_{k,t}(\theta)A_k)\,. \tag{19}$$

where K is the number of sampled semantic completions, $|y_k|$ is the length of the k-th completion, and A_k denotes its group-normalized advantage. The negative sign converts the clipped policy objective into a loss to be minimized. All tokens in the same semantic completion share ϵ_k, since retrieval drift is measured at the completion level. The final training objective is defined as $\mathcal{L} = \mathcal{L}_{\mathrm{GRPO}} + \mathcal{L}_{\mathrm{str}}$.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

Existing sketch-based 3D retrieval datasets emphasize geometry and lack aligned appearance and text. We introduce **ST-Tex3D v2**, built from SkSh-Lamps-Chairs sketches Qi et al. (2021) and corresponding ShapeNetCore meshes Chang et al. (2015). It contains 15,600 textured instances with three multi-view renderings and GPT-5-generated retrieval descriptions. We focus on chairs and lamps because large-scale instance-level freehand sketch–3D correspondences are mainly available for these categories. Geometry–sketch pairs and textures are independently split into training and test sets at a 7:3 ratio. Retaining only within-split combinations yields 7,644 training instances and 1,404 test instances, with all test shapes, sketches, and textures unseen during training. Details are provided in Appendix A.

4.1.2 Evaluation Metrics

We evaluate exact-instance retrieval using Top-K retrieval accuracy (R@K), where $K \in \{1, 5, 10\}$. R@K measures the percentage of queries for which the ground-truth textured 3D model appears among the top-K results. We report R@1, R@5, and R@10, and higher values indicate better retrieval performance.

4.1.3 Implementation Details

Our model is implemented in PyTorch and trained with DDP on four NVIDIA A40 GPUs (global batch size 32). The retrieval network and Qwen LoRA adapters are optimized separately using AdamW with learning rates of 1×10^{-5} and 2×10^{-5}, respectively, and a StepLR scheduler. The retrieval backbone uses CLIP ViT-B/16 (224×224, 512 dimensions), while the 3D branch employs PointNet++ with 512 points and 16 local groups, alongside three rendered texture views. Qwen3-VL-8B-Instruct is adapted with LoRA ($r = 8$, $\alpha = 16$) on the `q_proj`, `k_proj`, and `v_proj` layers, with all other parameters frozen. GRPO samples eight responses per query (maximum 48 tokens), using reward groups of one target, three same-shape texture negatives, and four same-texture shape negatives. Inference requires only one semantic completion per query, consuming 30.8 GB of GPU memory and 735.7 ms per query at batch size 4. All baselines use the same training data and, where architecturally permitted, the same 3D gallery encoder.

4.2 Comparison with State-of-the-Art Methods

We compare StructAlign with representative single-modal, sketch–text fusion, and MLLM-based retrieval methods. Whenever compatible with the original architectures, all baselines use the same data split, textured 3D gallery encoder, embedding dimension, and evaluation protocol. Single-modal baselines include Pivoting Chowdhury et al. (2023a), Doodle to Object Wang & Zhou (2023), TriCoLo Ruan et al. (2024), and OpenShape Liu et al. (2023). For sketch–text retrieval, we evaluate score-, feature-, and token-level fusion methods, including Late Score Fusion Schlachter et al. (2022a), Weighted Feature Sum, Concatenation + MLP, FiLM Perez et al. (2018), and Global Cross-Attention Vaswani et al. (2017), together with 3D adaptations of SceneDiff Zuo et al. (2024), Sketch2Word Koley et al. (2024a), and PASTA Lee et al. (2025). MLLM-based baselines include Qwen Global Feature, Qwen Completion + Text Encoder, Qwen3-VL-Embedding-8B Li et al.

Table 1: Comparison with single-modal retrieval methods, controlled sketch–text fusion baselines, and MLLM-based baselines on ST-Tex3D v2. R@K denotes Top-K retrieval accuracy. Red and blue cell backgrounds indicate the best and second-best results, respectively. The final row reports the absolute improvement of our method over the second-best result in each column.

Method	Query	Chairs			Lamps		
		R@1	R@5	R@10	R@1	R@5	R@10
Sketch-only Baselines							
Pivoting Chowdhury et al. (2023a)	S	12.43	31.84	46.59	13.62	41.95	56.41
Doodle to Object Wang & Zhou (2023)	S	11.78	32.26	44.88	12.91	39.84	53.72
Text-only Baselines							
TriCoLo Ruan et al. (2024)	T	12.96	31.72	45.18	13.11	35.47	50.26
OpenShape Liu et al. (2023)	T	13.72	31.28	46.73	13.86	34.91	51.24
Sketch–Text Fusion Baselines							
Late Score Fusion Schlachter et al. (2022b)	S+T	19.23	44.86	61.02	18.72	44.31	60.84
Weighted Feature Sum Schlachter et al. (2022a)	S+T	18.41	45.27	60.11	20.06	43.98	62.87
Concatenation + MLP	S+T	20.17	45.92	62.48	19.38	46.16	61.95
FiLM Modulation Perez et al. (2018)	S+T	21.06	46.54	64.23	21.44	47.83	64.96
Global Cross-Attention Vaswani et al. (2017)	S+T	20.52	47.14	65.09	20.91	48.26	66.18
SceneDiff-to-3D Zuo et al. (2024)	S+T	23.54	50.96	70.47	24.16	51.93	71.62
Sketch2Word-to-3D Koley et al. (2024b)	S+T	23.28	50.41	69.36	23.17	51.72	69.12
PASTA-style Semantic Prior Lee et al. (2025)	S+T	22.74	51.38	72.18	**25.02**	51.84	71.27
MLLM-based Baselines							
Qwen Global Feature Bai et al. (2025)	S+T	19.84	46.02	62.73	20.34	45.78	63.01
Qwen Completion + Text Encoder Bai et al. (2025)	S+T	21.57	48.62	65.88	22.13	49.44	67.31
Qwen3-VL-Embedding-8B Li et al. (2026b)	S+T	23.76	52.74	71.68	23.62	51.26	72.84
Qwen-DPO Bai et al. (2025); Rafailov et al. (2023)	S+T	24.86	52.11	71.93	24.79	**53.14**	72.26
UME-R1-to-3D Lan et al. (2026)	S+T	**25.31**	**53.06**	**72.62**	24.88	52.83	**73.12**
StructAlign (Ours)	**S+T**	**43.88**	**78.06**	**88.46**	**48.71**	**75.26**	**88.29**
Improv. over 2nd Best		↑18.57	↑25.00	↑15.84	↑23.69	↑22.12	↑15.17

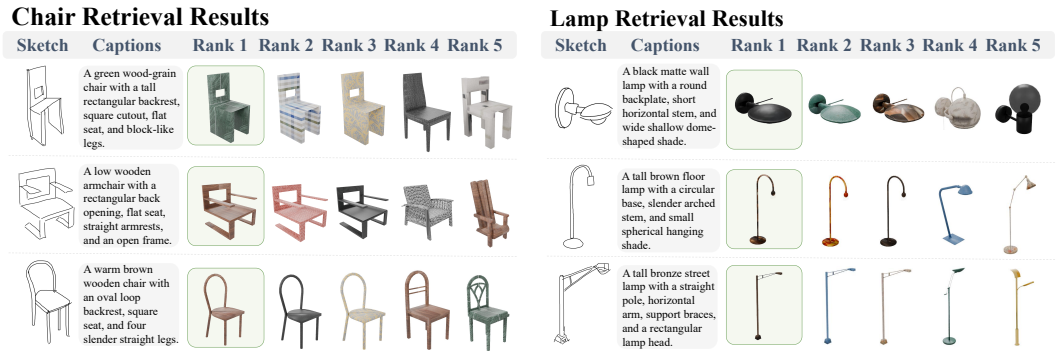

Figure 3: Retrieval results on ST-Tex3D v2.

(2026b), Qwen-DPO Rafailov et al. (2023), and UME-R1-to-3D Lan et al. (2026). For UME-R1-to-3D, we preserve its reasoning and RL pipeline while replacing the target representation with our unified textured 3D gallery encoder. Implementation details are provided in Appendix B.

4.2.1 Comparison Results.

As shown in Table 1, StructAlign consistently and substantially outperforms all competing methods on both the chair and lamp categories. Single-modal methods perform clearly worse than sketch–text

Table 2: Ablation study on the chair subset of ST-Tex3D v2.

Structure-Grounded Query Construction			
Configuration	**R@1**	**R@5**	**R@10**
w/o Foreground-Aware Encoding	39.62	73.41	85.76
w/o Background Aggregation	41.57	75.92	87.31
w/o Local Semantic Injection	34.86	67.95	81.73
w/o Injection Gating	37.94	71.66	84.28
w/o Structural Consistency	39.21	73.88	85.94

Retrieval-Oriented Policy Optimization			
Configuration	**R@1**	**R@5**	**R@10**
w/o Shape–Texture Ranking Reward	35.43	68.72	82.46
w/o Retrieval-Drift Calibration	40.68	74.96	86.54
StructAlign (Full)	**43.88**	**78.06**	**88.46**

approaches, confirming the complementary roles of the two modalities: sketches provide instance-level geometric constraints, whereas text specifies material, color, and surface appearance. Jointly exploiting both modalities is therefore essential for distinguishing textured 3D models with similar geometry but different appearances. Structured cross-modal methods, including SceneDiff-to-3D, Sketch2Word-to-3D, and the PASTA-style Semantic Prior, generally outperform simple score- and feature-level fusion strategies. Nevertheless, they remain considerably behind StructAlign. Among the MLLM-based baselines, the reinforcement-learning-based UME-R1-to-3D and Qwen-DPO achieve the strongest overall performance. This demonstrates the suitability of reinforcement learning for retrieval tasks, since the quality of different semantic completions can be directly evaluated through their induced candidate rankings, allowing the reward signal to align naturally with the downstream retrieval objective. However, these general frameworks do not explicitly address the substantial modality gap between sparse geometric sketches and appearance-oriented textual descriptions. As shown in Fig. 3, StructAlign consistently outperforms the strongest baselines across R@1–R@10 for both chairs and lamps. These gains stem from preserving foreground sketch structure, selectively injecting local semantics, and optimizing completions with shape–texture rewards and retrieval-drift-calibrated updates.

4.3 Ablation Study

We ablate StructAlign on the chair subset of ST-Tex3D v2 using the same data split and textured 3D encoder, with further analyses in Appendix C. As shown in Table 2, removing foreground-aware encoding or replacing background aggregation with token removal degrades performance, highlighting the importance of informative strokes and background context. Among query-construction variants, removing local semantic injection causes the largest drop (9.02, 10.11, and 6.73 points in R@1, R@5, and R@10). Injection gating and structural consistency further help preserve sketch geometry during semantic injection. Removing the shape–texture ranking reward reduces R@1, R@5, and R@10 by 8.45, 9.34, and 6.00 points, respectively, demonstrating the value of joint geometry–texture supervision. Retrieval-drift calibration also improves performance by adapting policy clipping to retrieval-space changes.

5 Conclusion

We presented **StructAlign**, a structure-preserving framework for sketch–text textured 3D retrieval. It treats the sketch as a geometric anchor and selectively injects multimodal semantics through foreground-aware encoding, gated local attention, and Structural KL. A shape–texture ranking reward and retrieval-drift-calibrated GRPO further align semantic completion with downstream retrieval. Experiments and ablations on ST-Tex3D v2 demonstrate consistent improvements over existing methods and confirm the importance of preserving reliable sketch structure during multimodal fusion. The current evaluation is limited to two object categories and retrieval-oriented machine-generated descriptions. Future work will extend the framework to more object categories, natural user descriptions, noisier sketches, and more efficient semantic models.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025.

Patrice Bechard, Chao Wang, Amirhossein Abaskohi, Juan A Rodriguez, Christopher Pal, David Vazquez, Spandana Gella, Sai Rajeswar, and Perouz Taslakian. Starflow: Generating structured workflow outputs from sketch images. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1628–1645, 2026.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Jiaxin Chen and Yi Fang. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 605–620, 2018.

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, pp. 100–116. Springer, 2018.

Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Democratising 2d sketch to 3d shape retrieval through pivoting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23275–23286, 2023a.

Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Scenetrilogy: On human scene-sketch and its complementarity with photo and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10972–10983, 2023b.

Rishi Gupta, Mukilan Karuppasamy, Shyam Marjit, Aditay Tripathi, and Anirban Chakraborty. O3slm: Open weight, open data, and open vocabulary sketch-language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4511–4519, 2026.

Zhizhong Han, Mingyang Shang, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Y2seq2seq: Cross-modal representation learning for 3d shape and text by joint reconstruction and prediction of view and word sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 126–133, 2019.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. You'll never walk alone: A sketch and text duet for fine-grained image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16509–16519, 2024a.

Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. You'll never walk alone: A sketch and text duet for fine-grained image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16509–16519, 2024b.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations*, 2026.

Ehsan Latif, Zirak Khan, and Xiaoming Zhai. Sketchmind: a multi-agent cognitive framework for assessing student-drawn scientific sketches. *Advances in Neural Information Processing Systems*, 38:82803–82829, 2026.

Seunggwan Lee, Hwanhee Jung, Byoungsoo Koh, Qixing Huang, Sang Ho Yoon, and Sangpil Kim. PASTA: Part-aware sketch-to-3d shape generation with text-aligned prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18585–18595, 2025.

Changxing Li, Donglin Zhang, Zhikai Hu, Xiao-Jun Wu, and Josef Kittler. Sketchmind: Understanding abstract sketches with mllms for fine-grained sketch-based image retrieval. In *Proceedings of the ACM Web Conference 2026*, pp. 2453–2464, 2026a.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026b.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3d shape representation towards open-world understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44860–44879, 2023.

AprilPyone MaungMaung, Makoto Shing, Kentaro Mitsui, Kei Sawada, and Fumio Okura. Text-guided scene sketch-to-photo synthesis. *arXiv preprint arXiv:2302.06883*, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Toward fine-grained sketch-based 3d shape retrieval. *IEEE transactions on image processing*, 30:8595–8606, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, and Angel X. Chang. TriCoLo: Trimodal contrastive loss for text to shape retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5815–5825, 2024.

Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A sketch is worth a thousand words: Image retrieval with text and sketch. In *European Conference on Computer Vision*, pp. 251–267. Springer, 2022.

Kristofer Schlachter, Benjamin Ahlbrand, Zhu Wang, Valerio Ortenzi, and Ken Perlin. Zero-shot multi-modal artist-controlled retrieval and exploration of 3d object sets. In *SIGGRAPH Asia 2022 Technical Communications*, New York, NY, USA, 2022a. Association for Computing Machinery. doi: 10.1145/3550340.3564216.

Kristofer Schlachter, Benjamin Ahlbrand, Zhu Wang, Ken Perlin, and Valerio Ortenzi. Zero-shot multi-modal artist-controlled retrieval and exploration of 3d object sets. In *SIGGRAPH Asia 2022 Technical Communications*, pp. 1–4, 2022b.

Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, and Yi Chang. Parts2words: Learning joint embedding of point clouds and texts by bidirectional matching between parts and words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6884–6893, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Bingrui Wang and Yuan Zhou. Doodle to object: Practical zero-shot sketch-based 3d shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2474–2482, 2023. doi: 10.1609/aaai.v37i2.25344.

Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1875–1883, 2015.

Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8929–8939, 2023.

Ying Zang, Yidong Han, Chaotao Ding, Jianqi Zhang, and Tianrun Chen. Magic3dsketch: Create colorful 3d models from sketch-based 3d modeling guided by text and language-image pretraining. *Neurocomputing*, pp. 131925, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Ran Zuo, Haoxiang Hu, Xiaoming Deng, Cangjun Gao, Zhengming Zhang, Yukun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenediff: Generative scene-level image retrieval with text and sketch using diffusion models. 2024.