JanusMesh: Fast and Zero-Shot 3D Visual Illusion Generation via Cross-Space Denoising

Siang-Ling Zhang, Huai-Hsun Cheng, Tsung-Ju Yang, and Yu-Lun Liu

National Yang Ming Chiao Tung University

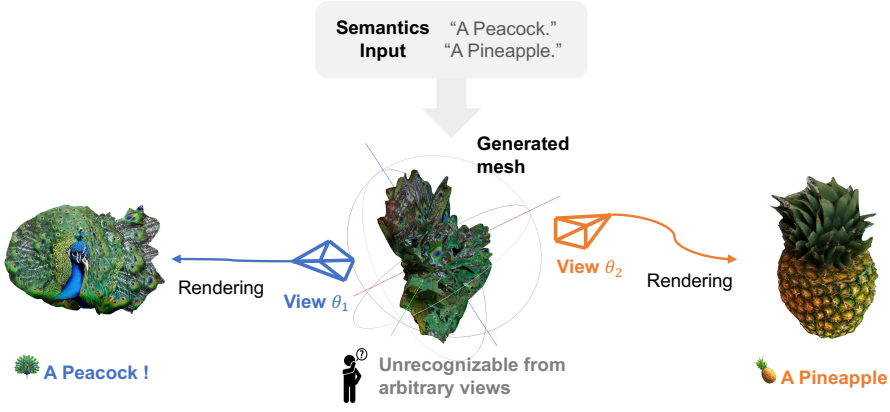

Fig. 1: Zero-shot 3D Visual Illusion Generation. Given two different text prompts, our method creates a single 3D mesh that embodies a dual-semantic visual illusion. The generated shape appears unrecognizable from arbitrary viewpoints, but it perfectly reveals the two target semantics (*e.g.*, a peacock and a pineapple) when observed from specific camera angles (View θ_1 and View θ_2). Our framework achieves this intricate 3D illusion efficiently without requiring per-shape optimization.

Abstract. Creating 3D visual illusions, a single 3D mesh that reveals entirely different semantics from various viewing angles, is a fascinating but tough challenge. Existing optimization-based methods are slow and can produce oversaturated colors. In contrast, naive stitching approaches fail to produce geometrically coherent objects. This results in visible unnatural seams and semantic leaks. In this paper, we present a fast and training-free framework for generating text-driven 3D visual illusions. Our approach decouples the generation into two stages. First, we propose a cross-space dual-branch denoising process. This process dynamically decodes 3D latents into voxel space for CLIP-guided orientation alignment and Signed Distance Field (SDF) blending, which ensures seamless geometric fusion. Second, we introduce a view-conditioned texture synthesis module that projects and aggregates view-specific 2D diffusion priors onto the fused geometry. Extensive experiments demonstrate that our method generates highly realistic, dual-semantic 3D illusions in just 3–5 minutes. It significantly outperforms existing methods in geometric integrity, semantic recognizability, and efficiency. Project page: `https://siang1105.github.io/JanusMesh.github.io/`

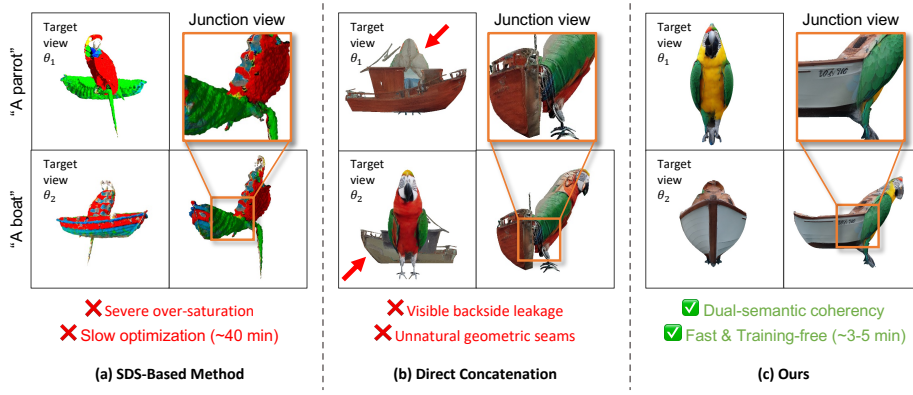

Fig. 2: Comparison of 3D visual illusion generation methods. (a) **SDS-Based Methods** suffer from severe over-saturation and slow optimization. (b) **Direct Concatenation** exposes unnatural geometric seams and semantic leakage at target views (**red arrows**). (c) **Our method** creates a seamless, dual-semantic coherent 3D mesh. Unlike previous approaches, our method does not require training. It generates high-quality 3D visual illusions in just 3–5 minutes while completely preventing geometric interference between the two semantics.

Keywords: 3D Visual Illusion · Dual-Branch Denoising · View-Conditioned Texturing

1 Introduction

Visual illusions have long fascinated human perception by challenging our understanding of physical reality. While diffusion models have enabled computational optical illusions in 2D, bringing this concept into the 3D world—creating a single object that presents entirely different semantics depending on the viewing angle—remains a formidable challenge. As illustrated in Fig. 1, our goal is zero-shot *3D visual illusion generation*: synthesizing a single 3D mesh that presents, for instance, a peacock from θ_1 and a pineapple from θ_2, while appearing as an abstract geometry from arbitrary angles.

Earlier 3D optical illusion works focused on 2D projections or surface patterns such as shadow art, wireframe silhouettes, or height fields. With text-to-image diffusion models, zero-shot 2D illusion methods (e.g., Visual Anagrams [21]) synchronize noise predictions across views to create multi-interpretation images. Extending this to true 3D meshes, existing approaches rely on SDS to optimize a 3D representation so its renders match different prompts at different viewpoints [39].

However, existing methods fall short in producing high-quality 3D illusions efficiently. As shown in Fig. 2, optimization-based approaches (e.g., Shape From Semantics [39]) require ∼40 minutes per shape and suffer from severe color over-saturation. A naive Direct Concatenation of two separately generated objects produces unnatural geometric seams and visible backside leakage, breaking the perceptual illusion.

We present a zero-shot two-stage framework that generates coherent 3D visual illusions in 3–5 minutes. Stage 1 employs a dual-branch denoising process using TRELLIS [73], decoding latents into voxel space at each step, aligning objects via CLIP-guided orientation search, and merging them via SDF blending before re-encoding. Stage 2 performs view-conditioned texturing by projecting Stable Diffusion predictions onto the fused mesh. Our method demonstrates superior geometry coherence, texture realism, and semantic recognizability over existing baselines.

In summary, our main contributions are threefold:

- A zero-shot framework extending generative multi-view illusions from 2D to fully textured 3D meshes.
- A training-free two-stage architecture featuring dual-branch denoising with SDF blending and CLIP-guided alignment for geometric integrity, coupled with view-conditioned texturing for dual-semantic coherency.
- A rigorous evaluation protocol incorporating CLIP, GPT-4.1-mini, FID/KID, and a novel Object Detection metric, with experiments validating our method over baselines and demonstrating scalability to three-object illusions.

2 Related Work

Computational Optical Illusions. Prior work has explored appearance-varying 3D objects, including shadow art [56, 64, 69], wire art [26, 60], view-dependent heightfields [58], SDF-based anamorphic packing [12], and spatially ambiguous geometry [15]; in 2D, spatial-frequency decomposition [20, 57] shifts perceived content with viewing distance, while progressive vector sketching [10] transforms perceived semantics through sequential stroke addition. These approaches produce 2D projections, surface patterns, or—as with adversarially injected viewpoint-dependent content in 3D Gaussian Splatting [33]—view-specific artifacts, whereas our work generates a fully textured 3D mesh with distinct semantics from different viewpoints.

Illusion Generation with Diffusion Models. Text-to-image diffusion models [23, 63, 66] have enabled new illusion synthesis. SDS-based methods [3, 59] optimize for multiple prompts but converge slowly. Visual Anagrams [21] introduced zero-shot illusions via per-view noise averaging, extended to frequency decompositions [20], phase-transfer [19], and audio-visual spectrograms [9]; in 3D, Illusion3D [18] and LookingGlass [5] lift these priors into NeRF and anamorphic images, respectively. Our work extends the zero-shot spirit of [21] to native 3D latent space, producing a fully textured mesh.

Text-to-3D Generation. Optimization-based methods [7, 45, 59, 70] distill 2D diffusion priors via SDS, with improvements via interval score matching [44], rectified-flow distillation [75], and Gaussian acceleration [67, 77]. Feed-forward methods combine multi-view diffusion [48, 50, 52, 65] with reconstruction networks [25, 71, 74]. Application-driven variants extend generation to scene scale and alternative styles: training-free pipelines complete fully textured room-scale

meshes from sparse images [41], diffusion priors enable city-scale 3D scene creation with iterative geometry and texture refinement [37], and differentiable mesh optimization under 2D pixel-art supervision produces stylized 3D content [27]. Native 3D generative models [32, 36, 43, 72, 81] learn latent representations directly from 3D data; TRELLIS [73] encodes geometry and appearance in a sparse structured latent space via rectified flow [49], which we repurpose for dual-semantic mesh generation.

Synchronized Diffusion Denoising. Merging denoising trajectories enables compositional generation [47], seamless panoramas [1], and perceptual synchronization [38], with advanced samplers addressing compositional failures [16, 35]. Recent works average clean-image predictions rather than noise [34] and apply spatial guidance in 3D latents [17]. Closely related, multi-view diffusion outpainting coordinates denoising across views with geometry-aware strategies for sparse-view reconstruction [28], and 3D-aware 360° video diffusion decouples texture refinement from a 3D cache that enforces geometric consistency [8]. Unlike these consistency-enforcing frameworks, our dual-branch denoising uniquely enforces divergent semantics at target viewpoints via SDF fusion.

3D Texture Synthesis. Diffusion-based mesh texturing spans depth-conditioned inpainting [62], ControlNet-enhanced generation [6, 79], multi-view consistent UV-latents [4, 11, 51], and single-pass feed-forward models [78] with material [14] or appearance [76] extensions. Crucially, all these methods apply a single prompt uniformly. Our view-conditioned synthesis instead assigns distinct prompts to different angular sectors, back-projecting viewpoint-specific clean images via cosine-weighted blending.

CLIP-Guided 3D Understanding and Generation. CLIP's [61] render-caption similarity enables zero-shot 3D generation [30], mesh stylization [54], NeRF manipulation [31, 68], and latent score distillation [53]. Leveraging this render-and-score paradigm, our CLIP-guided Orientation Search automatically selects the relative rotation that maximizes silhouette alignment between the two objects. This resolves geometric mismatches that would otherwise cause SDF fusion failures.

3 Method

3.1 Preliminaries

TRELLIS: Structured 3D Latent Representation. TRELLIS [73] is a two-stage Rectified Flow [46] 3D generator that first predicts a low-resolution sparse voxel structure, then refines it with high-dimensional appearance features. We perform geometry blending specifically during this first structural stage. Because direct spatial transformations distort latent distributions, we draw inspiration from LookingGlass [5]: at each denoising step, we decode the latent into voxel space, perform SDF blending, and re-encode the fused result back to the latent space. This cross-space denoising ensures the geometric validity of the blended mesh.

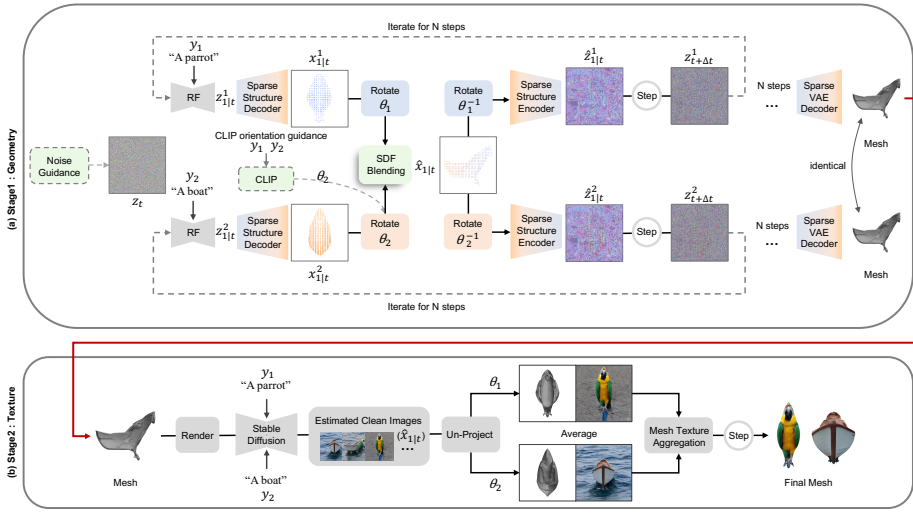

Fig. 3: Pipeline overview. (a) Stage 1 employs dual-branch denoising. At each step, latents are decoded to voxel space, rotation-aligned, and fused via SDF blending (Fig. 4), then re-encoded to continue denoising, producing a single unified 3D mesh. **(b) Stage 2** applies view-conditioned texturing to the fused mesh. Estimated clean images $\hat{x}_{1|t}$ are predicted via Stable Diffusion, un-projected from viewpoints θ_1 and θ_2, and iteratively aggregated via Mesh Texture Aggregation, producing a single object with distinct semantics at each target viewpoint.

Visual Anagrams. Visual Anagrams [21] generates 2D illusions by averaging noisy estimates across transformed views in a canonical space. Because this restricts transformations to be orthogonal, SyncTweedies [34] relaxes the constraint by averaging estimated clean latents instead: $\hat{\mathbf{z}}^i_{1|t} = \pi_i(\frac{1}{N}\sum_j \pi_j^{-1}(\mathbf{z}^j_{1|t}))$. We extend this synchronization principle from 2D pixels to 3D voxels, blending predicted clean geometries via SDF averaging to generate view-dependent dual-semantic meshes.

3.2 Overview

Given prompts y_1, y_2 and target viewpoints θ_1, θ_2, our framework generates a single 3D mesh exhibiting y_1 at θ_1 and y_2 at θ_2. A successful illusion requires: (1) **Semantic Recognizability** at target views; (2) **Geometric Integrity** from any viewpoint; and (3) **Illusion Effect** (concealing opposing semantics at non-target views), all evaluated in Sec. 4.1. As shown in Fig. 3, our pipeline achieves this via Stage 1 dual-branch geometric denoising (Sec. 3.3) and Stage 2 view-conditioned texturing (Sec. 3.4).

3.3 Dual-Branch Geometry Generation (Stage 1: Geometry)

Dual-Branch Denoising and Clean Latent Estimation. Stage 1 builds on TRELLIS's Rectified Flow, starting from a shared initial noise z_t and performing two independent denoising branches conditioned on y_1 and y_2, estimating the

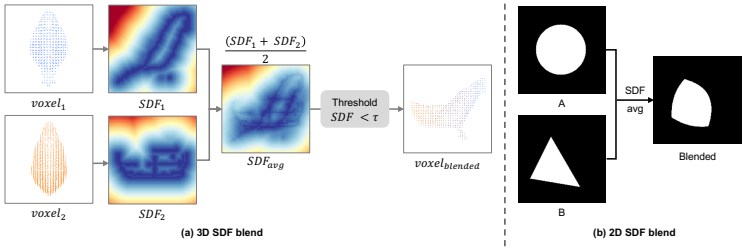

Fig. 4: SDF blending. (a) Given two rotation-aligned voxels, we compute their SDFs, take the element-wise average, and binarize with threshold τ to obtain the blended voxel. (b) 2D illustration: averaging the SDFs of a circle (A) and a triangle (B) produces a smooth intermediate contour lying geometrically between the two.

clean latent at each timestep t:

$$x^1_{1|t} = z_t + u_\theta(z_t;\, t,\, y_1)(1-t), \tag{1}$$

$$x^2_{1|t} = z_t + u_\theta(z_t;\, t,\, y_2)(1-t), \tag{2}$$

where u_θ is the Rectified Flow Network. We apply Classifier-Free Guidance (CFG) [24] with Interval CFG to improve generation quality while avoiding over-saturation at extreme noise levels.

Geometry Blending in Voxel Space. The clean latent estimates $x^1_{1|t}$ and $x^2_{1|t}$ are decoded via the Sparse Structure Decoder into voxel representations v_1 and v_2. We rotate v_2 by θ_2 to align both voxels in a common reference frame. Since directly averaging binary occupancy grids lacks geometric continuity, we convert both into Signed Distance Fields (SDFs), average them element-wise, and binarize with threshold τ to obtain the blended geometry $\hat{x}_{1|t}$:

$$\mathrm{SDF}_{\mathrm{blend}} = \frac{\mathrm{SDF}(v_1) + \mathrm{SDF}(v_2)}{2}, \tag{3}$$

$$\hat{x}_{1|t} = [\mathrm{SDF}_{\mathrm{blend}} < \tau]. \tag{4}$$

The blending process is illustrated in Fig. 4. After blending, $\hat{x}_{1|t}$ is rotated by $-\theta_2$ to restore v_2's original coordinate frame, then re-encoded into $\hat{z}^1_{1|t}$ and $\hat{z}^2_{1|t}$ via the Sparse Structure Encoder for the next denoising step.

3.4 View-Conditioned Texture Synthesis (Stage 2: Texture)

View-Aware Texture Prediction. Because the fused Stage 1 mesh contains unnatural geometry, direct TRELLIS texturing fails. Therefore, we treat texturing as an independent stage using a depth-conditioned ControlNet [80] (Stable Diffusion [63]) to predict clean images $\hat{x}_{1|t}$. At each denoising step, we render the mesh from θ_1 and θ_2, predict textures, and un-project them onto the 3D surface. Finally, Mesh Texture Aggregation merges these multi-view contributions using cosine-weighted blending (based on surface normals), ensuring high-quality, seamless textures.

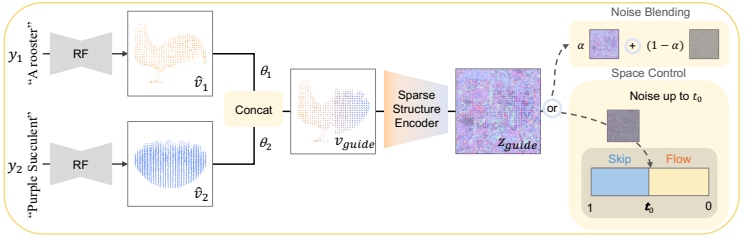

Fig. 5: Noise Guidance. Given two prompts y_1 and y_2, two single-semantic voxels $\hat{v}_1$ and $\hat{v}_2$ are independently generated and concatenated at θ_1 and θ_2 to form v_{guide}, encoded into z_{guide} via the Sparse Structure Encoder. **Noise Blending Guidance** mixes z_{guide} with pure noise via $\alpha \cdot z_{\text{guide}} + (1-\alpha) \cdot z_{\text{noise}}$, injecting a mild spatial prior. **Space Control Guidance** interpolates between z_{guide} and noise at timestep t_0, providing stronger structural constraints for geometrically challenging pairs.

View-Dependent Texture Selection. We determine which branch's texture to use based on the current camera angle. Taking $\theta_1 = 0°$ as an example, viewpoints within 270°–90° adopt the texture estimate from y_1, while the remaining angles adopt that from y_2. Although switched via a hard cutoff, no visible seam appears in practice, as cosine-weighted blending naturally smooths the transition at the boundary.

3.5 Noise Guidance

The starting point of Rectified Flow denoising—pure random noise—lacks spatial structural constraint for dual-semantic fusion, making it prone to geometric interference and difficult convergence. We propose two noise guidance strategies that inject a spatial prior before denoising begins, improving geometric fusion quality, as illustrated in Fig. 5.

Noise Blending Guidance. We pre-generate two single-semantic voxels $\hat{v}_1$ and $\hat{v}_2$, halve and concatenate them at target angles θ_1 and θ_2 to form a guidance voxel v_{guide}, which is encoded into a guidance latent via the Sparse Structure Encoder. The guidance latent is then combined with pure Gaussian noise via a weighted sum as the initial denoising latent:

$$z_{\text{init}} = \alpha \cdot \text{Encoder}(v_{\text{guide}}) + (1-\alpha) \cdot z_{\text{noise}}, \tag{5}$$

where α controls the guidance strength, balancing structural prior and generation diversity.

Space Control Guidance. Inspired by SpaceControl [17], we adapt the principle of manipulating the starting denoising timestep to our dual-semantic fusion setting. Similarly to Noise Blending Guidance, we first generate v_{guide} and encode it into a guidance latent z_{guide}, and interpolate between the guidance latent and pure noise at timestep t_0:

$$z\{t_0\} = t_0 \cdot \text{Encoder}(v_{\text{guide}}) + (1-t_0) \cdot z_{\text{noise}}. \tag{6}$$

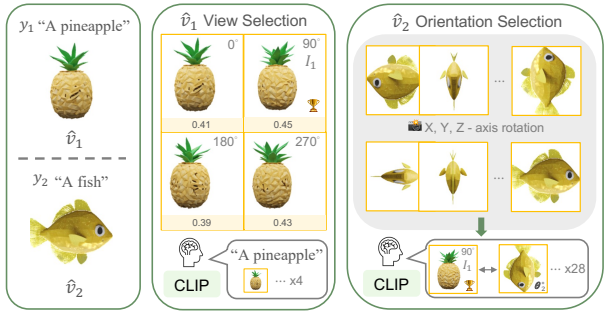

Fig. 6: CLIP-guided Orientation Search. Anchor View Selection identifies the best representative view I_1 for $\hat{v}_1$ from 4 orthogonal renders via CLIP text-image similarity (e.g., *"pineapple"* at 90°). **Cross-object Matching** then evaluates 28 3D rotations of $\hat{v}_2$, selecting the angle θ_2^* that maximizes CLIP image-image similarity to I_1. This optimally aligns their silhouettes prior to SDF blending.

A larger t_0 imposes stronger structural constraints, while a smaller t_0 preserves more generation diversity. In our 25-step setting, $t_0 = 10$ means the first 10 steps are guided before normal denoising resumes, striking a balance between structural constraint and generation freedom.

3.6 CLIP-guided Orientation Search

Existing multi-view illusion methods typically assume fixed viewpoints (e.g., 0° and 180°); however, different objects may exhibit significant orientation differences in their canonical poses. Directly fusing two geometrically misaligned objects at default angles often results in a blended mesh that fails to present clear semantic silhouettes. To address this, we propose a CLIP-based adaptive orientation search that automatically selects the optimal fusion orientation before denoising begins, as illustrated in Fig. 6.

Anchor View Selection for Object 1. We independently generate single-semantic voxels $\hat{v}_1$ and $\hat{v}_2$ from prompts y_1 and y_2 respectively. For object 1, we render 4 candidate views of $\hat{v}_1$ at 90° intervals around the Z-axis and select the view with the highest CLIP text-image similarity to y_1 as the representative view I_1.

Cross-Object Orientation Matching for Object 2. For object 2, we sample discrete rotation combinations along the (X, Y, Z) axes, generating 28 candidate renders, and select the rotation with the highest CLIP image-image similarity to I_1 as the fusion orientation θ_2^*. Candidate angles are sampled at 90° intervals, covering $\{0°, 90°, 180°, 270°\}$ per axis, balancing search efficiency and angular coverage. Denser sampling (e.g., 45°) could further improve alignment precision, which we leave as future work.

3.7 Extension to Three-Object 3D Illusions

Our framework naturally scales to three-object illusions by adding a third Stage 1 denoising branch. Given three prompts and a shared noise z_t, we fix target

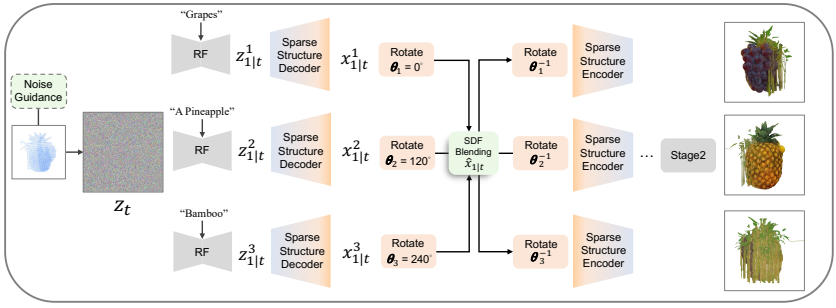

Fig. 7: Extension to three-object 3D illusion. Our framework scales to three semantics (e.g., *"Grapes"*, *"A Pineapple"*, *"Bamboo"*) by adding a third denoising branch, with rotation angles fixed at 0°, 120°, 240° to uniformly cover 360°. Noise Guidance steers all three branches toward their respective targets, ensuring each semantic is clearly presented at its target viewpoint.

angles at 0°, 120°, and 240°. At each step, the three predicted voxels are SDF-fused, inverse-rotated, and re-encoded. Due to complex geometric conflicts, Noise Guidance is mandatory here, utilizing a merged prior of three pre-generated single-semantic objects. Specifically, since averaging three geometries imposes stronger geometric conflicts than the two-object case, we adopt Space Control Guidance with an increased $t_0 = 20$ out of 25 steps, providing stronger structural constraints to ensure each semantic is preserved at its target viewpoint. Stage 2 remains identical to the two-object case. See Fig. 7 for the pipeline and Sec. 4.4 for results.

4 Experiments

4.1 Experimental Setup

Baselines. We compare against four baselines: **Shape from Semantics** [40], which addresses the same task via Score Distillation Sampling (SDS) but requires ∼40 minutes per object; and **Direct Concatenation**, which independently generates two objects via TRELLIS [73], halves each along the midplane, and stitches them into a single mesh. We also evaluate **TRELLIS** [73] and **DreamBeast** [42], both prompted with *"a single 3D object, front side is {y_1}, back side is {y_2}"* to elicit view-dependent semantics from a single generation. TRELLIS is a state-of-the-art feed-forward 3D generation model, while DreamBeast is an SDS-based method designed for generating fantastical creatures with part-level semantic control. These two baselines assess whether existing generation methods can produce coherent visual illusions without task-specific design.

Data. We collect object prompts from Shape from Semantics [40] and supplement with additional common objects, yielding 60 distinct objects across five categories: 16 birds, 19 mammals, 5 reptiles and aquatic animals, 9 plants, and 11 man-made artifacts. Two objects are randomly sampled per experiment to form a prompt pair.

Table 1: Runtime breakdown per stage and case on a single NVIDIA RTX 4090.

	Stage 1	Stage 2	Total
Case 1 & 2	∼1 min	∼2 min	∼3 min
Case 3 (w/ CLIP search)	∼3 min	∼2 min	∼5 min

Implementation Details. All experiments are conducted on a single NVIDIA RTX 4090 GPU. For SDF blending, we use Truncated SDFs with truncation distance clip_s = 12 and binarization threshold $\tau = 0.8$. Stage 1 runs for 25 denoising steps with one geometry blending operation per step, with CFG applied at guidance scale $\omega = 7.5$ within $t \in [0.5, 0.95]$ (Interval CFG); Stage 2 runs for 30 denoising steps. For CLIP-guided Orientation Search, we use OpenCLIP [29] ViT-B/32 pretrained on LAION-2B (`laion2b_s34b_b79k`). For Noise Blending Guidance, we set $\alpha = 0.3$; for Space Control Guidance, we set $t_0 = 10$ out of 25 steps. We organize our experiments into three cases based on the rotational configuration of the two objects; in all cases, object A is fixed at its canonical orientation:

- **Case 1:** Object B is also unrotated. The resulting illusion reveals the front face of object A at 0° and the back face of object B at 180°.
- **Case 2:** Object B is rotated by 180°, so that both the front of object A and the front of object B are visible at 0° and 180°, respectively.
- **Case 3:** The rotation angle of object B is automatically determined by CLIP-guided Orientation Search. Empirically, the selected angle tends to place the two objects approximately 180° apart in most cases.

Noise Blending and Space Control Guidance can be freely combined with Cases 1 and 2; Case 3 requires no additional guidance since the fusion angle is already adapted via CLIP. Table 1 provides a per-stage runtime breakdown: for Cases 1 and 2, Stage 1 takes ∼1 minute and Stage 2 takes ∼2 minutes, totaling ∼3 minutes. For Case 3, Stage 1 requires ∼3 minutes due to the additional CLIP-guided Orientation Search, while Stage 2 again takes ∼2 minutes, totaling ∼5 minutes. All configurations complete within 3–5 minutes, offering a significant efficiency advantage over the SDS-based baseline (∼40 minutes).

Metrics. We design six quantitative metrics and a user study to evaluate generation quality and illusion effect: *(1) CLIP Similarity.* We compute CLIP [29,61] text-image similarity across 1,000 renders per viewpoint (±20° jitter). Notably, Direct Concatenation artificially inflates this score by trivially preserving single-view appearances via naive stitching, making CLIP alone insufficient. *(2) GPT Accuracy (%).* To evaluate semantic clarity, we prompt GPT-4.1-mini to identify the respective semantics from side-by-side renders of both target viewpoints. Specifically, given a side-by-side image composed of two viewpoint renders (labeled A and B), we query the model with: *"Given the image with pieces A and*

B, choose which interpretation fits: Option 1: Left = $\{y_1\}$, Right = $\{y_2\}$; Option 2: Left = $\{y_2\}$, Right = $\{y_1\}$. Respond with Option 1 or Option 2 only." Accuracy is computed as the proportion of responses matching the ground-truth prompt assignment. This 2-way accuracy quantifies how well the object conveys its intended semantics.

(3) FID & KID. To measure visual realism, we compute FID [22] and KID [2] between 1,000 renders of our results and 1,000 reference images (20 views of 50 objects) from Objaverse 1.0 [13]. *(4) Object Detection Score.* To objectively measure geometric fusion, we render the midpoint angle and count objects using OWLv2 [55] (Fig. 8). We report the average object count (ideally 1) and the multi-object rate (proportion of renders with > 1 detection). *(5) View-Conditional CLIP Contrast.* To penalize cross-view semantic leakage, we compute CLIP similarity between each viewpoint render and the *opposite* prompt—i.e., the prompt that is *not* intended to be visible from that viewpoint. A lower score indicates that the unintended semantics are less recognizable from each view, reflecting a cleaner perceptual separation between the two target interpretations. This metric directly addresses the limitation of standard CLIP Similarity, which can be inflated by naive stitching methods that allow semantic bleed-through. *(6) Boundary Seam Score (Impact Factor).* To quantify geometric smoothness at the fusion boundary, we measure surface curvature discontinuity around the seam region. For each boundary vertex v, we compute the *Curvature Jump* as the mean absolute difference between v's curvature and that of its immediate neighbors. We then define *Boundary Avg* as the mean Curvature Jump over all boundary vertices, and *Global Avg* as the mean Curvature Jump over all vertices in the mesh. The *Impact Factor* is defined as:

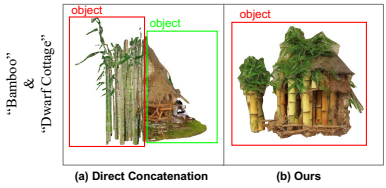

Fig. 8: Object detection at the junction viewpoint for *"Bamboo"* & *"Dwarf Cottage"*. (a) Direct Concatenation is detected as two separate objects (red and green boxes); (b) Ours is detected as a single unified object (red box only).

$$\text{Impact Factor} = \frac{\text{Boundary Avg}}{\text{Global Avg}}. \tag{7}$$

A value close to 1 indicates that the boundary region is geometrically indistinguishable from the rest of the surface, whereas larger values indicate a visually abrupt seam. Direct Concatenation produces high Impact Factor scores due to the sharp geometric discontinuity at the stitching boundary.

User Study. Beyond quantitative metrics, we conduct a user study to collect perceptual evaluations from human observers. A total of 50 participants took part in the study. Each participant was presented with rendered results from all three methods and asked to evaluate them across the following questions:

- **Q1:** How recognizable are the intended semantics at each target viewpoint? (1: unrecognizable, 2: partially recognizable, 3: clearly recognizable)

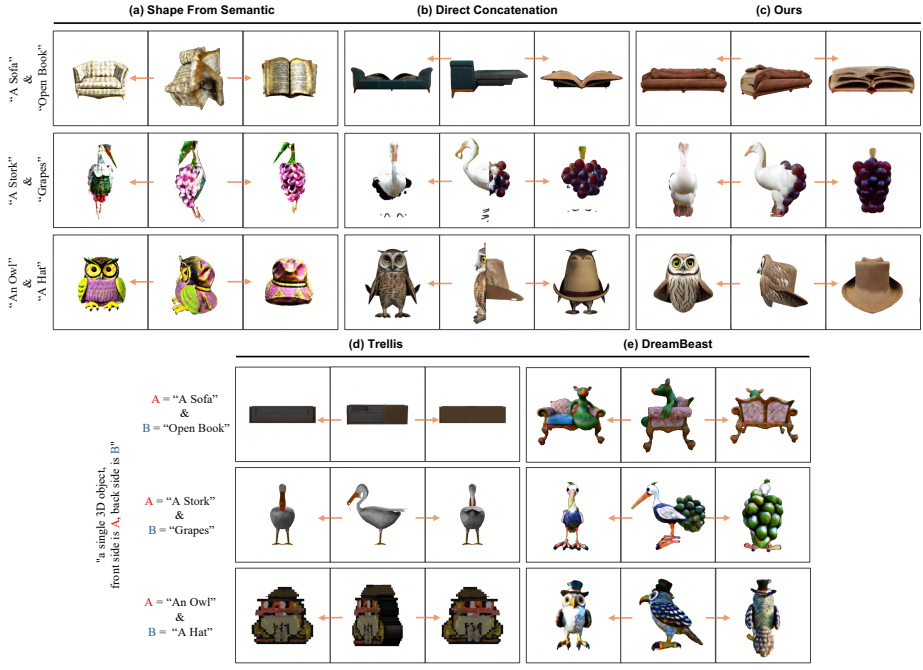

Fig. 9: Qualitative comparison with baselines. Left-to-right: View 1, blended mesh, View 2. **(a)** Shape From Semantics [40] suffers from over-saturation and geometry leakage (e.g., *"Stork"/"Grapes"*). **(b)** Direct Concatenation exposes unnatural junction seams and the opposing geometry at target views. **(c)** Ours produces a single coherent mesh with clear, view-dependent semantics and no leakage. **(d)** TRELLIS [73], prompted with *"a single 3D object, front side is A, back side is B"*, fails to produce view-dependent semantics and collapses to a generic shape that reflects neither target interpretation. **(e)** DreamBeast [42], despite its part-level semantic control, generates fantastical hybrid creatures that blend both semantics into every viewpoint rather than isolating them, failing to achieve the intended illusion effect.

- **Q2:** Which result better aligns with the intended semantics?
- **Q3:** Comparing CLIP-adaptive orientation (Case 3) versus fixed 0°/180° angles, which produces a more natural illusion effect?

4.2 Results and Analysis

Quantitative Comparison. Table 2 reports results over 50 randomly sampled prompt pairs. Our method outperforms or matches both baselines on the majority of metrics. Direct Concatenation achieves the highest CLIP score (29.030 vs. ours 28.170), but as discussed in Sec. 4.1, this is a systematic artifact of naively preserving per-viewpoint appearance rather than genuine illusion quality. Our method achieves the highest GPT Accuracy (84%), outperforming Direct Concatenation (76%) and Shape from Semantics (70%), lowest FID (185.555), and best Object Detection scores (avg. count 0.86, multi-object rate 18% vs. 2.1 and 56% for Direct Concatenation), collectively demonstrating superior semantic

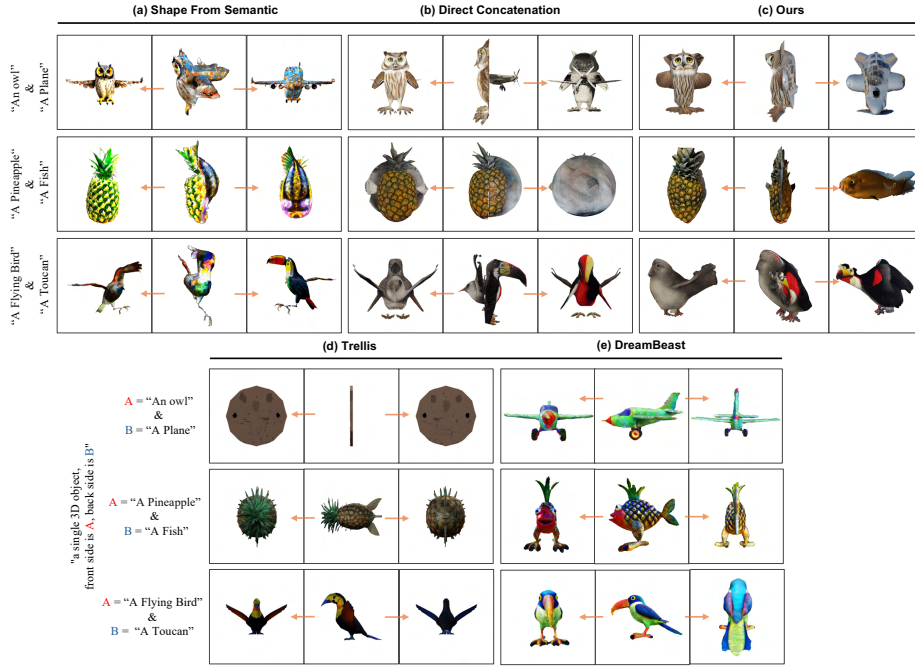

Fig. 10: Qualitative comparison (CLIP-guided orientation). Left-to-right: View 1, blended mesh, View 2. **(c) Ours** uses CLIP-guided search to adaptively align compatible silhouettes (e.g., sideways fish with pineapple, flying owl with plane), revealing clearer semantics at both viewpoints. Lacking adaptive rotation, **(a)** Shape From Semantics [40] and **(b)** Direct Concatenation suffer from silhouette misalignment and degraded fusion. **(d)** TRELLIS [73], prompted with *"a single 3D object, front side is A, back side is B"*, produces flat, degenerate geometry that fails to capture either semantic. **(e)** DreamBeast [42] blends both semantics uniformly across all viewpoints, generating fantastical hybrid creatures rather than achieving view-dependent separation.

recognizability, visual realism, and geometric coherence. Our method completes within 3–5 minutes, matching Direct Concatenation and offering a significant efficiency advantage over Shape from Semantics [40] (∼40 minutes).

Qualitative Comparison. Fig. 9 and Fig. 10 compare results under fixed-angle and CLIP-guided configurations respectively. Our method produces geometrically coherent results with clear per-viewpoint semantics, while Shape From Semantics [40] suffers from over-saturation and Direct Concatenation exposes visible junction seams.

User Studies. Our user studies strongly reinforce the quantitative findings. Participants overwhelmingly preferred our method over both baselines and rated our results as semantically recognizable, confirming the perceptual quality of our generated illusions. Additionally, the large majority found CLIP-guided orien-

Table 2: Comparison with the baselines.

Method	CLIP ↑	CLIP (opp.) ↓	GPT Acc. (%) ↑	FID ↓	KID ↓	Object Detection		Impact Factor	Runtime
						Avg. Obj. Count	Multi-Obj. Rate (%)		
Shape From Semantic [40]	27.460	19.72	70	194.136	0.051	0.64	2	0.973	∼40 min
Direct Concat	**29.030**	20.38	76	187.886	0.067	2.1	56	1.129	∼3-5 min
Trellis	22.180	22.68	60	**174.130**	**0.044**	0.58	8	0.952	∼2-3 min
DreamBeast	22.993	22.89	65	184.956	0.0935	0.76	18	0.974	∼3-4 hr
Ours	28.170	**19.26**	**84**	185.555	0.051	0.86	18	0.994	∼3-5 min

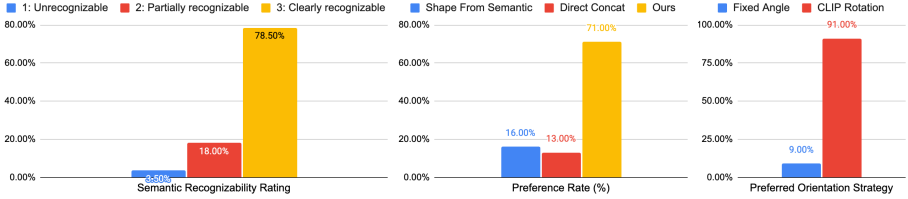

Fig. 11: User study. (Left) Semantic Recognizability: 78.5% of participants rated our results as clearly recognizable (score 3). **(Middle) Method Preference**: 71% of participants preferred our method over Shape From Semantics [40] (16%) and Direct Concatenation (13%). **(Right) Orientation Strategy**: 91% of participants found CLIP-guided orientation to produce a more natural illusion than fixed 0°/180° angles.

tation to produce more natural illusions than fixed angles, validating the effectiveness of our adaptive orientation search. Results are shown in Fig. 11.

4.3 Ablation Studies

Geometry Blending Strategy. We compare five voxel fusion strategies. **Union** takes the logical OR of two voxel grids, directly superimposing both structures without any shape coordination. **Blur Average** applies 3D Gaussian smoothing before element-wise averaging, introducing soft boundaries but losing fine geometric details. **Minkowski Blend** dilates each voxel with a spherical structuring element before merging, which tends to inflate the overall mesh volume. **Polar Coord Blend** fuses voxels slice-by-slice via 2D polar-coordinate boundary averaging, implicitly assuming star-shaped slices and thus failing for concave or non-symmetric objects. **SDF Average (Ours)** averages the Truncated SDFs of both occupancy grids and binarizes the result; the zero iso-surface naturally corresponds to the intermediate shape between the two objects, producing a geometrically stable and coherent blended surface. Visual comparisons are shown in Fig. 12.

Noise Guidance. We compare three configurations: no guidance, Noise Blending Guidance, and Space Control Guidance. The optimal strategy depends on the geometric characteristics of the object pair: Space Control imposes stronger constraints via the guided latent state for the first t_0 steps, making it better suited for pairs with large silhouette discrepancies; Noise Blending provides a milder prior that benefits pairs with similar silhouettes but distinct semantics; for geometrically compatible pairs, no guidance suffices. We therefore treat noise guidance as an optional, pair-dependent auxiliary rather than a mandatory component. Qualitative comparisons are shown in Fig. 13.

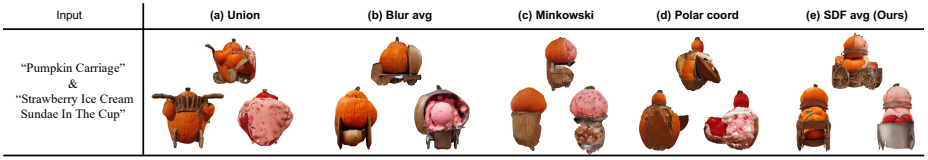

Fig. 12: Ablation on geometry blending (*"carriage"/"sundae"*). Left-to-right: View 1, blended mesh, View 2. **(a) Union:** yields conflicting junctions. **(b) Blur Avg:** loses fine details. **(c) Minkowski:** over-expands geometry. **(d) Polar Coord:** distorts asymmetric objects. **(e) SDF Avg (Ours):** optimally balances geometric integrity and semantic preservation.

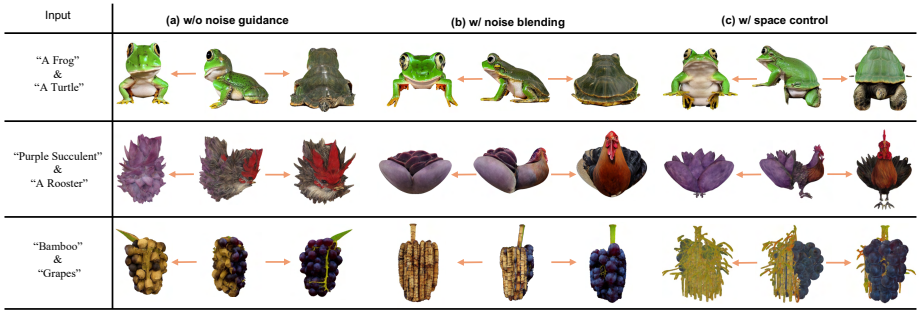

Fig. 13: Ablation on noise guidance across varying geometric compatibilities: (a) no guidance, (b) Noise Blending, and (c) Space Control. **(c)** handles large geometric discrepancies (*"Bamboo"/"Grapes"*) best via strong spatial constraints. **(b)** is optimal for similar silhouettes but distinct semantics (*"Succulent"/"Rooster"*), avoiding semantic loss in (a) and residual artifacts in (c). For compatible pairs (*"Frog"/"Turtle"*), all settings perform comparably. Thus, optimal intervention strength depends on geometric alignment.

View-Conditioned Texture Synthesis. We ablate the necessity of Stage 2 by comparing results with and without view-conditioned texturing (Fig. 14(a)(b)). Without Stage 2, TRELLIS fails to interpret the unnatural fused geometry, producing semantically incoherent textures at both viewpoints. Our view-conditioned synthesis assigns each viewpoint its own texture prediction, ensuring the rendered appearance closely matches the corresponding prompt at each target angle.

CLIP-Guided Orientation Search. We compare adaptive CLIP-guided rotation against fixed $0°/180°$ angles (Fig. 14(c),(d)). For canonically misaligned pairs (e.g., horizontal *"Rhinoceros"*/upright *"Pineapple"*), fixed angles cause silhouette misalignment; CLIP rotation optimally aligns them to recover distinct semantics. It similarly prevents semantic collapse for *"Succulent"/"Cactus"*. However, for already compatible pairs, CLIP misclassifications can yield suboptimal angles. Thus, we recommend adaptive rotation specifically for geometrically challenging pairs, defaulting to fixed angles for compatible ones.

4.4 Applications

Our framework demonstrates strong performance across diverse semantic combinations, from structurally similar to geometrically distinct object pairs, as shown

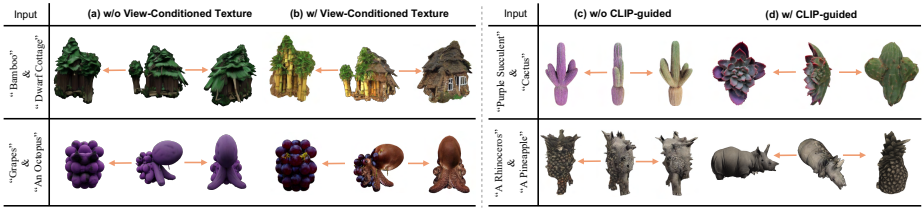

Fig. 14: Ablation on texture synthesis and orientation search. Each pair shows View 1, blended mesh, and View 2. **(a) w/o View-Conditioned Texture:** Standard texturing fails on fused geometries, blending semantics into indistinguishable appearances (e.g., uniform foliage or solid purple). **(b) w/ View-Conditioned Texture (Ours):** Accurately assigns distinct, prompt-driven textures to each viewpoint. **(c) w/o CLIP-guided:** Fixed 0°/180° angles cause silhouette misalignments, failing to reveal clear semantics. **(d) w/ CLIP-guided (Ours):** Adaptive rotation optimally aligns silhouettes, recovering distinct semantics at both views.

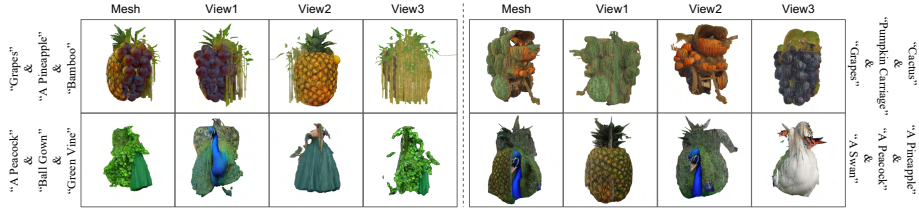

Fig. 15: Qualitative results of three-object 3D illusion generation. Each row shows the fused mesh and three target-viewpoint renders for a prompt triplet. Our method successfully generates a single coherent mesh that presents three distinct semantics at 0°, 120°, and 240°, respectively.

in Fig. 9 and Fig. 10. Beyond the two-object case, our method scales naturally to three-object illusion generation without architectural modifications (Sec. 3.7), as demonstrated in Fig. 15.

5 Conclusion

We present a zero-shot framework that extends visual illusions to true 3D geometry. Given text prompts, our method generates a single coherent mesh revealing distinct semantics from different viewpoints in under 5 minutes, without per-shape optimization. Furthermore, we introduce CLIP-guided Orientation Search for silhouette alignment and Noise Guidance to resolve geometric conflicts, demonstrating seamless scalability to three-object illusions without modifying the core fusion procedure.

Limitations. Our method inherits TRELLIS's failure cases for specific object categories (e.g., pigs, bats; see supp.). Furthermore, CLIP-guided Orientation Search struggles with three-object illusions because three silhouettes average into ambiguous shapes. Thus, we currently fix angles to 0°/120°/240°, leaving automated three-object alignment as future work.

Acknowledgements

This research was funded by the National Science and Technology Council, Taiwan, under Grants NSTC 112-2222-E-A49-004-MY2 and 113-2628-EA49-023-. The authors are grateful to Google, NVIDIA, and MediaTek Inc. for their generous donations. Yu-Lun Liu acknowledges the Yushan Young Fellow Program by the MOE in Taiwan.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
3. Burgert, R., Li, X., Leite, A., Ranasinghe, K., Ryoo, M.: Diffusion illusions: Hiding images in plain sight. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
4. Cao, T., Kreis, K., Fidler, S., Sharp, N., Yin, K.: Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4169–4181 (2023)
5. Chang, P., Sancho, S., Tang, J., Gross, M., Azevedo, V.: Lookingglass: Generative anamorphoses via laplacian pyramid warping. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24–33 (2025)
6. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 18558–18568 (2023)
7. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22246–22256 (2023)
8. Chen, T.H., Chen, Y.H., Tu, T., Lee, J.Y., Wu, C.Y., Lin, F., Zhang, H., Paz, D., Huang, X., Guo, Y., et al.: Pantheon360: Taming digital twin generation via 3d-aware 360deg video diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11138–11149 (2026)
9. Chen, Z., Geng, D., Owens, A.: Images that sound: Composing images and sounds on a single canvas. Advances in Neural Information Processing Systems **37**, 85045–85073 (2024)
10. Cheng, H.H., Zhang, S.L., Liu, Y.L.: Stroke of surprise: Progressive semantic illusions in vector sketching. arXiv preprint arXiv:2602.12280 (2026)
11. Cheng, W., Mu, J., Zeng, X., Chen, X., Pang, A., Zhang, C., Wang, Z., Fu, B., Yu, G., Liu, Z., et al.: Mvpaint: Synchronized multi-view diffusion for painting anything 3d. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 585–594 (2025)
12. Debnath, S., Tiwari, A., Sadekar, K., Raman, S.: Rasp: revisiting 3d anamorphic art for shadow-guided packing of irregular objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5849–5858 (2025)
13. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)

14. Deng, K., Omernick, T., Weiss, A., Ramanan, D., Zhu, J.Y., Zhou, T., Agrawala, M.: Flashtex: Fast relightable mesh texturing with lightcontrolnet. In: European conference on computer vision. pp. 90–107. Springer (2024)
15. Dodik, A., Yu, I., Chandra, K., Ragan-Kelley, J., Tenenbaum, J., Sitzmann, V., Solomon, J.: Meschers: Geometry processing of impossible objects. ACM Transactions on Graphics (TOG) **44**(4), 1–10 (2025)
16. Du, Y., Durkan, C., Strudel, R., Tenenbaum, J.B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., Grathwohl, W.S.: Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In: International conference on machine learning. pp. 8489–8510. PMLR (2023)
17. Fedele, E., Engelmann, F., Huang, I., Litany, O., Pollefeys, M., Guibas, L.: Spacecontrol: Introducing test-time spatial control to 3d generative modeling. arXiv preprint arXiv:2512.05343 (2025)
18. Feng, Y., Sanjay, V., Lutz, S., AlBahar, B., Ge, S., Huang, J.B.: Illusion3d: 3d multiview illusion with 2d diffusion priors. arXiv preprint arXiv:2412.09625 (2024)
19. Gao, X., Yang, S., Liu, J.: Ptdiffusion: Free lunch for generating optical illusion hidden pictures with phase-transferred diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18240–18249 (2025)
20. Geng, D., Park, I., Owens, A.: Factorized diffusion: Perceptual illusions by noise decomposition. In: European Conference on Computer Vision. pp. 366–384. Springer (2024)
21. Geng, D., Park, I., Owens, A.: Visual anagrams: Generating multi-view optical illusions with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24154–24163 (2024)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
24. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
25. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
26. Hsiao, K.W., Huang, J.B., Chu, H.K.: Multi-view wire art. ACM Trans. Graph. **37**(6), 242 (2018)
27. Huang, Y.C., Chan, J., Chien, H.J., Liu, Y.L.: Voxify3d: Pixel art meets volumetric rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15398–15410 (2026)
28. Huang, Y.C., Chien, H.J., Lin, C.Y., Chen, Y.H., Liu, Y.L.: Gamo: Geometry-aware multi-view diffusion outpainting for sparse-view 3d reconstruction. arXiv preprint arXiv:2512.25073 (2025)
29. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`, if you use this software, please cite it as below.
30. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 867–876 (2022)

31. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5885–5894 (2021)
32. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
33. Ke, B.H., Xie, Y.Z., Liu, Y.L., Chiu, W.C.: Stealthattack: Robust 3d gaussian splatting poisoning via density-guided illusions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27400–27411 (2025)
34. Kim, J., Koo, J., Yeo, K., Sung, M.: Synctweedies: A general generative framework based on synchronized diffusions. arXiv preprint arXiv:2403.14370 (2024)
35. Kim, S., Lee, K., Choi, J.S., Jeong, J., Sohn, K., Shin, J.: Collaborative score distillation for consistent visual synthesis. arXiv preprint arXiv:2307.04787 (2023)
36. Lan, Y., Zhou, S., Lyu, Z., Hong, F., Yang, S., Dai, B., Pan, X., Loy, C.C.: Gaussiananything: Interactive point cloud latent diffusion for 3d generation (2025)
37. Lee, J.Y., Liu, Y.R., Tsai, S.R., Chang, W.C., Wu, C.H., Chan, J., Zhao, Z., Lin, C.H., Liu, Y.L.: Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. arXiv preprint arXiv:2510.15869 (2025)
38. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. Advances in Neural Information Processing Systems **36**, 50648–50660 (2023)
39. Li, L., Wang, C., Zhou, Y., Deng, B., Zhang, J.: Shape from semantics: 3d shape generation from multi-view semantics. arXiv preprint arXiv:2502.00360 (2025)
40. Li, L., Wang, C., Zhou, Y., Deng, B., Zhang, J.: Shape from semantics: 3d shape generation from multi-view semantics (2025), `https://arxiv.org/abs/2502.00360`
41. Li, M.F., Ku, Y.F., Yen, H.X., Liu, C., Liu, Y.L., Chen, A.Y., Kuo, C.H., Sun, M.: Genrc: Generative 3d room completion from sparse image collections. In: European Conference on Computer Vision. pp. 146–163. Springer (2024)
42. Li, R., Han, J., Melas-Kyriazi, L., Sun, C., An, Z., Gui, Z., Sun, S., Torr, P., Jakab, T.: Dreambeast: Distilling 3d fantastical animals with part-aware knowledge transfer. In: 2025 International Conference on 3D Vision (3DV). pp. 1243–1252. IEEE (2025)
43. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., et al.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
44. Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y.: Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6517–6526 (2024)
45. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023)
46. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
47. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European conference on computer vision. pp. 423–439. Springer (2022)

48. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9298–9309 (2023)
49. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
50. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
51. Liu, Y., Xie, M., Liu, H., Wong, T.T.: Text-guided texturing by synchronized multi-view diffusion. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
52. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024)
53. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12663–12673 (2023)
54. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13492–13502 (2022)
55. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36**, 72983–73007 (2023)
56. Mitra, N.J., Pauly, M.: Shadow art. ACM Transactions on Graphics **28**(5), 156–1 (2009)
57. Oliva, A., Torralba, A., Schyns, P.G.: Hybrid images. ACM Transactions on Graphics (TOG) **25**(3), 527–532 (2006)
58. Perroni-Scharf, M., Rusinkiewicz, S.: Constructing printable surfaces with view-dependent appearance. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–10 (2023)
59. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
60. Qu, Z., Yang, L., Zhang, H., Xiang, T., Pang, K., Song, Y.Z.: Wired perspectives: Multi-view wire art embraces generative ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6149–6158 (2024)
61. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
62. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
63. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
64. Sadekar, K., Tiwari, A., Raman, S.: Shadow art revisited: a differentiable rendering based approach. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 29–37 (2022)
65. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)

66. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
67. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
68. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3835–3844 (2022)
69. Wang, C., Deng, B., Zhang, J.: Neural shadow art. arXiv preprint arXiv:2411.19161 (2024)
70. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in neural information processing systems **36**, 8406–8441 (2023)
71. Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., Zhu, J.: Crm: Single image to 3d textured mesh with convolutional reconstruction model. In: European conference on computer vision. pp. 57–74. Springer (2024)
72. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., Yao, Y.: Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. Advances in Neural Information Processing Systems **37**, 121859–121881 (2024)
73. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)
74. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)
75. Yang, X., Chen, C., Yang, X., Liu, F., Lin, G.: Text-to-image rectified flow as plug-and-play priors. arXiv preprint arXiv:2406.03293 (2024)
76. Yeh, Y.Y., Huang, J.B., Kim, C., Xiao, L., Nguyen-Phuoc, T., Khan, N., Zhang, C., Chandraker, M., Marshall, C.S., Dong, Z., et al.: Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4304–4314 (2024)
77. Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6796–6807 (2024)
78. Yu, X., Yuan, Z., Guo, Y.C., Liu, Y.T., Liu, J., Li, Y., Cao, Y.P., Liang, D., Qi, X.: Texgen: a generative diffusion model for mesh textures. ACM Transactions on Graphics (TOG) **43**(6), 1–14 (2024)
79. Zeng, X., Chen, X., Qi, Z., Liu, W., Zhao, Z., Wang, Z., Fu, B., Liu, Y., Yu, G.: Paint3d: Paint anything 3d with lighting-less texture diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4252–4262 (2024)
80. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
81. Zhao, Z., Liu, W., Chen, X., Zeng, X., Wang, R., Cheng, P., Fu, B., Chen, T., Yu, G., Gao, S.: Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. Advances in neural information processing systems **36**, 73969–73982 (2023)