RM3D: A TRAINING-FREE RETRIEVAL-AND-MORPHING PIPELINE FOR SINGLE IMAGE 3D GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present RM3D, a training-free retrieval-and-morphing framework for single-image 3D generation that uses a retrieved 3D asset without per-instance optimization, task-specific fine-tuning, or additional trainable weights. Existing retrieval-augmented methods provide external 3D evidence at inference time, but still optimize each input or train a reference-conditioned network. RM3D instead reuses a frozen native-3D generator as both a shape prior and a mechanism for latent morphing. Given one image, it retrieves a semantically compatible mesh, loads its precomputed structured latent, and morphs that latent toward the input image inside the frozen sampler. To preserve the retrieved geometry while adapting its appearance, RM3D aligns source and target condition tokens through Hungarian matching, uses correspondence-matched key/value memory in early transformer blocks, and switches to target-only key/value memory in later blocks. Across the quantitative comparisons, RM3D ranks first among the compared methods on every reported appearance and geometry metric. Qualitative comparisons also show that it better preserves input appearance and cross-view structure. A controlled scaling experiment further shows that no metric worsens as the retrieval database grows, indicating that RM3D can improve through a larger asset collection without retraining.

1 INTRODUCTION

Textured 3D assets are fundamental components of games, film, product design, AR/VR, simulation, and embodied AI, and their manual authoring requires substantial time and specialized expertise. Generating 3D assets from a single image is therefore an active research problem that is fundamentally underconstrained: an image fixes only the visible hemisphere, leaving unobserved geometry and appearance to be inferred from information beyond the single input image.

Most existing systems infer unobserved content using learned priors, whether by per-instance distillation of a 2D diffusion prior (Poole et al., 2023; Lin et al., 2023), multi-view diffusion followed by reconstruction (Liu et al., 2023c;b), direct feed-forward regression (Hong et al., 2024; Xu et al., 2024; Tang et al., 2024), or native-3D generation over a 3D representation (Nichol et al., 2022; Jun & Nichol, 2023; Xiang et al., 2025; Hunyuan3D Team et al., 2025). Objectives and representations have improved across these approaches, but the main strategy is still to train a larger backbone on a larger 3D dataset. Curated 3D datasets remain orders of magnitude smaller than image datasets, so coverage remains limited in parts of the 3D asset distribution needed in practical applications, and scaling such priors typically requires further training.

Retrieval-augmented generation offers an alternative strategy: retrieving external evidence at inference time rather than encoding that information in model weights. This retrieval-based paradigm is particularly well suited to 3D generation: repositories such as ObjaverseXL (Deitke et al., 2023a) already contain millions of assets, and a retrieved mesh provides an explicit, complete shape rather than a shape embedding. RetDream (Seo et al., 2024) retrieves 3D assets to guide score distillation, and Phidias (Wang et al., 2025) conditions a 3D diffusion model on a retrieved reference, outperforming its reference-free counterpart. Both still require additional optimization: RetDream

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.

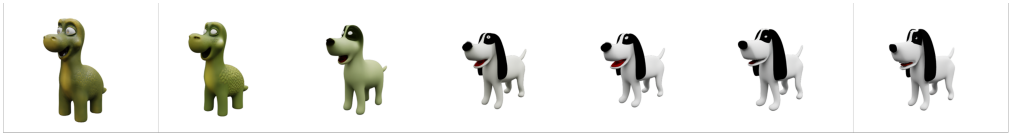

Figure 1: RM3D morphs a retrieved asset toward the input image inside a frozen sampler without updating weights.

performs per-instance optimization, whereas Phidias fine-tunes a reference-conditioned network to process the retrieved asset.

A retrieved asset can instead be incorporated with *no task-specific post-training*: no weight updates, no additional trainable parameters, and only fixed pretrained networks. RM3D adopts this setting. We use Interp3D's frozen three-pass morphing framework (Liu et al., 2026) as a starting point; configuration A retains only its basic image-endpoint interpolation procedure, omitting Interp3D's progressive semantic, structural, and texture alignment. Our method differs in how it selects the endpoints and controls the sampling trajectory: we retrieve the highest-similarity asset through a joint image–shape embedding space (Liu et al., 2023a) and use its *exact structured latent cached in advance* — precomputed offline by the frozen TRELLIS encoder (Xiang et al., 2025) — as the source endpoint, while the target is sampled from the input image alone (Figure 1). Expanding the database requires preprocessing but does not require model retraining.

A morph between two endpoints yields an intermediate state, whereas reconstruction must recover the target. We therefore add two parameter-free inference-time operations. Hungarian alignment establishes group-level correspondence between the retrieved asset's and the input image's DINOv2 conditions before interpolation, so that interpolation does not depend on the raw token order. We apply the alignment jointly with a correspondence-matched K/V memory in the early blocks. Target-only attention blocks then restrict the late denoising blocks' attention to the target's key/value memory.

Our contributions are threefold.

(i) To our knowledge, we introduce the first **post-training-free** retrieval-and-morphing pipeline for single-image 3D generation: retrieval supplies an exact database latent as the source endpoint of a frozen morphing sampler, without task-specific training or additional trainable parameters.

(ii) We introduce **Hungarian alignment** of condition tokens with correspondence-matched K/V and **target-only attention blocks** to steer reconstruction toward the input image without trainable parameters.

(iii) Extensive experiments demonstrate the superior qualitative and quantitative performance of RM3D.

2 Related Work

2.1 3D shape retrieval

Cross-modal retrieval provides the source asset for RM3D. CLIP-style contrastive pretraining (Radford et al., 2021) generalizes to 3D through joint embedding spaces aligning shapes with images and text (Liu et al., 2023a), built on self-supervised visual features (Oquab et al., 2024) and generated 3D–text pairs (Luo et al., 2023), and available 3D datasets expanded from curated CAD collections (Chang et al., 2015) to web-scale repositories (Deitke et al., 2023b;a). We adopt OpenShape unchanged and retrieve by cosine similarity, holding the encoder fixed so our database-scaling result measures database size rather than encoder quality. The nearest neighbour is never the object in the image, so our method focuses on how the retrieved asset is used.

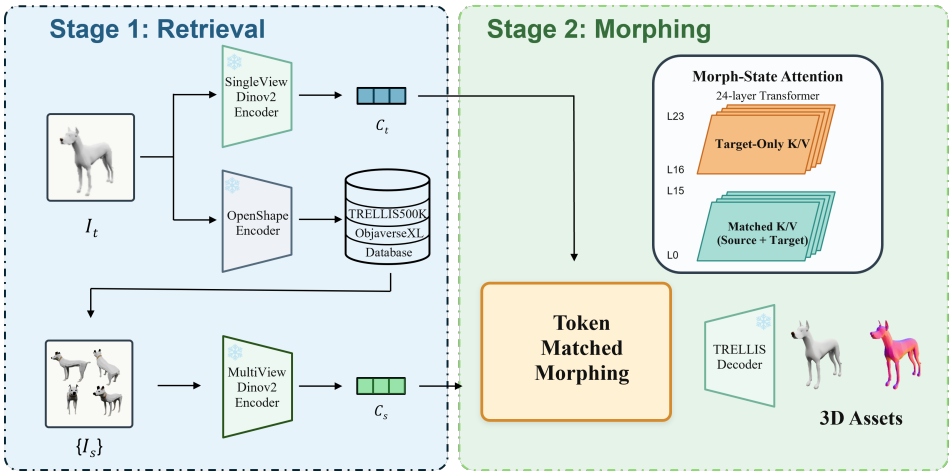

Figure 2: RM3D pipeline. The input image is encoded by OpenShape and used to retrieve the top-1 compatible asset from a filtered ObjaverseXL database; the retrieved asset contributes its *exact* cached SLAT as the source endpoint, while the input image is sampled through the frozen TRELLIS image-to-3D prior to form the target endpoint. Hungarian token alignment aligns the two DINOv2 conditions before interpolation. Each Euler step runs three passes (source, target, interpolation); blocks 0–15 attend to a correspondence-matched fusion of both endpoint memories, and the target-only attention blocks 16–23 attend to the target memory alone. The frozen SLAT decoder emits 3D Gaussians and a textured mesh. Every module is inference-time and parameter-free.

2.2 Feed-forward image-to-3D generation

Early single-image and text-to-3D generation optimized per instance, distilling 2D or multi-view diffusion priors into a 3D representation (Poole et al., 2023; Lin et al., 2023; Shi et al., 2024), with later work improving the objective and shape parameterization (Wang et al., 2023; Chen et al., 2023a; Qian et al., 2024; Xu et al., 2023); these optimization-based methods are general but require minutes to hours per asset. View-conditioned diffusion improved novel-view synthesis (Liu et al., 2023c; 2024b), while feed-forward pipelines reconstruct from predicted views (Liu et al., 2023b; Long et al., 2024; Liu et al., 2024a) or regress geometry and appearance directly, from LRM's triplane radiance field (Hong et al., 2024; Mildenhall et al., 2020) to meshes (Xu et al., 2024), 3D Gaussians (Tang et al., 2024; Kerbl et al., 2023), orthographic views (Wang et al., 2024) and sub-second inference (Tochilkin et al., 2024).

Native-3D generators instead model an explicit 3D representation, and their development has largely followed advances in representation design: point clouds (Nichol et al., 2022) and implicit functions (Jun & Nichol, 2023) were followed by compressed 3D latents inspired by latent diffusion on images (Rombach et al., 2022). TRELLIS (Xiang et al., 2025) couples a sparse voxel scaffold with per-voxel appearance features in a structured latent (SLAT) decoding to meshes, 3D Gaussians (Kerbl et al., 2023) or radiance fields, sampled by a rectified-flow transformer (Lipman et al., 2023; Liu et al., 2023d). Recent systems such as CLAY (Zhang et al., 2024b), Hunyuan3D 2.1 (Hunyuan3D Team et al., 2025), and SAM 3D (Chen et al., 2026) extend learned image-to-3D generation through larger models, datasets, and production-oriented outputs. All rely on training for their performance and are limited by the available 3D data. We use one, TRELLIS, entirely frozen.

2.3 Reference- and retrieval-augmented 3D generation

Retrieval-augmented generation is established in language modelling (Lewis et al., 2020), where a very large database lets a smaller model match much larger ones (Borgeaud et al., 2022) — the same trade-off that we explore in 3D — and it transferred to image generation (Chen et al., 2023b; Sheynin et al., 2023). In 3D, RetDream (Seo et al., 2024) retrieves 3D assets to initialize and regularize score distillation, and Phidias (Wang et al., 2025) is closest to our goal, conditioning a 3D diffusion model on a retrieved reference and clearly outperforming its reference-free counterpart.

Table 1: Single-image 3D generation on 500 randomly sampled Toys4k test instances. Appearance is evaluated over six rendered viewpoints, including front and back views; geometry uses 10K unit-box-normalized points. **Bold**: best.

	Appearance (six-view 3DGS renders)					Geometry	
Method	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	CLIP-P↑	CD↓	FScore↑
Phidias	18.16	0.854	0.198	0.851	0.882	0.1614	0.531
TRELLIS-S	19.96	0.894	0.095	0.952	0.952	0.1134	0.661
Hunyuan3D-2.1	19.71	0.888	0.115	0.893	0.893	0.1191	0.652
SAM 3D	18.23	0.877	0.143	0.878	0.876	0.1135	0.661
Ours (RM3D)	**20.98**	**0.904**	**0.092**	**0.953**	**0.953**	**0.1092**	**0.698**

The critical difference is training: RetDream still optimizes per instance, and Phidias fine-tunes a reference-conditioned network to use the retrieved asset. RM3D removes post-training entirely, so a larger database can improve results without retraining.

This is possible because a diffusion or flow sampler (Ho et al., 2020; Song et al., 2021) can be directly controlled at inference time. In 2D, training-free editing manipulates cross-attention (Hertz et al., 2023; Chefer et al., 2023), injects self-attention features (Tumanyan et al., 2023), borrows a source image's keys and values (Cao et al., 2023), or shares attention across a batch (Hertz et al., 2024); modifying the sampling trajectory instead of the weights predates these methods (Meng et al., 2022). Our target-only attention blocks follow this strategy, and the reported effect of layer depth on feature injection (Tumanyan et al., 2023) is why we restrict substitution to late blocks; the contrast with ControlNet, which freezes the backbone but *trains* an adapter (Zhang et al., 2023), distinguishes our post-training-free setting from adapter training. Diffusion morphing is likewise established in 2D (Zhang et al., 2024a; Yang et al., 2024), though both require endpoint-specific optimization. Interp3D (Liu et al., 2026) instead provides a training-free native-3D framework with aligned condition interpolation, SLAT-guided structure interpolation, and texture fusion. Our configuration A retains only its basic three-pass image-endpoint scaffold as a simplified baseline. Concurrent work also uses a frozen TRELLIS with inference-time attention biasing, but trains its fusion branch (Lin et al., 2026).

3 METHOD

RM3D reconstructs a textured 3D asset from a single target image. It retrieves a compatible source from an offline asset database, constrains sparse geometry using the source's exact SLAT, and reconstructs target appearance through source-guided asymmetric attention. All pretrained weights remain frozen and RM3D adds no trainable parameters.

3.1 PRELIMINARIES

TRELLIS (Xiang et al., 2025) represents an asset by a structured latent (SLAT) $Z = (P, F)$, where $P \subset \mathbb{Z}^3$ contains occupied sparse coordinates and $F \in \mathbb{R}^{|P|\times 8}$ contains their features. Its frozen pipeline first predicts an occupied structure and then samples SLAT features with two rectified-flow models (Liu et al., 2023d; Lipman et al., 2023); frozen decoders produce 3D Gaussians and a textured mesh. Starting from the three-pass sampling framework used by Interp3D (Liu et al., 2026), RM3D modifies endpoint construction, coordinate support, and attention memories during inference.

3.2 SINGLE IMAGE RETRIEVAL

For each offline database asset A_m, we precompute an OpenShape (Liu et al., 2023a) embedding e_m, an exact SLAT $Z_m^{\text{exact}} = (P_m, F_m)$, and cached view-059 DINOv2 condition tokens c_m. Given

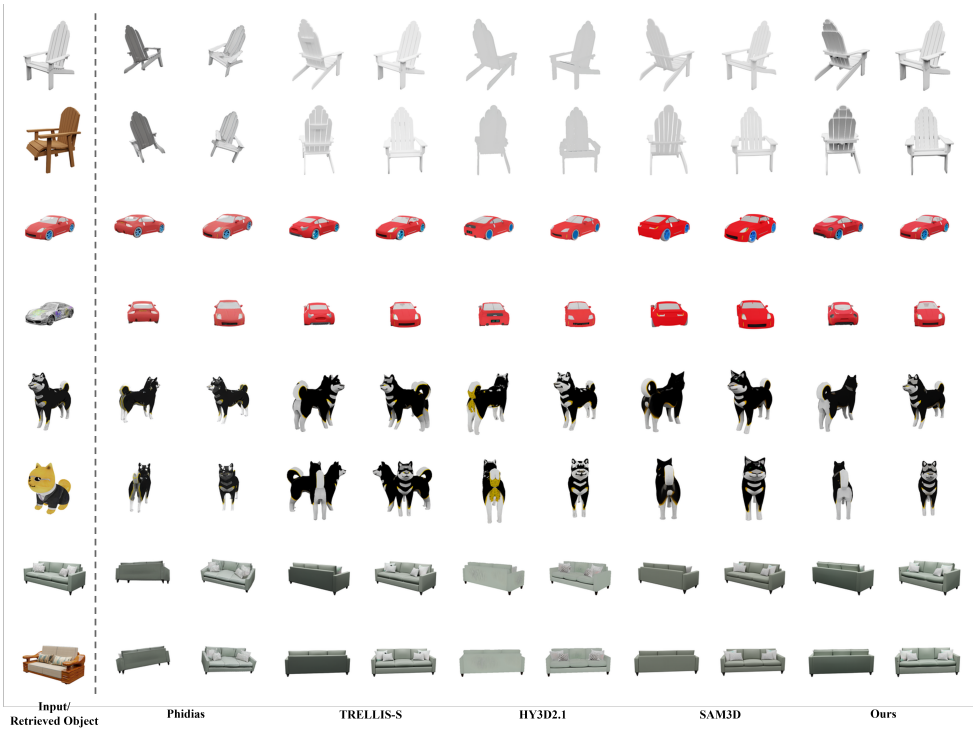

Figure 3: Qualitative comparison on four of the 500 sampled test instances. In each two-row block, the first column contains the input (top) and retrieved object (bottom); each method cell shows four views selected from the six-view evaluation set, which includes front and back viewpoints. RM3D better preserves input appearance and cross-view structure.

the target image I_t, retrieval selects

$$r = \arg\max_m \frac{e_t^\top e_m}{\|e_t\|_2 \|e_m\|_2}, \qquad e_t = \phi_{\text{img}}(I_t). \tag{1}$$

During evaluation, RM3D uses the precomputed top-1 retrieval result to load $Z_s^{\text{exact}} = (P_s, F_s)$ and the cached source condition c_s.

At generation time, foreground processing and resizing map I_t to 518×518, after which frozen DINOv2 produces $c_t = \text{LN}(\psi_{\text{DINO}}(I_t))$, and the frozen single-image prior samples

$$Z_t = (P_t, F_t) \sim p_\theta(Z \mid c_t). \tag{2}$$

Z_t supplies target descriptors and support for the subsequent morphing process. The online condition consists of the single target image I_t.

3.3 Exact SLAT Guided Morphing

Source-SLAT-Anchored Geometry.

At each structure-flow step, RM3D aligns grouped source and target condition tokens with a cosine-cost Hungarian assignment (Kuhn, 1955). Over $N = 25$ steps, the group size $g_n = \lceil 32(0.001)^{n/(N-1)} \rceil$ decreases from 32 to 1, and the condition at ratio α is

$$c_\alpha^{(n)} = (1-\alpha)\tilde{c}_s^{(n)} + \alpha \tilde{c}_t^{(n)}. \tag{3}$$

All condition tokens are retained; the assignment changes only their group order.

The exact source SLAT also supplies geometry-aware descriptors. Coordinates from P_s and P_t are mapped from the 64^3 sparse grid to 16^3 voxels by $\nu(p) = \lfloor p/4 \rfloor$, and features falling in the

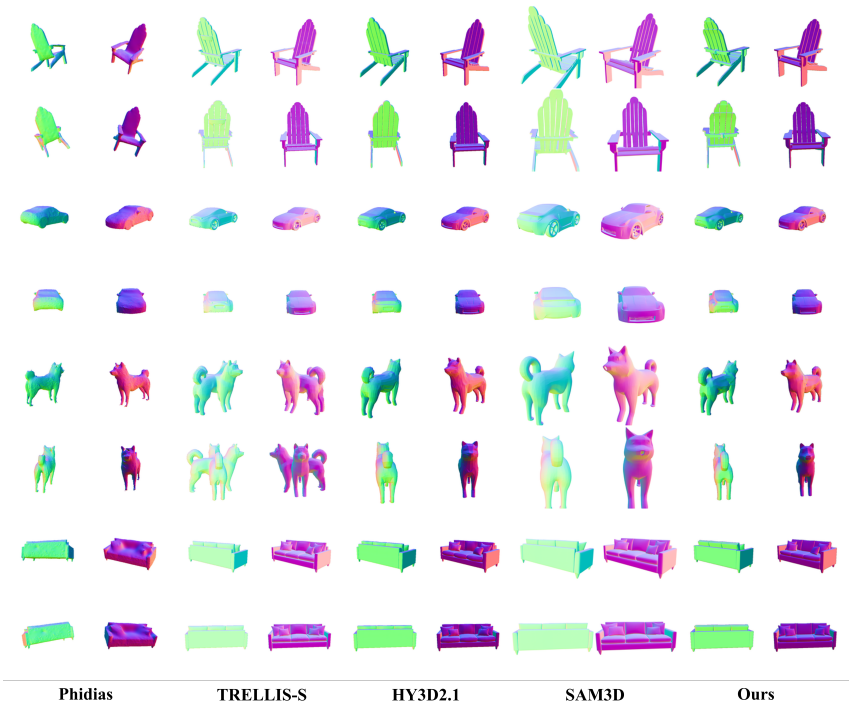

Figure 4: Normal maps for the four mesh-exporting methods on the same four of the 500 randomly sampled comparison instances as Figure 3, showing the front and back viewpoints from the six-view evaluation set.

same voxel are mean-pooled. Progressively finer groups use $g_n^{\text{geo}} = \lceil 64(0.001)^{n/(N-1)} \rceil$, with assignment

$$\Pi_{\text{geo}}^{(n)} = \arg\min_{\Pi} \sum_{g} \left[1 - \cos\left(\bar{f}_s^{\,g}, \bar{f}_t^{\,\Pi(g)} \right) \right]. \tag{4}$$

Pairs below $\tau_{\text{geo}} = 0.6$ are rejected. The surviving assignment reorders the target endpoint memories into source-aligned locations at each flow step. Exact SLAT features serve as correspondence descriptors.

Source, target, and current structure streams start from the same dense Gaussian noise. The current Query attends independently to memory banks that combine the source with the current state and the aligned target with the current state:

$$A_s^l = \text{Attn}\big(Q_c^l, [K_s^l; K_c^l], [V_s^l; V_c^l]\big), \tag{5}$$

$$A_t^l = \text{Attn}\big(Q_c^l, [\mathcal{R}_{\Pi_{\text{geo}}^{(n)}}(K_t^l); K_c^l], [\mathcal{R}_{\Pi_{\text{geo}}^{(n)}}(V_t^l); V_c^l]\big). \tag{6}$$

Each block interpolates the full attention outputs,

$$H_{\text{geo}}^l = (1 - \alpha_g) A_s^l + \alpha_g A_t^l, \qquad \alpha_g = 0.9. \tag{7}$$

After 25 Euler steps, occupancy decoding yields the sparse support $P_g = P_{0.9}$. If the predicted support is empty, RM3D uses the nearest nonempty sampled support and records this fallback. The resulting coordinate support is P_g.

Target Reconstruction with a Source Prior.

The reconstruction initializes new source, target, and current streams from Gaussian noise on P_s, P_t, and P_g, respectively. At every step and block, endpoint passes cache (K_s^l, V_s^l) and (K_t^l, V_t^l) before the current reconstruction pass. Its Query always comes from the evolving state on P_g. Cross-attention receives the full target condition $c_\star = c_t$ ($\alpha_c = 1$), while the geometry mixing weight remains $\alpha_g = 0.9$ throughout reconstruction.

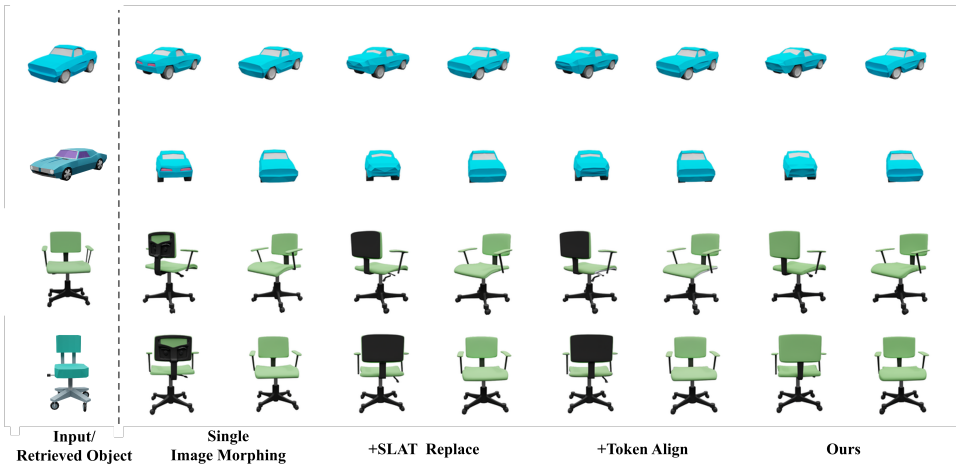

Figure 5: Qualitative ablation on two of the 200 randomly sampled test instances. For each instance, the first column contains the input image (top) and retrieved object (bottom); each method group spans two columns showing the front and back viewpoints from the six-view evaluation set. **A**: Single Source Image to Target Image baseline. **B**: +exact retrieved SLAT source endpoint. **C**: +grouped Hungarian condition-token alignment and correspondence-aware K/V matching and fusion. **D**: +target-only attention refinement (full RM3D). The dark seat backs in A and B are corrected only once D is enabled.

For each current key $K_c^l[u]$, hard endpoint matching chooses

$$j_e^\star(u) = \arg\max_j \cos\big(K_c^l[u], K_e^l[j]\big)\,, \qquad e \in \{s, t\}. \tag{8}$$

Hard matching evaluates three candidates and copies the highest-scoring token when its similarity exceeds $\tau = 0.6$; otherwise it retains the current token. This produces matched keys $\widehat{K}_s^l, \widehat{K}_t^l$ and a target value $\widehat{V}_t^l$ selected by the same target key index.

For blocks $l = 0, \ldots, 15$, the active key memory is

$$K_{\text{mix}}^l = \frac{1}{11}\widehat{K}_s^l + \frac{10}{11}\widehat{K}_t^l + \frac{10}{11}K_c^l, \tag{9}$$

$$K_{\text{out}}^l = \text{RMSNorm}_h\left(\frac{\|K_c^l\|_F}{\|K_{\text{mix}}^l\|_F}K_{\text{mix}}^l\right), \qquad V_{\text{out}}^l = \widehat{V}_t^l. \tag{10}$$

The global norm of the fused key is matched to K_c^l before per-head RMS normalization. Accepted source keys contribute with relative weight 0.1 and affect where current Queries attend. Accepted target matches supply the output values, while rejected matches use the current Value.

Blocks $l = 16, \ldots, 23$ perform target-focused refinement:

$$K_{\text{out}}^l = \text{RMSNorm}_h(K_t^l), \qquad V_{\text{out}}^l = V_t^l. \tag{11}$$

These blocks use target K/V, while their Queries retain the geometry, early-block updates, and residual state accumulated on P_g.

3.4 Flow Sampling and Decoding

All three streams evolve for $N = 25$ Euler steps under the fixed attention rules above:

$$x_{n+1} = x_n - (t_n - t_{n+1})\,\widehat{v}_\theta(x_n, t_n, c_t; \mathcal{M}_{s,n}, \mathcal{M}_{t,n}, P_g)\,, \tag{12}$$

where $\widehat{v}_\theta$ applies classifier-free guidance with strength 5 for $t \in [0.5, 1.0]$, i.e., $v_{\text{CFG}} = 6v_c - 5v_\varnothing$. The time grid uses $\rho = 3$ and the flow noise floor is $\sigma_{\min} = 10^{-5}$. Conditional endpoint caches are restored after each unconditional pass. The final sample is denormalized using the pretrained TRELLIS statistics and decoded as

$$Z_\star^{\text{tex}} = \text{Denorm}(x_{25}), \qquad G^\star = D_{\text{SLAT}}(Z_\star^{\text{tex}}). \tag{13}$$

The decoders output 3D Gaussians and a complete textured GLB mesh with a 4096^2 texture.

Table 2: Cumulative ablation on Toys4k. A: Single Source Image to Target Image baseline. B: A with its image-sampled Source SLAT replaced by the exact retrieved SLAT. C: B + grouped Hungarian condition-token alignment and correspondence-aware K/V matching and fusion. D: C + target-only attention refinement (full RM3D). All results use the same 200 randomly sampled Toys4k test instances. Appearance metrics are averaged over six rendered viewpoints, including front and back views; CD and FScore measure geometry. **Bold**: best.

	Appearance					Geometry	
Config	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	CLIP-P↑	CD↓	FScore↑
A	19.02	0.864	0.135	0.829	0.901	0.1212	0.618
B	19.49	0.883	0.121	0.881	0.924	0.1168	0.623
C	19.86	0.892	0.116	0.881	0.931	0.1161	0.631
D	**20.08**	**0.896**	**0.096**	**0.948**	**0.948**	**0.1127**	**0.643**

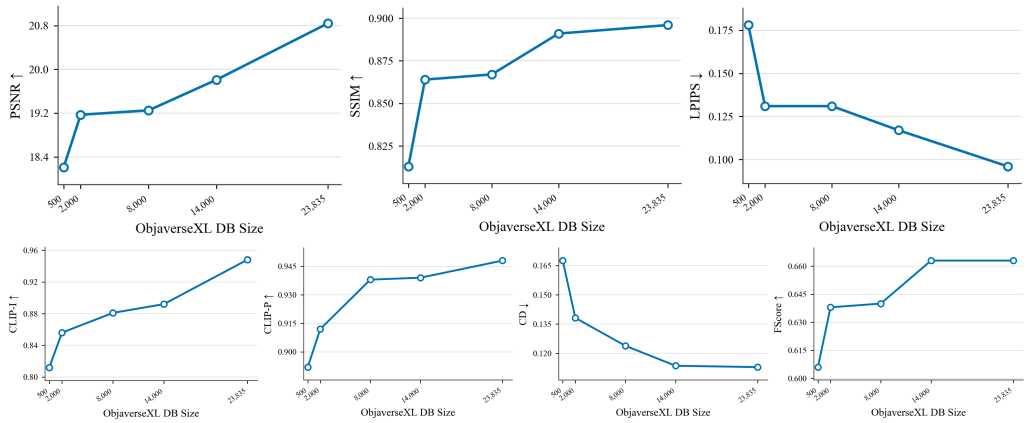

Figure 6: Scaling across five ObjaverseXL database sizes (500, 2,000, 8,000, 14,000, 23,835 assets), evaluated on the same 200 randomly sampled test instances using six rendered viewpoints, including front and back views.

4 Experiments

4.1 Dataset

Retrieval database. We build it from the ObjaverseXL (Sketchfab) (Deitke et al., 2023a) portion of TRELLIS-500K (Xiang et al., 2025), the collection curated for training TRELLIS. That split contains 168,307 instances, more than required for retrieval, so we retain only assets with AestheticScore above 6.5, giving 23,835 assets.

Test set. Toys4k (Stojanov et al., 2021) provides ground-truth meshes and multi-view renders on a fixed camera grid; we use 1,000 instances spanning 27 categories. No Toys4k asset appears in the retrieval database, so the retrieved source and the evaluation ground truth are always different objects.

4.2 Comparison

Baselines. TRELLIS-S, the Single Image version of TRELLIS (Xiang et al., 2025), uses the same frozen backbone and so gives the most direct measure of what retrieval-and-morphing adds. We also compare with Hunyuan3D-2.1 (Hunyuan3D Team et al., 2025), SAM 3D (Chen et al., 2026), and Phidias (Wang et al., 2025), the closest reference-augmented system, which fine-tunes a network to use the retrieved asset. All four rely on training for their performance; RM3D adds no trainable weights.

Metrics. All comparison metrics use a fixed random sample of 500 of the $1{,}000$ Toys4k test instances. Appearance is measured on renders produced from the model's native 3D Gaussians, reporting PSNR↑, SSIM↑, LPIPS↓, CLIP-I↑ (image–image against the input) and CLIP-P↑ (prompt–image against the class prompt); the two use separate code paths and separate inputs. Geometry is measured by Chamfer Distance (CD↓) and FScore↑ at $\tau{=}0.1$ on $10{,}000$ surface points normalised to a unit bounding box. Each run takes exactly one image as input, either view 059 or view 061, never both. We render and evaluate each reconstruction from six viewpoints, including both front and back views.

Qualitative Comparison. Figure 3 shows four examples; its first column contains the input and retrieved object, and each method cell contains four of the six evaluation views. RM3D better preserves input appearance and cross-view structure. The front and back normal maps in Figure 4 show cleaner, more complete surfaces with fewer unwanted folds.

Quantitative Comparison. Table 1 reports appearance on 3D-Gaussian renders and geometry against the Toys4k ground-truth point clouds. RM3D is best on every reported metric, including the updated 20.98 PSNR, 0.904 SSIM, 0.1092 CD, and 0.698 FScore. For geometry, RM3D attains the best CD (0.1092) and FScore (0.698) of the five methods, against 0.1614 and 0.531 for Phidias. The normal-map comparison is consistent with this ordering.

4.3 Ablation Study

Both ablation experiments use the same fixed random subset of 200 Toys4k test instances and the same six-view evaluation protocol, including front and back views.

Component Ablation. We compare four cumulative configurations that follow the mechanisms defined in Sec. 3.2 and Sec. 3.3. A runs the original Interp3D Source-to-Target morph, in which the Source endpoint SLAT is sampled from the Source image. B changes only this source-side input by replacing the image-sampled SLAT with the retrieved asset's cached exact SLAT $Z_s^{\text{exact}} = (P_s, F_s)$. The Target endpoint and all sampling settings remain unchanged, including Gaussian initialization of the flow streams; here, "endpoint" denotes the Source SLAT supplied to the morph rather than the initial flow state. C further adds grouped Hungarian condition-token alignment and correspondence-aware K/V matching and fusion. D adds target-only attention refinement, forming the full RM3D model. Table 2 reports the quantitative results. In Figure 5, the first column contains the input image (top) and retrieved object (bottom), while each configuration occupies two columns showing the front and back viewpoints. Configuration D performs best on all seven metrics. B improves every metric over A; C further improves PSNR, SSIM, LPIPS, CLIP-P, CD, and FScore while retaining the same 0.881 CLIP-I. D improves every metric over C, including LPIPS ($0.116{\rightarrow}0.096$) and CLIP-I ($0.881{\rightarrow}0.948$). Across A–D, CD decreases from 0.1212 to 0.1127, while FScore increases from 0.618 to 0.643, with geometry improving at each cumulative step.

Scaling with the Retrieval Database. Because RM3D is training-free, the only way to improve RM3D without retraining is to enlarge the database. Figure 6 evaluates five database sizes from 500 to $23{,}835$ assets on this 200-instance subset, all else fixed. No metric worsens as the database grows, with plateaus rather than gains at some steps. The first expansion to $2{,}000$ assets gives the largest gains in SSIM, LPIPS, CD, and FScore, because below a few thousand assets retrieval often fails to return a within-category source; larger databases continue to improve the semantic and geometric metrics, while the perceptual metrics plateau. Since no weights change between points, these changes reflect retrieval alone, with no clear saturation at the largest database size.

5 Conclusion

We presented RM3D, a training-free retrieval-and-morphing framework for single-image 3D generation. It uses a retrieved asset's exact structured latent as the source endpoint for morphing toward the input image inside a frozen native-3D generator, and reduces the resulting bias with two parameter-free modules, Hungarian token alignment and target-only attention blocks. It adds no trainable weights, yet matches or exceeds strong trained baselines on Toys4k, and no metric degrades as the database grows. Quality does depend on retrieval, and the three-pass sampler requires more computation than a single forward pass.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, with a focus on improving wording, grammar, and readability. These tools also assisted with adjusting the document layout and refining table formatting to improve alignment, consistency, and visual presentation.

REPRODUCIBILITY STATEMENT

We will publicly release the source code for RM3D upon acceptance of the paper. This release is intended to support independent reproduction of the method and facilitate further research on training-free, retrieval-guided single-image 3D generation.

REFERENCES

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning (ICML)*, 2022.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4), 2023.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *International Conference on Learning Representations (ICLR)*, 2023b.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7220–7232, 2026.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, et al. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yicong Hong, Kai Zhang, Jiuxiang Gu, et al. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024.

Hunyuan3D Team, Shuhui Yang, Mingxin Yang, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025.

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chen-Hsuan Lin, Jun Gao, Luming Tang, et al. Magic3d: High-resolution text-to-3d content creation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Siyou Lin, Zhou Xue, Hongwen Zhang, et al. Mix3r: Mixing feed-forward reconstruction and generative 3d priors for joint multi-view aligned 3d reconstruction and pose estimation. *arXiv preprint arXiv:2605.03359*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, et al. Openshape: Scaling up 3d shape representation towards open-world understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, et al. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, et al. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023c.

Xiaolu Liu, Yicong Li, Qiyuan He, et al. Interp3d: Correspondence-aware interpolation for generative textured 3d morphing. *arXiv preprint arXiv:2601.14103*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023d.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, et al. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations (ICLR)*, 2024b.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, et al. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Junyoung Seo, Susung Hong, Wooseok Jang, Ines Hyeonsu Kim, Min-Seop Kwak, Doyup Lee, and Seungryong Kim. Retrieval-augmented score distillation for text-to-3d generation. In *International Conference on Machine Learning (ICML)*, 2024.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, et al. Knn-diffusion: Image generation via large-scale retrieval. In *International Conference on Learning Representations (ICLR)*, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, et al. Mvdream: Multi-view diffusion for 3d generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1798–1808, 2021.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision (ECCV)*, 2024.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, et al. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, et al. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, et al. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision (ECCV)*, 2024.

Zhenwei Wang, Tengfei Wang, Zexin He, et al. Phidias: A generative model for creating 3d content from text, image, and 3d conditions with reference-augmented diffusion. In *International Conference on Learning Representations (ICLR)*, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, et al. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, et al. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Zhaoyuan Yang, Zhengyang Yu, Zhiwei Xu, Jaskirat Singh, Jing Zhang, Dylan Campbell, et al. Impus: Image morphing with perceptually-uniform sampling using diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Kaiwen Zhang, Yifan Zhou, Xudong Xu, Xingang Pan, and Bo Dai. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, et al. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics*, 43(4), 2024b.

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 (ICCV)*, 2023.