SHAPE MEETS SURFACE: COMPOSITIONAL SKETCH-TEXT RETRIEVAL OF TEXTURED 3D OBJECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sketches provide an intuitive way to express object geometry, but they are insufficient for specifying textured appearance such as material, color, pattern, and surface style. Text offers complementary appearance descriptions, yet it is often less precise in constraining fine-grained geometry. In this paper, we study a new task: sketch-text guided textured 3D model retrieval, where the goal is to retrieve a textured 3D model that simultaneously satisfies sketch-based geometric intent and text-based appearance intent. To support this task, we construct ST-Tex3D, a tri-modal benchmark containing 12,000 textured 3D instances across the chair and lamp categories, with aligned SVG sketches, ShapeNetCore meshes, multi-view renderings, and retrieval-oriented textual descriptions. We propose a unified retrieval framework that models sketch structure, textual appearance, 3D geometry, and multi-view texture evidence. Specifically, the framework distills high-level semantics for sketch encoding, generates intent-aware sketch-text query tokens, extracts geometry and texture representations for each candidate 3D model, and performs query-aware evidence aggregation with fine-grained late-interaction matching. Experiments on ST-Tex3D validate the effectiveness of the proposed task formulation, dataset, and retrieval framework for texture- and style-sensitive 3D model retrieval. Our code and models will be released upon acceptance.

1 INTRODUCTION

Sketches are highly expressive. With only a few strokes, users can efficiently convey object contours, parts, and spatial layouts. This structural expressiveness has been widely explored in fine-grained image retrieval and has further promoted research on sketch-based 3D shape retrieval Chowdhury et al. (2023); Wang & Zhou (2023); Liang et al. (2024). However, sketches are expressive but not complete. Although sketches are well suited for describing geometric structure, they cannot reliably specify appearance-level attributes such as texture, material, color, and surface style. Existing sketch-based 3D shape retrieval methods usually take sketches as the only query modality and mainly define relevance according to category consistency and geometric similarity. Therefore, when multiple 3D models have similar geometric structures but differ in textured appearance, retrieval based only on sketches is insufficient to satisfy users' fine-grained intent. Text provides a complementary signal. However, existing methods usually take text as the only query condition. Although language is flexible in describing material, color, and style, it remains insufficient in precisely constraining local geometric structure, part layout, and shape proportion Liu et al. (2023); Xue et al. (2024); Luo et al. (2023). These observations indicate that, in textured 3D model retrieval, sketches and text are not substitutes for each other, but complementary cues: sketches constrain geometric structure, while text constrains textured appearance. Based on this motivation, we define a new task: sketch–text guided textured 3D model retrieval. Given a freehand sketch and a textual description, the goal of this task is to retrieve a textured 3D model that simultaneously satisfies the sketch-based geometric intent and the text-based appearance intent. This task differs from sketch-based or text-based 3D shape retrieval. In this task, conditioned on the geometric structure specified by the sketch, text provides appearance constraints such as texture, material, color, and style.

A major obstacle to studying this task is the lack of suitable datasets. Existing sketch–shape datasets mainly focus on matching sketches with 3D geometry Wang & Zhou (2023). Text–3D datasets support semantic alignment between language and 3D objects, but do not provide sketch queries Chen et al. (2018); Achlioptas et al. (2019); Luo et al. (2023). Although large-scale 3D datasets contain

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.

textures, materials, or metadata, they are not organized as joint retrieval supervision in the form of sketch–text–textured 3D model triplets Chang et al. (2015); Wu et al. (2015); Deitke et al. (2023b;a); Collins et al. (2022); Fu et al. (2021). Therefore, existing datasets cannot directly evaluate the retrieval ability of models under a joint query scenario where sketches specify geometric structure and text specifies appearance attributes.

To fill this gap, we construct ST-Tex3D, a tri-modal dataset for sketch–text guided textured 3D model retrieval, containing 12,000 textured 3D instances. This benchmark integrates SVG sketches, ShapeNetCore OBJ meshes, texture variants, multi-view renderings, and retrieval-oriented textual descriptions Qi et al. (2021); Chang et al. (2015). These textual descriptions emphasize discriminative attributes, including shape structure, material type, surface texture, color distribution, and overall style, thereby supporting texture- and style-sensitive retrieval under similar geometric structures. Furthermore, we propose a unified framework for sketch–text guided textured 3D model retrieval. The framework first enhances sketch representations through semantic distillation to alleviate the sparsity and ambiguity of sketches. Then, it jointly encodes sketch structure, sketch semantics, and textual appearance descriptions to generate multi-granular query intents. For each candidate 3D model, the framework simultaneously extracts geometric evidence and multi-view texture evidence. Finally, through query-aware evidence aggregation and fine-grained late-interaction matching, it jointly aligns geometric structure and textured appearance. Our main contributions are summarized as follows:

• We propose the task of sketch–text guided textured 3D model retrieval, extending traditional sketch-based 3D shape retrieval from focusing only on geometric matching to also considering texture-sensitive 3D model retrieval.

• We construct ST-Tex3D, a tri-modal dataset for sketch–text guided textured 3D model retrieval. The dataset contains 12,000 textured 3D instances, and provides SVG sketches, meshes, texture variants, multi-view renderings, and retrieval-oriented textual descriptions.

• We propose a unified retrieval framework that jointly models sketch geometry, textual appearance, 3D structure, and multi-view texture evidence. The framework achieves textured 3D model retrieval through semantic distillation, query intent modeling, evidence aggregation, and fine-grained late-interaction matching.

2 Related Work

Sketch-Based 3D Shape Retrieval. Sketch-based 3D shape retrieval (SBSR) retrieves 3D models using freehand sketches as queries. Early methods matched sketches with rendered views using hand-crafted line descriptors Yoon et al. (2010); Eitz et al. (2012). Deep approaches later learned cross-modal representations through Siamese networks, metric learning, and adversarial domain adaptation Wang et al. (2015); Dai et al. (2017); Chen & Fang (2018). To reduce 2D–3D view ambiguity, subsequent methods introduced multi-view aggregation and sketch-guided view attention Xie et al. (2017); Qi et al. (2021). Recent studies further extend SBSR to instance-level, weakly paired, and zero-shot settings through pivot learning, domain disentanglement, prototype contrastive learning, and teacher–student transfer Chowdhury et al. (2023); Xu et al. (2022); Wang & Zhou (2023); Liang et al. (2024). Nevertheless, existing methods remain largely sketch-only and geometry-driven, with appearance attributes such as texture, material, and color insufficiently modeled.

Text-Based 3D Shape Retrieval. Text-based 3D shape retrieval aligns natural language with 3D objects for text-to-shape retrieval, generation, and open-vocabulary 3D understanding. Early work learned from annotated text–3D pairs, covering global object-level alignment between free-form text and colored 3D shapes Chen et al. (2018), sequence-level alignment between view and word sequences Han et al. (2019), referential grounding for distinguishing similar 3D objects Achlioptas et al. (2019), and part–word correspondence through bidirectional matching between point-cloud parts and text words Tang et al. (2023). With vision-language models, text–3D alignment has moved toward large-scale cross-modal pretraining. One line of research transfers 2D image-text semantics to 3D representations through multi-view depth projections or image-depth contrastive pretraining Zhang et al. (2022); Huang et al. (2023). Another line learns unified image–text–3D representations through object-level triplet pretraining, multi-source 3D data scaling, automatic language generation, and large-scale 3D-text construction Xue et al. (2023); Liu et al. (2023); Xue et al. (2024); Luo

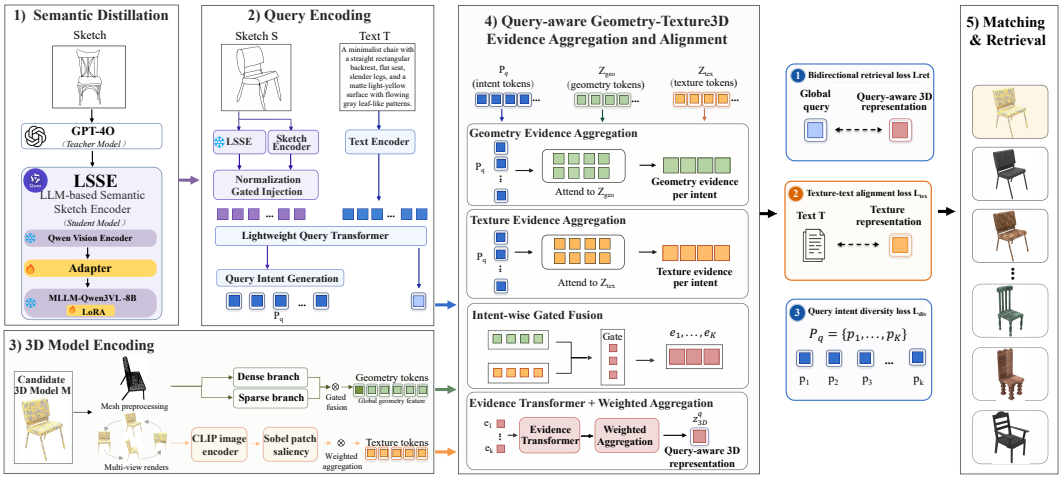

Figure 1: Overview of the proposed sketch-text guided textured 3D model retrieval framework. The model jointly encodes sketch-text queries and textured 3D candidates, aggregates geometry-texture evidence, and retrieves the best-matched 3D model.

et al. (2023). Recent methods further improve fine-grained text–3D alignment through multi-scale reasoning and dynamic attention, while large-scale binding models extend unified representations across multiple modalities Li et al. (2025); Fan et al. (2026); Zhang et al. (2025). These methods advance text–3D representation learning but typically treat text as the sole query, whereas our setting combines sketch-based geometry with text-specified appearance, including color, texture, material, and surface style.

Datasets for Multimodal 3D Retrieval. Existing 3D datasets cover sketch–shape retrieval, text–3D alignment, and textured 3D assets. Sketch-based datasets, including SHREC'13 Li et al. (2013), SHREC'14 Li et al. (2014), and Doodle2Object Wang & Zhou (2023), support sketch-to-geometry matching; text–3D datasets, such as Text2Shape Chen et al. (2018), ShapeGlot Achlioptas et al. (2019), and Cap3D Luo et al. (2023), provide language supervision for 3D semantic alignment; and large-scale asset collections, including ShapeNet Chang et al. (2015), ModelNet Wu et al. (2015), Objaverse Deitke et al. (2023b), Objaverse-XL Deitke et al. (2023a), ABO Collins et al. (2022), and 3D-FUTURE Fu et al. (2021), provide geometry, category labels, metadata, textures, or material information. However, these resources are not organized as aligned triplets of freehand sketch, texture description, and textured 3D target. They therefore cannot directly support the joint query setting studied here, where sketches specify geometry and text specifies texture or material appearance.

3 Methodology

3.1 Semantic Distillation for Sketch Encoding

Sketches are sparse and ambiguous, making it difficult to infer reliable high-level semantics from line drawings alone. We therefore introduce an LLM-based Semantic Sketch Encoder (LSSE) to enhance sketch representations with distilled semantic knowledge. During distillation, GPT-4o is used as the teacher. Given a sketch S of the corresponding 3D model, the teacher generates a semantic supervision text Y. The student is built upon Qwen3-VL-8B with LoRA and a lightweight TokenConv Adapter. After distillation, the student is frozen and used as LSSE.

Given a sketch S, LSSE produces a sequence of sketch tokens, which are projected into the sketch feature space:

$$Z_{\text{sem}} = \text{LSSE}(S) = \{z_j\}_{j=1}^{N_z}, \bar{Z}_{\text{sem}} = \phi_{\text{sem}}(Z_{\text{sem}}). \tag{1}$$

where Z_{sem} denotes the semantic token sequence and $\phi_{\text{sem}}(\cdot)$ is a learnable alignment projection.

Let $Z_s^0 = \{s_i^0\}_{i=1}^{N_s}$ be the patch tokens produced by the sketch backbone. To inject semantic cues while preserving contour and local structural information, we use the sketch tokens as queries and the aligned semantic tokens as keys and values, followed by a gated residual update:

$$A_s = \mathrm{CrossAttn}(Z_s^0, \bar{Z}_{\mathrm{sem}}, \bar{Z}_{\mathrm{sem}}), Z_s = \mathrm{LN}(Z_s^0 + G_s \odot A_s). \tag{2}$$

where A_s is the semantic information aggregated from LSSE tokens, G_s is a learnable injection gate, and Z_s is the final semantic-enhanced sketch representation. This design enriches sparse sketch features with high-level semantics while avoiding the suppression of original geometric cues, and Z_s is used for subsequent sketch-text query encoding.

3.2 Sketch-Text Query Encoding and Intent Generation

Given the semantic-enhanced sketch tokens Z_s and the text description T, we encode the text into token-level features $Z_t = [t_{\mathrm{cls}}, t_1, \ldots, t_{N_t}]$. The sketch and text tokens are then concatenated and fed into a lightweight query Transformer to obtain a unified sketch-text representation:

$$Z_q = \mathrm{Transformer}_q\left([Z_s; Z_t]\right), z_q = \mathrm{Norm}\left(W_q^c Z_q^{\mathrm{cls}}\right). \tag{3}$$

where Z_q is the fused token sequence, Z_q^{cls} is its global token, and z_q denotes the global sketch-text query representation. To model retrieval intents, we introduce K learnable intent queries $Q_0 = \{q_m^0\}_{m=1}^K$, which attend to the fused sketch-text tokens and generate query-aware intent tokens:

$$P_q = \mathrm{LN}\left(Q_0 + \mathrm{Attn}(Q_0, Z_q, Z_q)\right), \tag{4}$$

where $P_q = \{p_m\}_{m=1}^K$ denotes the intent token set. These tokens are designed to capture complementary cues, such as global shape, local parts, category semantics, color, material, and texture, and are later used to extract query-relevant geometry and texture evidence from candidate 3D models.

3.3 Geometry-Texture Encoding for 3D Models

To represent each candidate textured 3D model, we extract complementary evidence from both geometry and appearance. Specifically, we use a geometry-aware 3D encoder to capture global shape and local structural cues from point clouds, and a texture-aware multi-view encoder to extract color, material, texture, and surface-style information from rendered views. The resulting geometry and texture tokens are later used for query-aware evidence aggregation.

3.3.1 Geometry-Aware 3D Encoding

Given an input mesh model $\mathcal{M} = (\mathcal{V}, \mathcal{F})$ where $\mathcal{V}$ and $\mathcal{F}$ are the vertex and face sets, we first normalize it by centering and scale normalization, and then sample points on the mesh surface according to triangle areas. Each sampled point is represented with geometric attributes including 3D coordinates, surface normals, local curvature, and local density. These attributes provide both global shape information and local geometric cues such as edges, boundaries, and part connections. To encode the sampled point cloud, we adopt a dense-sparse PointNet++ architecture. The dense branch processes the full point cloud to capture the 3D structure, while the sparse branch focuses on structurally salient points selected according to curvature, normal variation, and local sparsity. These salient points correspond to regions with strong geometric changes, such as boundaries and part junctions. The token sequences from the dense and sparse branches are fused through a gated fusion module and projected into a unified feature space. This produces the geometric token sequence

$$Z_{\mathrm{geo}} = [z_{\mathrm{geo}}^{\mathrm{cls}}, z_1^{\mathrm{geo}}, \ldots, z_G^{\mathrm{geo}}], \tag{5}$$

where $z_{\mathrm{geo}}^{\mathrm{cls}}$ serves as the global geometric representation, and the remaining tokens encode local geometric structures.

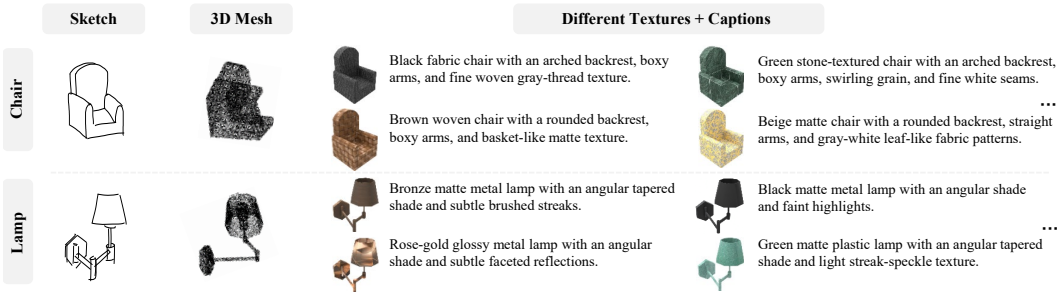

Figure 2: Examples of ST-Tex3D. Each 3D mesh is paired with multiple texture variants and retrieval-oriented captions.

3.3.2 Texture-Aware Multi-View Encoding

Geometry alone cannot describe appearance-related properties such as color, material, texture, and local surface patterns. Therefore, we further extract texture-aware visual representations from multi-view renderings of each 3D model. Given a model $\mathcal{M}$, we render it from N_v fixed viewpoints and use a CLIP image encoder to obtain view-level global tokens and patch-level visual tokens. To emphasize texture boundaries, local patterns, and high-frequency appearance regions, we compute patch-level saliency from image gradients and aggregate salient patch tokens into local texture representations. For each view, the salient texture representation is fused with the corresponding global visual token to obtain a view-level texture feature. We then aggregate features from all views according to their saliency responses, producing a global texture token z_{tex}. In addition, highly salient patch tokens from all views are selected and grouped into local texture tokens to support fine-grained appearance matching. The final texture token sequence is defined as

$$Z_{\text{tex}} = [z_{\text{tex}}, \ell_1, \ldots, \ell_{K_{\text{loc}}}], \tag{6}$$

where z_{tex} denotes the global texture representation and ℓ_k denotes the k-th local texture token. Finally, Z_{geo} and Z_{tex} are used as geometric and texture evidence for subsequent query-aware 3D representation learning.

3.4 Query-Aware 3D Evidence Aggregation and Matching

The sketch-text query may require different types of 3D evidence. Sketches mainly constrain global contour and local geometry, while text descriptions often specify category, color, material, texture, or other fine-grained appearance attributes. Therefore, instead of simply concatenating geometry and texture features, we aggregate 3D evidence according to the query intent tokens and further perform fine-grained token-level matching.

3.4.1 Query-Aware Geometry-Texture Evidence Aggregation

Given the query intent tokens $P_q = \{p_m\}_{m=1}^{K}$, the geometric token sequence Z_{geo}, and the texture token sequence Z_{tex}, we first map them into a shared feature space. Each query intent token is then used to separately attend to geometry and texture tokens, producing intent-specific 3D evidence:

$$e_m^{\text{geo}} = \text{CrossAttn}\left(p_m, Z_{\text{geo}}, Z_{\text{geo}}\right), e_m^{\text{tex}} = \text{CrossAttn}\left(p_m, Z_{\text{tex}}, Z_{\text{tex}}\right). \tag{7}$$

Here, e_m^{geo} and e_m^{tex} denote the geometry and texture evidence associated with the m-th query intention, respectively. This design allows different query intents to focus on different 3D regions or appearance details.

Since the relative importance of geometry and texture varies across queries, we introduce a query-conditioned gate to adaptively fuse the two types of evidence:

$$e_m = \text{LN}\left(\delta_m \odot e_m^{\text{geo}} + (1 - \delta_m) \odot e_m^{\text{tex}}\right), \tag{8}$$

where δ_m is predicted from the query intent and its corresponding 3D evidence. The fused token e_m represents the query-relevant 3D evidence for the m-th intention. In this way, the model can emphasize geometric cues for structure-oriented queries and texture cues for appearance-oriented queries.

After obtaining the fused evidence tokens for all query intentions, we use a lightweight evidence Transformer to model their interactions and aggregate them into a query-aware 3D representation z^q_{3D}. This representation preserves both intent-specific local evidence and global 3D context, making it suitable for fine-grained sketch-text to textured 3D model retrieval.

3.4.2 Fine-Grained Late-Interaction Matching

Although z^q_{3D} provides a query-aware global representation, global aggregation may weaken local intent–evidence correspondence. We therefore introduce a bidirectional token-level late-interaction score. Let $C_{mr} = \langle \hat{p}_m, \hat{e}_r \rangle$ denote the cosine similarity between the m-th query intent token and the r-th 3D evidence token, where hats denote ℓ_2 normalization. The fine-grained matching score is

$$s_{\mathrm{fine}}(q, \mathcal{M}) = \frac{1}{2K} \left(\sum_{m=1}^{K} \max_{r} C_{mr} + \sum_{r=1}^{K} \max_{m} C_{mr} \right). \tag{9}$$

The two terms perform query-to-evidence and evidence-to-query matching, respectively. Let $s_{\mathrm{global}}(q, \mathcal{M})$ denote the cosine similarity between the global query representation z_q and the query-aware 3D representation z^q_{3D}. The final retrieval score equally combines the two similarities:

$$s(q, \mathcal{M}) = \frac{1}{2} \left[s_{\mathrm{global}}(q, \mathcal{M}) + s_{\mathrm{fine}}(q, \mathcal{M}) \right]. \tag{10}$$

3.5 Training Objectives

We train the model with retrieval, texture–text alignment, and query-intent diversity objectives. For a mini-batch of B matched query–3D pairs, let $S = [s_{ij}]$ denote the query–3D similarity matrix combining global and fine-grained matching, and $A = [a_{ij}]$ the text–texture similarity matrix. Since both use symmetric cross-entropy matching, we define

$$\mathcal{L}_{\mathrm{sym}}(M) = \frac{1}{2} \left[\mathrm{CE}(M, \mathbf{y}) + \mathrm{CE}(M^\top, \mathbf{y}) \right], \quad \mathbf{y} = (1, \ldots, B), \tag{11}$$

where diagonal pairs are positive matches. The retrieval and texture–text alignment losses are therefore given by $\mathcal{L}_{\mathrm{ret}} = \mathcal{L}_{\mathrm{sym}}(S)$ and $\mathcal{L}_{\mathrm{tex}} = \mathcal{L}_{\mathrm{sym}}(A)$, respectively.

To reduce redundancy among query intent tokens, we further use

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{B} \sum_{i=1}^{B} \left\| \bar{P}^i_q \left(\bar{P}^i_q \right)^\top - I_K \right\|^2_F, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{ret}} + \mathcal{L}_{\mathrm{tex}} + \mathcal{L}_{\mathrm{div}}, \tag{12}$$

where $\bar{P}^i_q$ is the row-normalized intent-token matrix and I_K is the identity matrix.

4 ST-Tex3D Dataset

Existing sketch-based 3D retrieval datasets mainly evaluate geometric similarity and lack appearance-aware supervision. We therefore construct ST-Tex3D, a tri-modal benchmark of freehand sketches, textual descriptions, and textured 3D models, as shown in Fig. 2. By augmenting sketch–3D pairs with controlled texture variants and retrieval-oriented descriptions, ST-Tex3D supports fine-grained discrimination among geometrically similar models with different appearances.

ST-Tex3D is built from 1,200 freehand sketches in AmateurSketch-3D Qi et al. (2021), including 800 chairs and 400 lamps, together with their corresponding ShapeNet meshes Chang et al. (2015). We focus on chairs and lamps because they provide genuine instance-level sketch–3D pairs at sufficient scale; extending to other categories would largely require synthetic sketches, which may reduce realism and retrieval difficulty Lee et al. (2025); Chowdhury et al. (2023); Chen et al. (2024).

For each category, we manually collect ten category-specific textures and apply the same texture set to every geometry, producing 12,000 textured 3D instances. This controlled design preserves geometry while varying appearance, enabling direct evaluation of texture-sensitive retrieval. Each instance is rendered from three fixed viewpoints. GPT-5 generates one retrieval-oriented description for each instance from a single rendered view and its material label. The descriptions emphasize shape, material, texture, color, and style. Appendix A provides the prompt, texture construction and mapping procedures, and rendering settings.

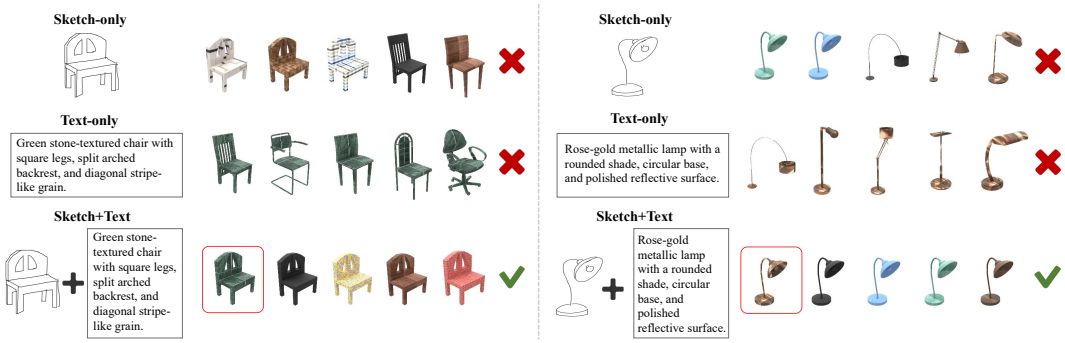

Figure 3: Comparison of sketch-only, text-only, and sketch-text retrieval. Sketch-text queries jointly constrain geometric structure and appearance attributes, leading to more accurate textured 3D model retrieval results.

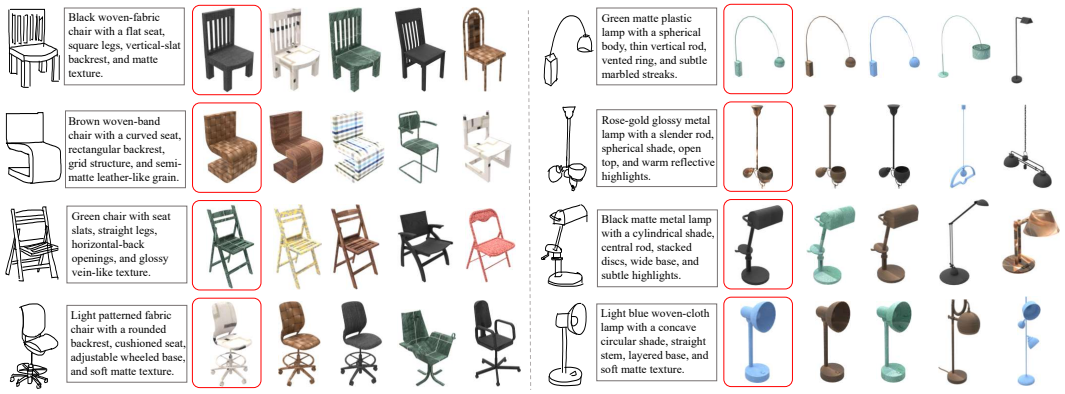

Figure 4: Results of sketch-text guided textured 3D model retrieval on the chair and lamp categories.

5 Experiments

5.1 Dataset and Evaluation Metrics

We evaluate our method on ST-Tex3D. For each category, the base geometries and texture set are independently partitioned into disjoint training and test subsets at a 7:3 ratio. We construct the training set using only training geometries paired with training textures, and the test set using only test geometries paired with test textures. Therefore, neither the geometry nor the texture of a test instance appears during training. This yields 5,880 training and 1,080 test instances; cross-split geometry–texture combinations are excluded to enforce strict generalization to both unseen shapes and unseen textures. Retrieval is performed within each category unless otherwise specified. Given the i-th sketch–text query q_i, the model computes a global query-aware similarity and a fine-grained late-interaction score for each candidate. The two scores are equally weighted to produce the final retrieval score, and all candidates are ranked accordingly. Recall@K measures whether the ground-truth target model appears among the top-K retrieved results. We report Recall@1, Recall@5, and Recall@10. Recall@1 reflects the model's ability to rank the correct target first, whereas Recall@5 and Recall@10 measure whether the target can be retrieved within a short candidate list.

5.2 Implementation Details

Our model is implemented in PyTorch and trained for 200 epochs on four NVIDIA A40 GPUs using AdamW with a global batch size of 32 (8 per GPU). The learning rate is set to 1×10^{-4} and decayed by a factor of 0.5 every 40 epochs using a StepLR scheduler. We use $K = 8$ learnable intent queries and represent each textured 3D model using three fixed-view renderings. We use the CLIP ViT-B/16 image encoder for sketches and its paired text encoder, with a shared feature dimension of 512 and

Method	Query	Chairs			Lamps		
		R@1	R@5	R@10	R@1	R@5	R@10
Sketch-only Baselines							
PointNet++ Qi et al. (2017)	S	5.34	18.62	29.17	3.86	26.41	39.28
MVCNN Su et al. (2015)	S	6.71	21.45	33.08	4.32	29.76	42.63
MVAttn Qi et al. (2021)	S	8.26	24.39	36.52	6.57	33.12	46.94
GANSketch Wang et al. (2022)	S	7.48	22.81	34.67	5.23	31.58	44.49
DeepSDF Park et al. (2019)	S	9.15	26.34	39.72	8.41	36.26	49.35
PSGN Fan et al. (2017)	S	10.06	28.57	42.14	10.18	38.73	52.27
Pivot-SBSR Chowdhury et al. (2023)	S	12.43	32.18	46.59	13.62	42.89	56.41
Text-only Baselines							
CLIP-MV Radford et al. (2021)	T	14.87	36.42	50.31	19.54	49.73	62.26
OpenShape Liu et al. (2023)	T	16.92	39.85	53.64	23.28	53.47	65.83
RMARN Li et al. (2025)	T	15.61	38.04	51.72	21.36	51.18	63.94
3DAlign-DAER Fan et al. (2026)	T	18.92	43.57	57.66	26.18	57.84	70.11
OmniBind Zhang et al. (2025)	T	17.63	41.28	55.21	24.41	55.16	67.42
Sketch–Text Fusion Baselines							
MVAttn + CLIP-MV Qi et al. (2021); Radford et al. (2021)	S+T	17.46	40.27	54.18	30.37	58.29	70.64
Pivot-SBSR + CLIP-MV Chowdhury et al. (2023); Radford et al. (2021)	S+T	18.38	42.91	55.73	33.86	61.54	72.92
Ablation of Our Method							
w/o LSSE and Gated Semantic Injection	S+T	21.83	47.64	61.39	30.84	57.61	68.92
w/o Texture-aware Encoding	S+T	20.57	45.28	59.46	28.73	54.96	66.47
w/o Query-aware Evidence Aggregation	S+T	23.19	50.74	63.81	34.91	60.84	71.26
w/o Texture-Text Alignment Loss	S+T	22.61	49.57	62.94	33.58	59.47	70.53
w/o Intent Diversity Loss	S+T	24.87	53.46	66.18	36.72	63.28	73.64
Ours	**S+T**	**29.12**	**58.34**	**70.41**	**40.17**	**67.42**	**79.86**

Table 1: **Comparison and ablation on sketch–text to textured 3D model retrieval.** We report R@1, R@5, and R@10 on Chairs and Lamps. S, T, and S+T denote sketch, text, and sketch–text queries, respectively.

a maximum text length of 77. For 3D inputs, we center and scale-normalize each mesh and sample 512 points from its surface. The geometry branch adopts a dual-branch dense–sparse PointNet++ encoder with 16 local groups to extract hierarchical point-cloud features. Evaluation is conducted on four NVIDIA A40 GPUs, with an average retrieval time of approximately 509 ms per query.

5.3 Comparisons with State-of-the-Art Models

We compare our method with representative sketch-based, text-based, and sketch-text fusion baselines on the proposed textured 3D retrieval benchmark. We make adaptations to these methods for evaluation on ST-Tex3D, with implementation details provided in the Appendix B. For sketch-based retrieval, we evaluate adapted variants of PointNet++, MVCNN, MVAttn, GANSketch, DeepSDF, PSGN, and Pivot-SBSR using only the sketch query. For text-based methods, including CLIP-MV, OpenShape, RMARN, 3DAlign-DAER, and OmniBind, we use only the textual description as the query to evaluate the capability of language in modeling appearance attributes such as texture, material, and color. In addition, we construct two sketch-text fusion baselines, MVAttn+CLIP-MV and Pivot-SBSR+CLIP-MV, by fusing sketch-based geometric similarity and text-based appearance similarity to obtain the final retrieval results.

As shown in Fig. 3, sketch-only methods retrieve geometrically similar candidates but lack appearance-level cues to distinguish differences in texture, material, or color. Text-only methods better capture appearance semantics, but lack geometric constraints from sketches. Sketch-text fusion baselines improve over single-modality methods, confirming the complementarity between sketch and text. As shown in Table 1, on Chairs, our method obtains 29.12%, 58.34%, and 70.41% on R@1, R@5, and R@10, outperforming the strongest fusion baseline Pivot-SBSR+CLIP-MV by

10.74, 15.43, and 14.68 percentage points. On Lamps, it reaches 40.17%, 67.42%, and 79.86%, with gains of 6.31, 5.88, and 6.94 percentage points. These results indicate that textured 3D model retrieval requires unified modeling of sketch geometry, textual appearance intent, 3D structure, and multi-view texture evidence. Fig. 4 shows qualitative retrieval results on Chairs and Lamps. Each query consists of a freehand sketch and a description, and the textured 3D models are ranked from left to right. The red boxes highlight the top-ranked results. The examples show that our method can retrieve models that match both the sketch-specified geometry and the text-specified appearance.

5.4 Ablation Studies

We conduct ablation studies on Chairs and Lamps to evaluate the contribution of each component, as shown in Table 1. All variants are trained and tested under the same setting, with only the corresponding module or loss removed. Removing LSSE with gated semantic injection decreases R@1, R@5, R@10 by 7.29, 10.70, 9.02 points on Chairs and 9.33, 9.81, 10.94 points on Lamps, respectively. This confirms that distilled sketch semantics are beneficial for resolving the ambiguity of sparse sketches. Removing texture-aware multi-view encoding leads to the largest degradation, with R@1, R@5, R@10 dropping by 8.55, 13.06, 10.95 points on Chairs and 11.44, 12.46, 13.39 points on Lamps. This demonstrates the importance of explicitly modeling appearance cues such as color, material, and texture for textured 3D retrieval. Removing query-aware geometry-texture evidence aggregation reduces R@1, R@5, R@10 by 5.93, 7.60, 6.60 points on Chairs and 5.26, 6.58, 8.60 points on Lamps, indicating that query-conditioned evidence selection is more effective than static geometry-texture fusion. For training objectives, removing the texture-text alignment loss $\mathcal{L}_{\mathrm{tex}}$ decreases R@1, R@5, R@10 by 6.51, 8.77, 7.47 points on Chairs and 6.59, 7.95, 9.33 points on Lamps, showing its effectiveness in aligning textual appearance descriptions with texture evidence. Removing the intent diversity regularization $\mathcal{L}_{\mathrm{div}}$ also degrades performance, with R@1, R@5, R@10 dropping by 4.25, 4.88, 4.23 points on Chairs and 3.45, 4.14, 6.22 points on Lamps, suggesting that diverse intent tokens help capture complementary retrieval cues. Overall, all ablated variants perform worse than the full model, validating the effectiveness of the proposed semantic sketch enhancement, texture-aware 3D encoding, query-aware evidence aggregation, and auxiliary training objectives.

6 Discussion and Limitations

Our results show that sketch and text provide complementary constraints, while intent-dependent geometry–appearance alignment is more effective than static fusion. Several limitations remain. Following prior fine-grained sketch-based 3D retrieval protocols, we evaluate chairs and lamps in separate category-specific galleries, leaving unified multi-category retrieval unassessed. Moreover, automatically generated descriptions may not fully capture the linguistic diversity and ambiguity of real user queries.

7 Conclusion

In this paper, we introduced sketch–text guided textured 3D model retrieval, a new setting that requires models to jointly satisfy sketch-specified geometric constraints and text-specified appearance constraints. Unlike conventional sketch-based 3D shape retrieval, which mainly focuses on geometry, our task further requires fine-grained discrimination of texture, material, color, and surface style among geometrically similar models. To support this task, we constructed ST-Tex3D, a sketch–text–textured-3D tri-modal benchmark for chairs and lamps, comprising SVG sketches, ShapeNetCore meshes, texture variants, multi-view renderings, and retrieval-oriented descriptions. We further proposed a unified framework combining semantic sketch distillation, sketch-text query intent modeling, geometry-texture 3D encoding, query-aware evidence aggregation, and fine-grained late-interaction matching. By enhancing sparse sketches with distilled semantic cues and aligning query intents with geometric and texture evidence, our method retrieves textured 3D models that preserve the sketch-specified structure while matching the text-specified appearance. Experimental results demonstrate its effectiveness for texture- and style-sensitive retrieval among geometrically similar 3D models. Future work will extend ST-Tex3D to broader categories, mixed-category galleries, and human-provided sketch–text queries.

REFERENCES

Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. Shapeglot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8938–8947, 2019.

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.

Minglin Chen, Weihao Yuan, Yukun Wang, Zhe Sheng, Yisheng He, Zilong Dong, Liefeng Bo, and Yulan Guo. Sketch2nerf: multi-view sketch-guided text-to-3d generation. *arXiv preprint arXiv:2401.14257*, 2024.

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, 2023.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21126–21136, 2022.

Guoxian Dai, Jin Xie, Fan Zhu, and Yi Fang. Deep correlated metric learning for sketch-based 3d shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023a.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023b.

Mathias Eitz, Ronald Richter, Tamy Boubekeur, Kristian Hildebrand, and Marc Alexa. Sketch-based shape retrieval. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 605–613, 2017.

Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, Jian Wang, and Keze Wang. 3dalign-daer: Dynamic attention policy and efficient retrieval strategy for fine-grained 3d-text alignment at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 265–273, 2026.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129 (12):3313–3337, 2021.

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.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22157–22167, 2023.

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.

Bo Li, Yijuan Lu, Afzal Godil, Tobias Schreck, Masaki Aono, Henry Johan, Jose M Saavedra, and Shoki Tashiro. Shrec'13 track: Large scale sketch-based 3d shape retrieval. 2013.

Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Hongbo Fu, Takahiko Furuya, Henry Johan, et al. Shrec'14 track: Extended large scale sketch-based 3d shape retrieval. In *Eurographics workshop on 3D object retrieval*, volume 2014, pp. 121–130, 2014.

Wenrui Li, Wei Han, Yandu Chen, Yeyu Chai, Yidan Lu, Xingtao Wang, and Xiaopeng Fan. Riemann-based multi-scale attention reasoning network for text-3d retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18485–18493, 2025.

Shuang Liang, Weidong Dai, Yiyang Cai, and Chi Xie. Sketch-based 3d shape retrieval via teacher–student learning. *Computer Vision and Image Understanding*, 239:103903, 2024.

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. *Advances in neural information processing systems*, 36:44860–44879, 2023.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pre-trained models. *Advances in Neural Information Processing Systems*, 36:75307–75337, 2023.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

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.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 945–953, 2015.

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.

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.

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.

Jiayun Wang, Jierui Lin, Qian Yu, Runtao Liu, Yubei Chen, and Stella X Yu. 3d shape reconstruction from free-hand sketches. In *European Conference on Computer Vision*, pp. 184–202. Springer, 2022.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Jin Xie, Guoxian Dai, Fan Zhu, and Yi Fang. Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5068–5076, 2017.

Rui Xu, Zongyan Han, Le Hui, Jianjun Qian, and Jin Xie. Domain disentangled generative adversarial network for zero-shot sketch-based 3d shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2902–2910, 2022.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1179–1189, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024.

Sang Min Yoon, Maximilian Scherer, Tobias Schreck, and Arjan Kuijper. Sketch-based 3d model retrieval using diffusion tensor fields of suggestive contours. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 193–200, 2010.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8552–8562, 2022.

Ziang Zhang, Minjie Hong, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, Zhou Zhao, et al. Omnibind: Large-scale omni multimodal representation via binding spaces. In *International Conference on Learning Representations*, volume 2025, pp. 20831–20851, 2025.