# CAST3D: Customizing Arbitrary 2D Assets into 3D World

Anonymous ECCV 2026 Submission

Paper ID #8909


**Fig. 1:** High-quality compositional 3D assets generated by CAST3D. **Customized Composition**: Our method enables transforming arbitrary 2D assets into semantically coherent and geometrically consistent 3D objects under textual guidance, *e.g.* putting on diverse caps or backpacks for a boy. **Customized Editing**: Our method also allow components in the 2D assets, such as wings of a dragon, to be replaced by those from another asset while preserving overall consistency.

**Abstract.** High-quality 2D assets have become increasingly abundant and easy to edit, providing a rich foundation for creative content across art, design, and virtual environments. However, while diffusion-based 3D generation has achieved remarkable progress in single-object synthesis, leveraging such 2D assets for controllable 3D composition remains a challenging problem. To address this, we introduce CAST3D, a training-free framework that enables Customized Composition in 3D: transforming arbitrary 2D assets into a coherent 3D object or scene under textual guidance. CAST3D consists of two stages: 3D Layout Hinting and Compositional Generation. To maintain structural consistency and eliminate artifacts, we further design stochastic trajectory manipulation (STM) for structure-preserving modification and a connectivity-based pruning strategy for clean geometry integration. Extensive experiments demonstrate that CAST3D produces semantically consistent and visually faithful 3D compositions, bridging 2D asset creation and 3D world synthesis.

**Keywords:** 3D Generation · Training Free · Diffusion Models

# 1 Introduction

Recent advances in visual content generation have led to the proliferation of high-quality 2D assets across domains such as digital art, design, and virtual environments [8,21,22]. Their accessibility and high editability provide a rich foundation of visual knowledge that can support 3D content creation. However, compared with the efficiency of 2D generation, constructing 3D objects and scenes still requires complex structural representations [15, 28, 32] and geometric optimization [11, 44]. Effectively leveraging existing 2D assets for cross-dimensional 3D composition and customization remains a promising yet underexplored direction in generative modeling.

To address this gap in 3D generative modeling, we introduce a new task termed **Customized 3D Composition**, which aims to transform arbitrary 2D assets into semantically coherent and geometrically consistent 3D object or scene under textual guidance. For example, given images of a boy, a baseball cap, and a backpack, and the prompt "a boy wearing a *baseball cap* and a *backpack*," the goal is to produce a geometrically plausible and visually consistent 3D composition. This task is inherently challenging because the provided 2D assets may exhibit mismatched viewpoints, inconsistent scales, and missing occlusion information, making the transformation from discrete 2D observations to a coherent 3D structure non-trivial. We view this task as bridging multi-modal generative modeling and controllable 3D synthesis, extending diffusion-based 3D generation toward compositional reasoning.

Although diffusion-based 3D generation methods such as Trellis [59] and Hunyuan3D [50] have achieved remarkable progress in single-object synthesis, they are primarily optimized for isolated objects and lack controllable compositionality across assets. Also, existing approaches typically rely on solely textual or single-image guidance, limiting their ability to integrate personalized visual priors from heterogeneous 2D sources. As a result, current diffusion-based 3D models remain insufficient for capturing cross-asset semantic relationships and ensuring geometric consistency in multi-object composition.

Intuitively, customized 3D composition from 2D assets can be approached through two straightforward paradigms.

1) Uplift-then-Compose: Lift each 2D asset into 3D using image-to-3D methods such as Score Distillation Sampling [39], followed by spatial arrangement based on textual descriptions. However, independently generated assets often have inconsistent scales, orientations, or coordinate frames, leading to spatial misalignment and geometric overlap.
2) Customize-then-Generate: First synthesize a compositional 2D image using multi-concept diffusion models or VLMs [4, 34, 55, 57] and then generate the corresponding 3D object as a whole. This strategy relies heavily on the 2D generator's multi-concept consistency capability, and is likely to suffer from geometric distortions and inconsistencies due to modality conversion.

Together, these limitations highlight the need for a paradigm that jointly enables semantic reasoning and geometric composition, which forms our core.

To overcome these challenges, we propose CAST3D, a training-free framework for compositional 3D generation from arbitrary 2D assets. As shown in Fig. 1, CAST3D leverages the semantic and geometric priors of 3D latent diffusion models and constructs 3D compositions through two stages: 3D layout hinting and compositional generation. This progressive workflow preserves the structural priors encoded in diffusion models while allowing independent image-3D alignment for each asset, achieving both global semantic coherence and local geometric fidelity.

Aimed at further improving stability during structural modification and generation, CAST3D introduces Stochastic Trajectory Manipulation, dubbed STM, as its core innovation. STM adjusts the denoising trajectory in latent space to balance semantic modification and geometric stability, enabling concept-specific edits while avoiding the structural drift caused by strong conditioning. Additionally, a lightweight connectivity-based pruning step removes disconnected artifacts during composition, further enhancing visual coherence.

Our main contributions are summarized as follows:

- We introduce the Customized 3D Composition task and present CAST3D, a training-free progressive framework that bridges 2D asset utilization and controllable 3D generation;
- We design Stochastic Trajectory Manipulation (STM) as a novel mechanism for balancing semantic modification and geometric preservation, and adopt a connectivity-based pruning strategy for improved composition quality;
- We provide extensive experiments demonstrating that CAST3D produces semantically coherent, geometrically consistent, and visually faithful 3D compositions across diverse assets and scenes.

## 2 Related Works

### 2.1 Customized Image Composition

Despite the remarkable progress of diffusion-based models, synthesizing coherent images that compose multiple distinct concepts remains highly challenging, even for purely textual concepts [3,7,37]. This difficulty is further amplified in image-conditioned settings, *i.e.*, customized image composition, where models must align and integrate several visual inputs in a coherent manner.

CustomDiffusion [20] pioneers multi-concept customization through test-time finetuning and KV-merging to fuse multiple concepts. MS-Diffusion [55] introduces a grounding resampler as a dedicated image encoder together with explicit layout guidance, enabling effective cross-attention modulation and mitigating concept mixing. In parallel, several methods [18,38,60] utilize LoRA [14] modules to encode distinct concepts, applying multiple LoRAs at inference to inject multiple identities or attributes. Another line of work employs encoder-based conditioning, enriching visual guidance via refined attention mechanisms [12,36] or multi-stage training pipelines [34,58]. Related efforts further explore modulation space, learning disentangled concept representations to improve compositional control [4,9], or integrate with latest VLMs [57].

### 2.2 Personalized 3D Generation

As image generation techniques continue to advance, 3D generation has also witnessed substantial progress. Score-based diffusion distillation [39,53] bridges the gap between 2D and 3D, providing a viable pathway for personalized 3D synthesis using powerful 2D priors. DreamBooth3D [41] and Consist3D [35] further pushes this direction by introducing a partial-to-full training strategy that progressively incorporates concept-specific visual cues into multi-view supervision. TIP-Editor [63] performs test-time finetuning of diffusion models, enabling concept embedding with user-provided 3D masks that designate regions of interest. Subsequent works [46,56] leverage richer 3D priors through multi-view diffusion models [45] or 3D latent diffusion models [5,24,59,62] to further improve fidelity and controllability.

Although these methods produce impressive results, they remain limited in composing multiple 2D assets with distinct concepts, primarily due to the lack of spatial reasoning, especially in identifying appropriate layout hints for coherent spatial arrangement. To the best of our knowledge, no existing approach provides a training-free, multi-asset, and geometrically consistent solution for customized 3D composition, which motivates the development of CAST3D.

## 3 Preliminaries

We adopt Trellis [59] as our underlying 3D latent diffusion model. Trellis represents 3D content in a unified latent space called Structured Latents (SLAT), which bridges multiple 3D representations. A SLAT is a set of active voxels, each associated with a latent feature, formally $\boldsymbol{z} = (\boldsymbol{f}_i, \boldsymbol{p}_i)$, where voxel coordinates $\boldsymbol{p}_i$ capture the coarse spatial structure and latent features $\boldsymbol{f}_i$ encode fine-grained geometry and appearance.

Trellis' generation pipeline follows two stages with two Rectfied Flow Transformers. In the sparse structure stage, a dense latent field is first predicted and decoded into a $64^3$ occupancy grid. In the SLAT stage, noised SLATs are initialized from the occupancy grid and progressively denoised, yielding detailed geometry and appearance.

## 4 Methods

### 4.1 Problem Formulation

We aim to address the problem of Customized 3D Composition, where arbitrary 2D assets are transformed and combined into a coherent 3D asset under textual guidance, with each component in the resulting asset following the appearance of its 2D counterpart.

Given a source text prompt $\boldsymbol{c}$ and 2D asset prompts $\{(\boldsymbol{o}_i, \boldsymbol{I}_i)\}_{i=1}^N$ consisting of $N$ text-image pairs assigning certain terms $\boldsymbol{o}_i$ in text prompt $\boldsymbol{c}$ with asset $\boldsymbol{I}_i$. To reduce layout ambiguity and ease generation, a "main asset" or anchor is selected, denoted as $(\boldsymbol{o}_B, \boldsymbol{I}_B)$. More formally, the goal is to design a 3D generation pipeline $\mathcal{P}(\boldsymbol{c}, \{(\boldsymbol{o}_i, \boldsymbol{I}_i)\}_{i=1}^N, i_B)$, where $i_B$ denotes the index of the main asset.

### 4.2 Stochastic Trajectory Manipulation

To enable structure-preserving latent modifications while supporting semantically plausible modifications in Rectified Flow models, we introduce stochasticity to its sampling process and exploit it with a generic trajectory manipulation algorithm.

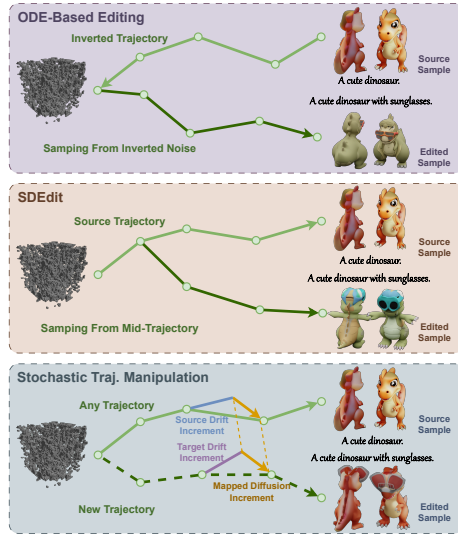


**Fig. 2:** Comparison of STM to similar algorithms. By mapping the diffusion increment from source trajectory to target trajectory, we achieved most consistent modifications compared to ODE-Based Methods and SDEdit [31].

**Stochastic Dynamics in FM** Rectified Flow models [25,27] leverage linear interpolation between data and prior distribution as forward process, formally $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}$. For backward process, the network optimizes a velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$, and solve the Probability Flow ODE [49] during inference:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{v}(\boldsymbol{x}_t, t)\,\mathrm{d}t \tag{1}$$

The deterministic denoising process, together with linear interpolation, enables high-fidelity generation and fast sampling with simple ODE solvers, such as Euler's method. However, due to linearization error, ODE-based inversion methods [6, 33, 48, 51, 52, 54] often require additional measures to maintain reconstruction quality. We take inspiration from SDE-based inversion techniques to guide denoising trajectory, which effectively mitigate these issues.

Following prior works [16, 29], we derive reverse-time SDE whose evolution yields the same marginal probability density as Eq. (1):

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{v}(\boldsymbol{x}_t, t) - \frac{1}{2} g^2(t) \nabla \log p_t(\boldsymbol{x}_t)]dt + g(t)\,\mathrm{d}\bar{\boldsymbol{w}} \tag{2}$$

where $p_t$ is the probability density generated by Eq. (1), $\mathrm{d}\bar{\boldsymbol{w}}$ is a Wiener process defined backward in time and $g(t)$ is an arbitrary diffusion coefficient. We set $g(t) = \sqrt{2\lambda_{\text{diff}} t}$ in our experiments, leaving $\lambda_{\text{diff}}$ as a hyperparameter to adjust stochasticity during sampling.

**Exploiting the Diffusion Term** Given Eq. (2) which enable stochastic sampling for Rectified Flow models, we solve the backward process with Euler-Murayama method, transforming it into a DDPM-like sampling process:

$$\begin{aligned} \boldsymbol{x}_{t+\Delta t} &\approx \boldsymbol{x}_t + \boldsymbol{v}(\boldsymbol{x}_t, t)\Delta t - \frac{1}{2} g^2(t) \nabla \log p_t(\boldsymbol{x}_t)\Delta t + g(t)\sqrt{|\Delta t|}\boldsymbol{\epsilon}_t \\ &\triangleq \boldsymbol{x}_t + f_t(\boldsymbol{x}_t)\Delta t + g_t\sqrt{|\Delta t|}\boldsymbol{\epsilon}_t \end{aligned} \tag{3}$$

**Algorithm 1** Stochastic Trajectory Manipulation

**Input:** source sample $\boldsymbol{x}_0^{\text{src}}$, $\{t_i\}_{i=1}^{n_{\max}}$, $\mathcal{F}$, $\mathcal{G}$
**Output:** edited sample $\boldsymbol{x}_0^{\text{tgt}}$
**Init:** $\boldsymbol{x}_{t_{\max}}^{\text{src}} \leftarrow \mathcal{G}(t_{\max}; \boldsymbol{x}_0^{\text{src}})$, $\boldsymbol{x}_{t_{\max}}^{\text{tgt}} \leftarrow \boldsymbol{x}_{t_{\max}}^{\text{src}}$
**for** $i = n_{\max}$ **to** 1 **do**
  $\Delta t = t_{i-1} - t_i$
  $\boldsymbol{x}_{t_{i-1}}^{\text{src}} \leftarrow \mathcal{G}(t_{i-1}; \boldsymbol{x}_0^{\text{src}})$
  Compute $f_{t_i}^{\text{src}}(\boldsymbol{x}_{t_i}^{\text{src}}), f_{t_i}^{\text{tgt}}(\boldsymbol{x}_{t_i}^{\text{tgt}}), g_t$
  $\boldsymbol{\epsilon}_{t_i}^{\text{src}} \leftarrow (\boldsymbol{x}_{t_{i-1}}^{\text{src}} - \boldsymbol{x}_{t_i}^{\text{src}} - f_{t_i}^{\text{src}}(\boldsymbol{x}_{t_i}^{\text{src}})\Delta t)/(g_t\sqrt{|\Delta t|})$
  $\boldsymbol{\epsilon}_{t_i}^{\text{tgt}} \leftarrow \mathcal{F}(\boldsymbol{\epsilon}_{t_i}^{\text{src}})$ // depend on mapping strategy
  $\boldsymbol{x}_{t_{i-1}}^{\text{tgt}} \leftarrow \boldsymbol{x}_{t_i}^{\text{tgt}} + f_{t_i}^{\text{tgt}}(\boldsymbol{x}_{t_i}^{\text{tgt}}) + g_t\sqrt{|\Delta t|}\boldsymbol{\epsilon}_{t_i}^{\text{tgt}}$
**end for**
**Return:** $\boldsymbol{x}_0^{\text{tgt}}$

where $\Delta t < 0$ and $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $f_t$ and $g_t$ are drift coefficient and diffusion coefficient respectively, corresponding to a standard SDE. Denoting the source and target denoising trajectory as $\boldsymbol{x}_t^{\text{src}}$ and $\boldsymbol{x}_t^{\text{tgt}}$, as well as corresponding noises $\boldsymbol{\epsilon}_t^{\text{src}}$ and $\boldsymbol{\epsilon}_t^{\text{tgt}}$, we further define a generic noise mapper function $\mathcal{F}$ that maps $\boldsymbol{\epsilon}_t^{\text{src}}$ to $\boldsymbol{\epsilon}_t^{\text{tgt}}$, and a trajectory generator $\mathcal{G}$ that generates samples at noise level $t$ based on clean samples and optionally other conditions. To enable flexible guidance of target trajectories, we formulate a generic algorithm that extracts the noise applied at each step of an arbitrary denoising path and transforms it accordingly. The simplified procedure is described in Algorithm 1.

Notably, the algorithm reduces to a special case of FlowEdit [19] where $n_{\text{avg}} = 1$ and $n_{\min} = 0$, as well as FlowAlign [17] given regularization ratio $\zeta$, with:

$$\mathcal{G}_{\text{FE}}(t; \boldsymbol{x}_0) = \mathcal{G}_{\text{FA}}(t; \boldsymbol{x}_0) = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}_t \tag{4}$$

$$\mathcal{F}_{\text{FE}}(\boldsymbol{\epsilon}_t^{\text{src}}) = \boldsymbol{\epsilon}_t^{\text{src}} + \frac{g(t)\Delta t}{2\sqrt{|\Delta t|}}(\nabla \log p_t^{\text{tgt}}(\boldsymbol{x}_t^{\text{tgt}}) - \nabla \log p_t^{\text{src}}(\boldsymbol{x}_t^{\text{src}})) \tag{5}$$

$$\begin{aligned}\mathcal{F}_{\text{FA}}(\boldsymbol{\epsilon}_t^{\text{src}}) = \boldsymbol{\epsilon}_t^{\text{src}} &+ (\frac{g(t)\Delta t}{2\sqrt{|\Delta t|}} - \frac{\zeta t^2}{1-t})(\nabla \log p_t^{\text{tgt}}(\boldsymbol{x}_t^{\text{tgt}}) - \nabla \log p_t^{\text{src}}(\boldsymbol{x}_t^{\text{src}})) \\ &+ \frac{\zeta}{1-t}(\boldsymbol{x}_t^{\text{src}} - \boldsymbol{x}_t^{\text{tgt}})\end{aligned} \tag{6}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ are i.i.d. Gaussians. This setting provides another interpretation for these two methods from the perspective of noises, and also shows that our method has the potential to be **inversion-free** under certain conditions. A detailed proof is provided in the Supplementary Material.

In our experiments, we primarily employ two trajectory generators:

$$\mathcal{G}_{\text{stc}}(t; \boldsymbol{x}_0) = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}_t \tag{7}$$

$$\mathcal{G}_{\text{trj}}(t; \boldsymbol{x}_0, \{\boldsymbol{x}_t^{\text{pre}}\}) = \boldsymbol{x}_{t+1}^{\text{pre}} + (\boldsymbol{x}_{t+1}^{\text{pre}} - \boldsymbol{x}_0)\Delta t/t \tag{8}$$

where $\{\boldsymbol{x}_t^{\text{pre}}\}$ is a sequence of samples extracted from another denoising trajectory. In our experiments, we adopt the identity mapping as the noise mapper:

$$\mathcal{F}(\boldsymbol{\epsilon}_t^{\text{src}}) = \boldsymbol{\epsilon}_t^{\text{src}} \tag{9}$$

Though several methods utilized SDE in flow matching, they primarily focus on improving sample quality [29,47], expanding search space [16,23] and deriving PF-ODEs [42]. To the best of our knowledge, we are the first to leverage SDE conversion in flow models to achieve stochastic inversion and denoising trajectory manipulation. As shown in Fig. 2, our proposed algorithm outperforms various existing asset variation methods [31, 42].

### 4.3 CAST3D

**Overview** As shown in Fig. 3, our method follows a two-stage pipeline, namely 3D Layout Hinting and Compositional Generation.

**3D Layout Hinting** In this stage, we first generate the main geometry $\boldsymbol{p}_B$ from the anchor image $\boldsymbol{I}_B$, and obtain a layout proposal $\boldsymbol{p}_C$ by applying STM to $\boldsymbol{p}_B$ with prompt $\boldsymbol{c}$. Denoting $\mathcal{D}_{\text{ss}}$ as the flow transformer in Trellis, the proposal process is formulated as:

$$\begin{aligned} \boldsymbol{p}_B &= \mathcal{D}_{\text{ss}}(\boldsymbol{p}_{\text{noise}}; \boldsymbol{I}_B) \\ \boldsymbol{p}_C &= \mathcal{D}_{\text{ss}}^{\text{STM}}(\mathcal{G}_{\text{trj}}(\cdot; \boldsymbol{p}_B, \{\boldsymbol{x}_t^B\}); \boldsymbol{c}, \boldsymbol{o}_B) \end{aligned} \tag{10}$$

where $\boldsymbol{p}_{\text{noise}}$ is Gaussian noise and $\boldsymbol{x}_t^B$ is trajectory of $\boldsymbol{p}_B$, with $\boldsymbol{c}$ and $\boldsymbol{o}_B$ acting as target and source condition used in STM respectively. We also jointly leverage text prompt $\boldsymbol{c}$ and $\boldsymbol{o}_B$, further amplifying modification strength, eventually obtaining a voxelized layout proposal $\boldsymbol{p}_C$. The process is performed entirely in the 3D modality, fully leveraging the spatial knowledge and reasoning capabilities of pretrained 3D diffusion models. This design avoids occlusion and multi-view inconsistency issues commonly encountered in 2D uplifting-based methods. Meanwhile, our STM strategy constrains the generation process, ensuring that the proposed layout remains well aligned with the provided 2D assets.

After proposal, the layouts are extracted at a per-component granularity to obtain layout hints. A text-based segmentation approach is required, with multiple available implementations, such as PartField [26] + MLLM [10]. In practice, for comparison on a fairer parameter scale, we utilize LangSAM [30] for 2D segmentation and modified FlashSplat [43] for 3D backprojection. Specifically, we first turn the voxelized layout proposal $\boldsymbol{p}_C$ into a textured 3D model represented by SLAT $\boldsymbol{z}_C$ using the text prompt $\boldsymbol{c}$. Denoting $\mathcal{D}_{\text{slat}}$ as the SLAT transformer, the generation can be formulated as:

$$\boldsymbol{z}_C = \mathcal{D}_{\text{slat}}(\boldsymbol{z}_{\text{noise}}; \boldsymbol{c}) \tag{11}$$

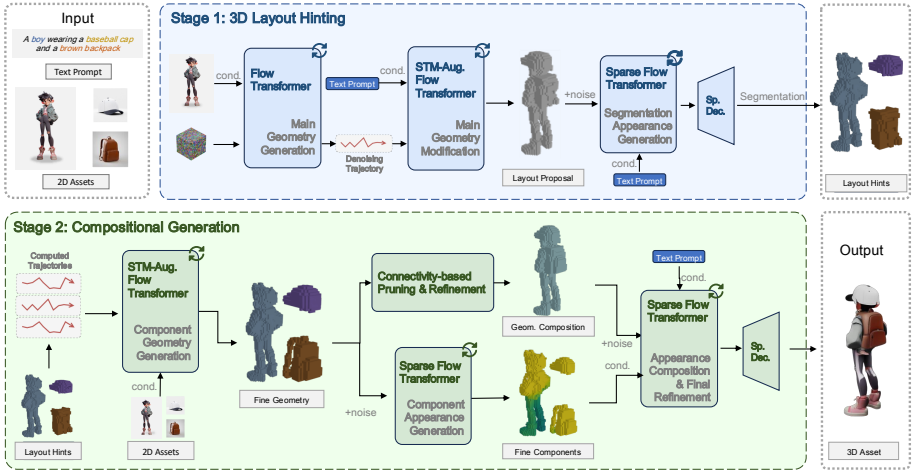


**Fig. 3:** Overview of our two-stage pipeline. **3D Layout Hinting**: A layout proposal is generated through applying STM on the main geometry, and hints are extracted with segmentation approaches. **Compositional Generation**: We generate the components based on layout hints and 2D assets, then compose the geometries and generate appearances of each component, finally collating together to obtain the 3D asset.

where $\boldsymbol{z}_{\text{noise}}$ is Gaussian noise structured as $p_C$.

We decode $\boldsymbol{z}_C$ into 3D Gaussians, render them into multiview images, and back-project the segmentation masks to accumulate votes for Gaussians belonging to each component. Formally, for the $j$-th Gaussian kernel, let $P_j^v$ denote the set of pixels it influences on view $v \in V$. Its membership to mask $M_i$ is determined by:

$$m_j^i = \text{sgn} \sum_{v \in V} \sum_{p \in P_j^v} \alpha_j T_j(p) [\mathbb{I}(p \in M_i^v) - \mathbb{I}(p \notin M_i^v)] \tag{12}$$

where $M_i^v$ is the mask obtained from SAM with prompt $\boldsymbol{o}_i$ within each asset $\{\boldsymbol{o}_i, \boldsymbol{I}_i\}$ on view $v$, and each mask is back-projected separately. We calculate the ratio of masked 3D Gaussians within each voxel, and mask voxels whose ratio exceeds a predefined threshold $\tau_{\text{mask}}$ to suppress artifacts introduced by segmentation noises. This produces a voxelized layout hint for each component, denoted as $\boldsymbol{h}_i$, which is used in the subsequent generation stage.

**Compositional Generation** In this stage, we first generate each component with respect to both the 2D asset $\boldsymbol{I}_i$ and the layout hint $\boldsymbol{h}_i$, producing fine geometry $\boldsymbol{p}_i$. Formally, this procedure can be expressed as:

$$\boldsymbol{p}_i = \mathcal{D}_{\text{ss}}^{\text{STM}}(\mathcal{G}_{\text{stc}}(\cdot; \hat{\boldsymbol{p}}_i); \boldsymbol{I}_i, \emptyset) \tag{13}$$

The fine geometries are then composed through a lightweight connectivity-based pruning strategy for better structural integrity. To be specific, we first

trivially compose the hints as $\boldsymbol{h}_C$, identify the difference between the main geometry $\boldsymbol{p}_B$ and $\boldsymbol{h}_C$, and obtain a composition anchor by restricting the differences only within a certain range of each component other than the main geometry. Formally, the composition anchor is obtained through:

$$\begin{aligned} \hat{\boldsymbol{p}}_{\text{diff}} &= \boldsymbol{p}_B \oplus \boldsymbol{h}_C \\ \boldsymbol{p}_{\text{diff}} &= \hat{\boldsymbol{p}}_{\text{diff}} \odot \sum_{i \neq i_B} \text{Expand}(\boldsymbol{p}_i) \\ \hat{\boldsymbol{p}}_B &= p_B \oplus \boldsymbol{p}_{\text{diff}} \end{aligned} \tag{14}$$

where $\oplus$ denotes XOR operation, $\odot$ denotes Hadmard Product, $\hat{\boldsymbol{p}}_B$ denotes the composition anchor and Expand$(\cdot)$ denote Gaussian blurring.

We further mitigate clipping artifacts by decomposing the residual voxels of the composition anchor relative to other components under 6-connectivity. The largest connected region of residual voxels and regions with large size or unchanged connectivity are preserved to maintain the overall structure while also keeping the floating and thin structures intact. After all the processing, we compose the geometry of all components and obtain coarse geometrical composition $\hat{\boldsymbol{p}}_{GC}$, and refine with STM using generator $\mathcal{G}_{\text{stc}}(\cdot; \hat{\boldsymbol{p}}_{GC})$ and target condition $\boldsymbol{c}$, producing the geometrical composition $\boldsymbol{p}_{GC}$ with smoother geometries.

Simultaneously, another branch generates the appearance of each component $\boldsymbol{z}_i$ conditioned on the image $\boldsymbol{I}_i$.

Finally, we perform appearance composition based on the geometrical composition $p_{GC}$ and component appearances $\boldsymbol{z}_{1:N}$. Specifically, we first locate the regions corresponding to each $\boldsymbol{z}_i$ and assign the optimal velocity $\boldsymbol{v}(\boldsymbol{x}_t, t) = (\boldsymbol{x}_t - \boldsymbol{x}_0)/t$ to ensure consistency with previously generated components. To further maintain smooth transitions, we apply a Gaussian kernel over regions with components correspondences to define soft boundaries, and interpolate the generated content with their associated $\boldsymbol{z}_i$ within these transition zones, formally:

$$\boldsymbol{z}_o = \mathcal{D}_{\text{slat}}(\boldsymbol{z}_{\text{noise}}; \boldsymbol{z}_{1:N}, \boldsymbol{c}) \tag{15}$$

where $\boldsymbol{z}_{\text{noise}}$ has same shape with $p_{GC}$. During sampling, we do not apply any guidance or modifications in the last $\tau_{\text{fine}}$ denoising timesteps to avoid blurry and rough intersections, as widely employed by previous methods.

# 5 Experiments

## 5.1 Experimental Settings

**Implementation Details** Our pipeline is built on Trellis and executed on a single NVIDIA RTX 4090 GPU. We collected 27 objects composed from 45 2D assets as test data. The sampling steps is set to 25, with a rescaling ratio of 3. The classifier-free guidance (CFG) scale for text-condition is set to 7.5 and for image-condition is set to 5.0 for all steps. We choose $n_{\max}$ ranging from $[15, 21]$ depending on the structural changes indicated by the prompt for STM in 3D

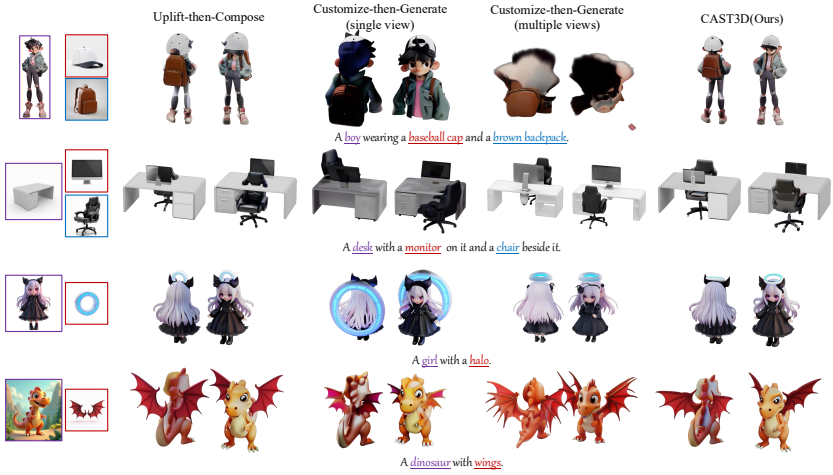


**Fig. 4: Qualitative comparison**. Compared to the baselines, our method generates 3D assets with both semantical coherence and geometrical consistency, achieving a better balance between faithfulness to the reference 2D assets and structural plausibility.

Layout Hinting, and $n_{\max}$ ranging from $[9, 14]$ in Compositional Generation. We set $\lambda_{\text{diff}}$ to 0.6, $\tau_{\text{mask}}$ to 0.9 and $\tau_{\text{fine}}$ to 0.15.

**Baseline Methods** In the absence of prior work on Customized Composition in 3D, we construct two baselines, Uplift-then-Compose and Customize-then-Generate, using SOTA foundational models. For Uplift-then-Compose, we adopt Trellis [59] as the 3D generator, consistent with our method. Since no open-source comparable approach supports text-guided composition of existing 3D objects, we reuse the layout hints extracted by our pipeline to ensure a fair comparison. Concretely, we compute the center of mass of the extracted hints and align them with the mean centroid of the corresponding 3D Gaussians. We also estimate the voxelized volume of each geometry and scale the uplifted asset to match the volumes. For the Customize-then-Generate baseline, we adopt the current state-of-the-art open-source method, Qwen-Image-Edit-2511 [57], as the multi-concept diffusion model to synthesize compositional 2D assets at a comparable parameter scale. We prompt the model to generate both single-view and multi-view images given the assets, and subsequently produce the corresponding 3D assets using the default implementation provided in the official Trellis repository.

**Evaluation Metrics** Due to the inherent complexity of our task, we face the trade-off between faithfulness and structural rationality which cannot be easily measured with existing approaches. We utilize Gemini 3 Pro [10] to evaluate the generation quality. Specifically, we define three evaluation metrics: Faithfulness, Aesthetic, and Structural Rationality, which respectively measure the similarity between each generated 3D object and its reference image, the overall visual

**Table 1: Quantitative comparison**. Our method achieves the best in terms of faithfulness, aesthetic and rationality as evaluated by most advanced VLMs. The baselines either fail on structural rationality or faithfulness preserving, as analyzed previously.

| Method | Faithfulness | Aesthetic | Rationality | Average |
|---|---|---|---|---|
| UtC. | <u>8.33</u> | <u>7.25</u> | 5.29 | <u>6.96</u> |
| CtG. (Single View) | 5.73 | 5.75 | <u>6.87</u> | 6.12 |
| CtG. (Multiple Views) | 6.77 | 5.89 | 6.03 | 6.23 |
| CAST3D | **9.19** | **8.68** | **8.65** | **8.84** |

quality of the generated 3D asset, and the structural plausibility of the composed objects. The prompts used are provided in the Supplementary Material.

### 5.2 Main Result

**Qualitive Comparison** As shown in figure Fig. 4, CAST3D is able to operate across various 2D assets, from synthesized to realistic. Our method produces high-quality 3D assets with close alignment to both the text prompt and reference 2D assets, with plausible scale and spatial positioning. In contrast, the Uplift-then-Compose baseline exhibits severe issue of clipping and artifacts, although reinforced with the extracted layout and volume information. In the meanwhile, the Customize-then-Generate baseline generally aligns well with the text prompt thanks to Qwen-Image-Edit model. However, with single view, the information loss due to conflicting viewpoints results in appearance deviations in the final 3D asset, particularly on occluded regions; with multiple views, the problem of multi-view inconsistency exhibits destructive impacts on cases like the first and last row. In addition, current multi-concept image diffusion models are primarily designed for identity preservation rather than structural preservation, which can further introduce substantial alterations to the subject and degrade overall faithfulness.

**Quantitative Comparison** We present the quantitative comparison in Tab. 1, it can be seen that our method obtains the results that are most consistent with the cognition of most advanced VLMs. Among them, the Uplift-then-Compose baseline receives a lower structural rationality score due to its inability to achieve natural fusion, while the Customize-then-Generate baseline receives lower scores for Faithfulness because it could not guarantee high-quality tracking of the input data. For Aesthetic score, Customize-then-Generate suffer from unintended dark patches in single view and collapses in multiple views, hindering the performance.

**User study** To further assess perceptual faithfulness and structural plausibility, we conducted a user study in which each round presented the textual prompt, reference 2D assets, and shuffled outputs from all methods. Participants rated each result on a 1-10 scale for structural rationality and faithfulness to the inputs. As shown in Tab. 2, CAST3D achieves the highest scores on both metrics,

**Table 2: User study**. Our method achieves the highest score in faithfulness and rationality, demonstrating its ability to generate visually plausible structure while maintaining overall faithfulness.

| Method | Faithfulness | Rationality | Average |
|---|---|---|---|
| UtC. | <u>7.41</u> | 5.92 | 6.66 |
| CtG. (Single View) | 6.49 | 6.34 | 6.41 |
| CtG. (Multiple Views) | 6.85 | <u>6.60</u> | <u>6.72</u> |
| CAST3D | **8.62** | **8.49** | **8.55** |


**Fig. 5: Qualitative Comparison with similar methods**. Tailor3D [40] exhibits significant blurries, while VoxHammer [24] suffers clipping and inconsistencies against 2D assets. Our method maintains the balance of faithfulness and quality successfully.

highlighting its ability to maintain both input fidelity and structurally coherence. Uplift-then-Compose prioritizes fidelity to the source assets and Customize-then-Generate favors structural plausibility, consistent with our analysis. Notably, in the Customize-then-Generate pipeline, human evaluators rate structural rationality higher for multi-view results than for single-view ones, contrary to VLMs' evaluation. We attribute this to VLMs concentrating more on details while human favoring overall structures, where generation with multi-views outperforms.

### 5.3 Additional Results

Aside from the main results, we also compared against similar methods, such as part-level generation [61] and 3D Editing [24,62], yet these approaches target fundamentally different tasks, with part-level generation prioritizing decomposing an asset into parts, while 3D editing perform variations on existing 3D assets. We also notice that some related approaches achieved inspiring results with image generated by industrial-level VLMs. Inspired by these approaches, we adapted Tailor3D [40] and VoxHammer [24] to our task by replacing some procedure in their pipeline with cutting-edge VLMs. Specifically, we first leverage Nano Banana Pro to fuse the 2D assets and produce multi-view images. For Tailor3D, we directly feed the forward and backward view. For VoxHammer, we first em-

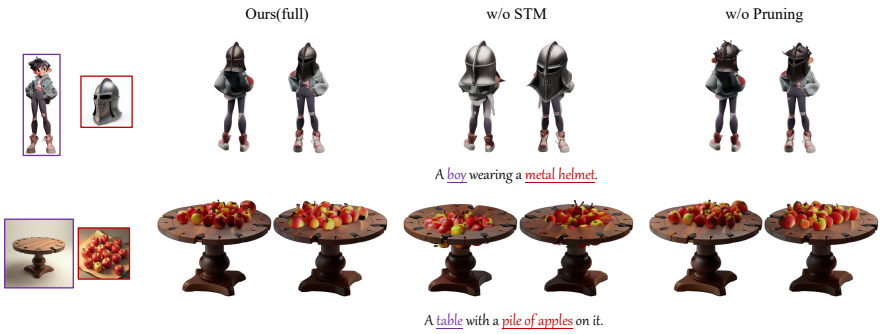


**Fig. 6: Ablation on STM and Pruning**. The results demonstrate the effectiveness of our STM algorithm and connectivity-based pruning strategy, which preserve structural consistency during modification and reduce artifacts during composition.

ploy OmniPart [61] on fused front image, and manually compose and scale the generated parts to form source 3D models and masks for editing, before fed into VoxHammer's pipeline. The result is shown in Fig. 5. These adapted methods perform short of our approach even if equipped with most advanced VLMs.

### 5.4 Ablation Study

We introduce Stochastic Trajectory Manipulation (STM) to preserve global structure while enabling semantic modifications. A lightweight connectivity-based pruning module further suppresses artifacts during composition. We conducted ablation experiments to validate their effectiveness.

Ablation results in Fig. 6 show that removing STM leads to noticeable inconsistencies in layout. Without STM's layout-preserving mechanism, the positional information drifts dramatically, eventually yielding inconsistent compositions. Meanwhile ablating the connectivity-based pruning module causes clipping artifacts after composition, thereby degrading the overall quality and producing visually unnatural results.

### 5.5 Case Studies

**Failure modes of image-based composition** With the results demonstrated in prior sections, we further studied the failure modes of image-based composition methods including most advanced VLMs. As shown in Fig. 7, even Nano Banana Pro demonstrated significant multiview inconsistency when faced with complex composition tasks. When conditioning on single image, inevitable occlusion causes information loss, creating "shadowed" regions (Trellis) or content drift (OmniPart). In the meantime, conditioning on multiview images faces multiview conflicts, which create artifacts or unrealistic structures, eventually hindering the performance. Meanwhile, our approach proposes layouts directly in the 3D domain, leveraging the rich spatial priors of 3D diffusion models and thereby avoiding these issues.

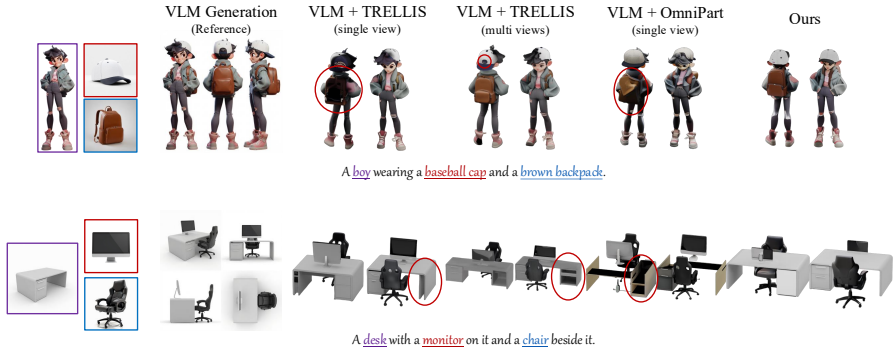


**Fig. 7: Failure modes of image-based compositions**. Cases with images composed by Nano Banana Pro. Unintended dark patches and content drift occur when conditioned on single view, while conditioning on multi views suffer from inconsistencies.


**Fig. 8: Flexibility and Robustness**. Our method tolerates layout hints of low quality or dissimilar shapes. **Left**: Even with imperfect hints due to model capacity, our second stage enables generating geometry consistent with the 2D asset. **Right**: Our method supports layout hints extracted from existing models for customized editing, where we alter the geometry while maintain the positional and scale information.

**Flexibility and Robustness** Thanks to the two-stage design, our method has high tolerance towards layout hints, our intermediate result. In Fig. 8, we fed low quality hints (left) and hints from existing models (right) to the second stage of our pipeline. As shown by our comparison between layout hints and generated geometry, our stochastic trajectory manipulation algorithm in the Compositional Generation stage is able to handle the discrepancies, following the 2D asset with respect to layout information within the hints.

## 6 Conclusion

In this paper, we introduce the task of Customized 3D Composition, which aims to faithfully transform multiple 2D assets into a coherent 3D object. To the best of our knowledge, this is the first attempt to address this task. Our proposed CAST3D tackles this challenge with our novel stochastic trajectory manipulation algorithm and connectivity-based pruning strategy in nearly pure 3D modality. Extensive experiments demonstrate that CAST3D produces coherent and faithful 3D compositions, and can be flexibly extended to other 3D generation scenarios.

# References


1. Anderson, B.D.: Reverse-time diffusion equation models. Stochastic Processes and their Applications **12**(3), 313–326 (1982). https://doi.org/https://doi.org/10.1016/0304-4149(82)90051-5, https://www.sciencedirect.com/science/article/pii/0304414982900515
2. Backhoff-Veraguas, J., Källblad, S., Robinson, B.A.: Adapted wasserstein distance between the laws of sdes. Stochastic Processes and their Applications **189**, 104689 (2025). https://doi.org/https://doi.org/10.1016/j.spa.2025.104689, https://www.sciencedirect.com/science/article/pii/S0304414925001309
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**, 1 – 10 (2023), https://api.semanticscholar.org/CorpusID:256416326
4. Chen, B., Zhao, M., Sun, H., Chen, L., Wang, X., Du, K., Wu, X.: Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. arXiv preprint arXiv:2506.21416 (2025)
5. Chi, Y., Li, X., Huang, Z., Rehg, J.M.: Vinedresser3d: Agentic text-guided 3d editing (2026), https://arxiv.org/abs/2602.19542
6. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=3lge0p5o-M-
7. Dahary, O., Patashnik, O., Aberman, K., Cohen-Or, D.: Be yourself: Bounded attention for multi-subject text-to-image generation (2024)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org, Vienna, Austria (2024)
9. Garibi, D., Yadin, S.S., Paiss, R., Tov, O., Zada, S., Ephrat, A., Michaeli, T., Mosseri, I., Dekel, T.: Tokenverse: Versatile multi-concept personalization in token modulation space. ACM Transactions on Graphics (TOG) **44**, 1 – 11 (2025), https://api.semanticscholar.org/CorpusID:275787987
10. Gemini Team, Google: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
11. Guo, M., Wang, B., He, K., Matusik, W.: Tetsphere splatting: Representing high-quality geometry with lagrangian volumetric meshes. arXiv preprint arXiv:2405.20283 (2024)
12. Hao, S., Han, K., Lv, Z., Zhao, S., Wong, K.Y.K.: ConceptExpress: Harnessing diffusion models for single-image unsupervised concept extraction. In: ECCV (2024)
13. Hitz, M., Robinson, B.A.: Bicausal optimal transport for sdes with irregular coefficients (2025), https://arxiv.org/abs/2403.09941
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
15. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) **42**, 1 – 14 (2023), https://api.semanticscholar.org/CorpusID:259267917

16. Kim, J., Yoon, T., Hwang, J., Sung, M.: Inference-time scaling for flow models via stochastic generation and rollover budget forcing. arXiv preprint arXiv:2503.19385 (2025)
17. Kim, J., Hong, Y., Park, J., Ye, J.C.: Flowalign: Trajectory-regularized, inversion-free flow-based image editing (2025), https://arxiv.org/abs/2505.23145
18. Kong, Z., Zhang, Y., Yang, T., Wang, T., Zhang, K., Wu, B., Chen, G., Liu, W., Luo, W.: Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In: European Conference on Computer Vision (2024), https://api.semanticscholar.org/CorpusID:268512816
19. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
20. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion (2023)
21. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
22. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
23. Li, J., Cui, Y., Huang, T., Ma, Y., Fan, C., Yang, M., Zhong, Z.: Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde (2025), https://arxiv.org/abs/2507.21802
24. Li, L., Huang, Z., Feng, H., Zhuang, G., Chen, R., Guo, C., Sheng, L.: Voxhammer: Training-free precise and coherent 3d editing in native 3d space. In: Thirteenth International Conference on 3D Vision (2026), https://openreview.net/forum?id=UhHNN5lW67
25. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
26. Liu, M., Uy, M.A., Xiang, D., Su, H., Fidler, S., Sharp, N., Gao, J.: Partfield: Learning 3d feature fields for part segmentation and beyond (2025)
27. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z
28. Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20654–20664 (2024)
29. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision (2024), https://api.semanticscholar.org/CorpusID:267027717
30. Medeiros, L.: Language segment anything (lang-sam). https://github.com/luca-medeiros/lang-segment-anything (2023)
31. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

33. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. arXiv preprint arXiv:2211.09794 (2022)
34. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., Li, M., Liu, M., Zhang, Y., Wu, S., Zhao, S., Zhang, J., He, Q., Wu, X.: Dreamo: A unified framework for image customization (2025), https://arxiv.org/abs/2504.16915
35. Ouyang, Y., Chai, W., Ye, J., Tao, D., Zhan, Y., Wang, G.: Chasing consistency in text-to-3d generation from a single image. CoRR **abs/2309.03599** (2023), https://doi.org/10.48550/arXiv.2309.03599
36. Parmar, G., Patashnik, O., Wang, K.C., Ostashev, D., Narasimhan, S., Zhu, J.Y., Cohen-Or, D., Aberman, K.: Object-level visual prompts for compositional image generation (2025), https://arxiv.org/abs/2501.01424
37. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7932–7942 (2023), https://api.semanticscholar.org/CorpusID:259108247
38. Po, R., Yang, G., Aberman, K., Wetzstein, G.: Orthogonal adaptation for modular customization of diffusion models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7964–7973 (2023), https://api.semanticscholar.org/CorpusID:265659333
39. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=FjNys5c7VyY
40. Qi, Z., Yang, Y., Zhang, M., Xing, L., Wu, X., Wu, T., Lin, D., Liu, X., Wang, J., Zhao, H.: Tailor3d: Customized 3d assets editing and generation with dual-side images (2024), https://arxiv.org/abs/2407.06191
41. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Mildenhall, B., Ruiz, N., Zada, S., Aberman, K., Rubenstein, M., Barron, J., Li, Y., Jampani, V.: Dreambooth3d: Subject-driven text-to-3d generation. ICCV (2023)
42. Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.S.: Semantic image inversion and editing using rectified stochastic differential equations. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=Hu0FSOSEyS
43. Shen, Q., Yang, X., Wang, X.: Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. European Conference of Computer Vision (2024)
44. Shen, T., Munkberg, J., Hasselgren, J., Yin, K., Wang, Z., Chen, W., Gojcic, Z., Fidler, S., Sharp, N., Gao, J.: Flexible isosurface extraction for gradient-based mesh optimization. ACM Transactions on Graphics (TOG) **42**, 1 – 16 (2023), https://api.semanticscholar.org/CorpusID:260167800
45. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: MVDream: Multi-view diffusion for 3d generation. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=FUgrjq2pbB
46. Shu, Z., Yu, J., Chao, K., Xin, S., Liu, L.: Gaussedit: Adaptive 3d scene editing with text and image prompts. IEEE Transactions on Visualization and Computer Graphics **31**(10), 7769–7780 (2025). https://doi.org/10.1109/TVCG.2025.3556745
47. Singh, S., Fischer, I.: Stochastic sampling from deterministic flow models. ArXiv **abs/2410.02217** (2024), https://api.semanticscholar.org/CorpusID:273098480

48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP
49. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=PxTIG12RRHS
50. Team, T.H.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation (2025)
51. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1921–1930 (2022), https://api.semanticscholar.org/CorpusID:253801961
52. Wallace, B., Gokul, A., Naik, N.V.: Edict: Exact diffusion inversion via coupled transformations. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22532–22541 (2022), https://api.semanticscholar.org/CorpusID:253761481
53. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
54. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. ArXiv **abs/2411.04746** (2024), https://api.semanticscholar.org/CorpusID:273877800
55. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. ArXiv **abs/2406.07209** (2024), https://api.semanticscholar.org/CorpusID:270379835
56. Wang, Y., Yi, X., Xu, Q., Zhou, Y., Chen, L., Zhang, H.: Personalize your gaussian: Consistent 3d scene personalization from a single image. ArXiv **abs/2505.14537** (2025), https://api.semanticscholar.org/CorpusID:278768450
57. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
58. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. ArXiv **abs/2506.18871** (2025), https://api.semanticscholar.org/CorpusID:279999713
59. 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 Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
60. Yang, Y., Wang, W., Peng, L., Song, C., Chen, Y., Li, H., Yang, X., Lu, Q., Cai, D., Wu, B., Liu, W.: Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. ArXiv **abs/2403.11627** (2024), https://api.semanticscholar.org/CorpusID:268532545
61. Yang, Y., Zhou, Y., Guo, Y.C., Zou, Z.X., Huang, Y., Liu, Y.T., Xu, H., Liang, D., Cao, Y.P., Liu, X.: Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. arXiv preprint arXiv:2507.06165 (2025)

62. Ye, J., Xie, S., Zhao, R., Wang, Z., Yan, H., Zu, W., Ma, L., Zhu, J.: NANO3d: A training-free approach for efficient 3d editing without masks. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=jov79sMFHn
63. Zhuang, J., Kang, D., Cao, Y.P., Li, G., Lin, L., Shan, Y.: Tip-editor: An accurate 3d editor following both text-prompts and image-prompts. ACM Trans. Graph. **43**(4) (Jul 2024). https://doi.org/10.1145/3658205, https://doi.org/10.1145/3658205