MVCUSTOM: MULTI-VIEW CUSTOMIZED DIFFUSION VIA GEOMETRIC LATENT RENDERING AND COMPLETION

Anonymous authors
Paper under double-blind review

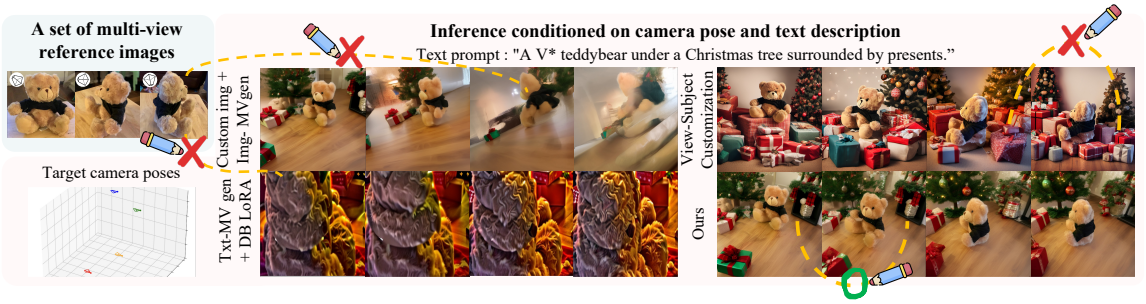

Figure 1: **Comparison between MVCustom and existing approaches extended to multi-view customization.** The light blue box shows the reference multi-view images and corresponding camera poses of a customized object. The 'X' marks indicate regions inconsistent with either the reference object's appearance or across views, while 'O' marks indicate well-maintained consistency. Our approach clearly outperforms existing methods by achieving accurate viewpoint alignment and robust multi-view consistency for both the customized object and novel surroundings generated from diverse textual prompts.

ABSTRACT

Multi-view generation with camera pose control and prompt-based customization are both essential elements for achieving controllable generative models. However, existing multi-view generation models do not support customization with geometric consistency, whereas customization models lack explicit viewpoint control, making them challenging to unify. Motivated by these gaps, we introduce a novel task, *multi-view customization*, which aims to jointly achieve multi-view camera pose control and customization. Due to the scarcity of training data in customization, existing multi-view generation models, which inherently rely on large-scale datasets, struggle to generalize to diverse prompts. To address this, we propose *MVCustom*, a novel diffusion-based framework explicitly designed to achieve both multi-view consistency and customization fidelity. In the training stage, MVCustom learns the subject's identity and geometry using a feature-field representation, incorporating the text-to-video diffusion backbone enhanced with dense spatio-temporal attention, which leverages temporal coherence for multi-view consistency. In the inference stage, we introduce two novel techniques: *depth-aware feature rendering* explicitly enforces geometric consistency, and *consistent-aware latent completion* ensures accurate perspective alignment of the customized subject and surrounding backgrounds. Extensive experiments demonstrate that *MVCustom* achieves the most balanced and consistent competitive performance across multi-view consistency, customization fidelity, demonstrating effective solution of multi-objective generation task.

Task	Method	Fidelity	Holistic	S.MV	H.MV
(a) Customization	DreamBooth, CustomDiffusion, etc.	O	O	X	X
(b) Subject-only text-to-MV gen.	FlexGen, Make-Your-3D, etc.	X	X	O	X
(c) Text-to-MV generation	CameraCtrl, ViewDiff, etc.	X	O	O	O
(d) Subject-only image-to-MV gen.	SV3D, SyncDreamer, etc.	X	X	O	X
(e) Image-to-MV gen.	SEVA, CAT3D, ViewCrafter, etc.	X	O	O	O
(f) Viewpoint-aware subject custom.	CustomDiffusion360, CustomNet	O	O	O	X
(g) Multi-view customization	**MVCustom (ours)**	O	O	O	O

Table 1: **Comparison of existing tasks and representative methods.** *Fidelity* refers to preserving object identity from reference images and alignment with textual prompts in customization. *Holistic* denotes whether both subjects and the surroundings described in a prompt are synthesized. *S.MV* evaluates whether subjects remain consistent across different viewpoints. *H.MV* consistency refers to whether both subjects and their surroundings are holistically consistent across viewpoints. *MV* stands for multi-view.

1 INTRODUCTION

As generative models advance rapidly, users are increasingly demanding fine-grained controllability. Among the essential elements, two forms of control are significant: camera control and customization. First, *camera control* is to generate images for specified viewpoints, which is essential in domains such as 3D understanding. In particular, ensuring camera pose control and multi-view consistency for both the subject and its surroundings is crucial for realistic and immersive content, as misalignment across views severely undermines geometric coherence. Second, *customization* is to capture user-specific subjects, or concepts, supporting personalized content generation and supporting applications such as creative media and design prototyping, *etc*.

While each form of control is valuable on its own, integrating them unlocks significantly richer applications. A unified framework that supports both capabilities enables 3D customization for virtual prototyping and personalized asset generation, where both user-specific fidelity and geometric consistency are indispensable. Moreover, it broadens the scope of controllable generative models, enabling realistic, immersive, and user-tailored content beyond the reach of existing approaches. To this end, we introduce the novel task of *multi-view customization*, which requires (1) generating images that adhere to specified camera parameters for consistent perspective alignment, (2) preserving subject identity provided by reference images, and (3) coherently adapting both subjects and their surrounding context to diverse textual prompts.

However, to the best of our knowledge, no prior method fully satisfies the requirements of the multi-view customization. As summarized in Tbl. 1, conventional customization methods (Lee et al., 2024; Ruiz et al., 2023; Kumari et al., 2024) preserve reference identity and align with prompts, but lack viewpoint control. Most multi-view generation methods focus only on subjects, neglecting consistent surroundings across views (cases b, d in Tbl. 1). Some holistic multi-view generation methods (He et al., 2024; Zhou et al., 2025) provide full-frame consistency but do not support personalization to novel reference concepts (cases c, e). Viewpoint-aware subject customization methods (Kumari et al., 2024; Yuan et al., 2023) remain subject-centric, leading to inconsistent surroundings across views (case f). These limitations underscore the need for a new approach explicitly designed for multi-view customization.

Directly adopting multi-view generation frameworks, which rely heavily on large-scale training data, is infeasible in the customization setting, where only a few reference images are available. A straightforward baseline applies conventional customization methods (Ruiz et al., 2023; Hu et al., 2021) directly to text-conditioned multi-view backbones (c in Tbl. 1), but this approach cannot preserve subject identity and reduces camera pose control ability. Another naive baseline generates a single customized image, then applies image-conditioned multi-view generation models (f in Tbl. 1), but the inherent ambiguity of a single view leads to inconsistent spatial relationships and degraded fidelity, as illustrated in Fig. 1.

To address these challenges, we propose *MVCustom*, a diffusion-based framework explicitly designed for robust multi-view customization. Our method separates training and inference stages to effectively handle limited data and ensure geometric consistency across diverse prompts. In the training stage, we leverage pose-conditioned transformer blocks (Kumari et al., 2024). However, a key change is using the video diffusion backbone enhanced with dense spatio-temporal attention to transfer temporal coherence into holistic-frames consistency, ensuring spatial coherence of both the subject and their surroundings across views. At inference, the key challenge is ensuring multi-view geometric consistency for novel prompts, particularly for the subject's surroundings that lack supervision from limited training data. To address this, we introduce two novel inference-stage techniques: *depth-aware feature rendering*, which explicitly enforces geometric consistency using inferred 3D scene geometry, and *consistent-aware latent completion*, which naturally completes previously unseen regions revealed by viewpoint shifts. Extensive comparisons demonstrate that MVCustom is the only approach that effectively integrates accurate multi-view generation and high-fidelity customization.

Our contributions are summarized as follows:

- We propose a novel task, *multi-view customization*, clearly define its requirements, and systematically analyze the limitations of existing methods and tasks.
- We introduce a video diffusion-based backbone enhanced with dense spatio-temporal attention modules, effectively transferring temporal coherence into multi-view consistency.
- To accommodate limited data in customization, we propose two novel inference-stage methods: *depth-aware feature rendering* for explicit geometric consistency, and *consistent-aware latent completion* for consistent and realistic completion of disoccluded regions.

2 Related work

Conventional text-based customization. Customization methods generate images guided by textual prompts while preserving identities from reference images, typically by learning concept-specific embeddings (Gal et al., 2022), fine-tuning models (Ruiz et al., 2023), or applying lightweight adaptations (Hu et al., 2021). Recent approaches further enhance text-image alignment (Alaluf et al., 2023; Li et al., 2024a) and multi-subject control (Kumari et al., 2023; Kwon & Ye, 2024). However, these methods typically lack explicit control over viewpoint. Some works achieve pose-variant compositions (Li et al., 2024b; Song et al., 2024), but do not support explicit camera pose control. Methods like CustomDiffusion360 (Kumari et al., 2024) and CustomNet (Yuan et al., 2023) incorporate viewpoint control yet remain predominantly subject-centric, neglecting to coherently represent their surroundings. In contrast, our proposed *MVCustom* explicitly ensures robust spatial coherence for both customized subjects and surroundings across diverse viewpoints.

Multi-view generation. Multi-view generation models (Zhao et al., 2025; Tang et al., 2024; Alper et al., 2025; Shin et al., 2023) focus on synthesizing consistent multiple views. However, these models typically require large datasets to learn 3D geometry and inpaint newly visible regions, making them unsuitable for customization with only a few reference images. An alternative approach may involve applying conventional customization methods directly onto multi-view generation backbones. Nevertheless, text-conditioned multi-view generation models (Höllein et al., 2024; Shi et al., 2023; Tang et al., 2023; Huang et al., 2024) are limited by the scarcity of paired text and multi-view data, leading to poor adaptability to diverse textual prompts. Another related approach utilizes multi-view diffusion models (Long et al., 2024) for novel-view synthesis from a single reference image, enabling subject-aware editing in multi-view settings (Liu et al., 2024). However, these methods primarily focus only subject editing. In contrast, our *MVCustom* framework explicitly addresses these challenges, combining effective 3D geometry learning with explicit inference-time geometric constraints, enabling robust multi-view consistency and precise alignment with diverse textual prompts.

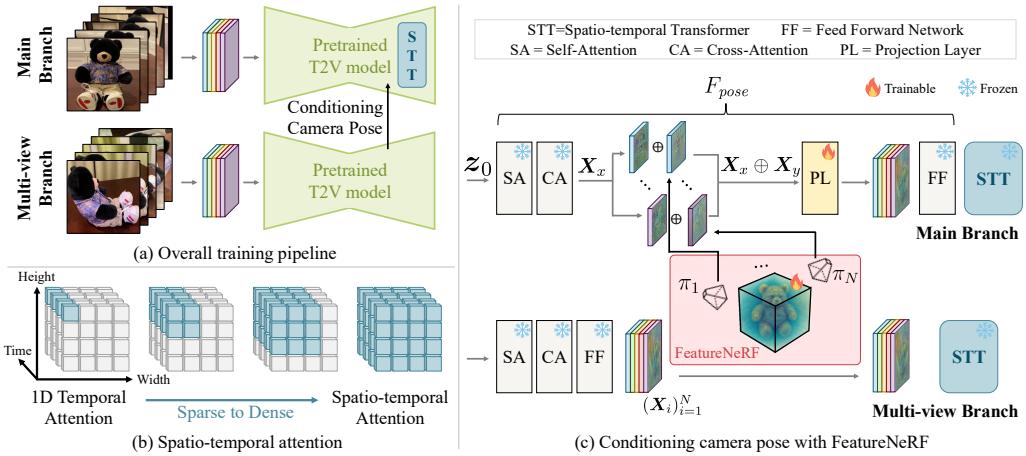

Figure 2: **Overview.** (a) The overall training pipeline, depicting how camera pose conditioning operates with two branches, the main and multi-view. (b) Visualization of our progressive attention mechanism. We gradually broaden the spatial attention field, enhancing geometric consistency. (c) The detailed illustration of the pose-conditioned transformer block. FeatureNeRF and a projection layer are trained to produce a feature map, obtained by concatenating the main-branch and multi-view feature map.

3 METHODOLOGY

In this section, we first introduce our multi-view customization task, explicitly incorporating camera viewpoint control (Sec. 3.1). Next, we describe pose-conditioned transformer blocks to reflect camera poses into the customized subject (Sec. 3.2). Then, we introduce our video diffusion backbone designed for large viewpoint changes (Sec. 3.3). Finally, we present our core contributions — *depth-aware feature rendering* and *consistent-aware latent completion* — to ensure multi-view consistency not only of the customized subject but also their surroundings under novel textual prompts (Sec. 3.4).

3.1 PROBLEM DEFINITION

We define *multi-view customization* as an extension of traditional customization that incorporates explicit control over camera viewpoints. Traditional customization aims to model the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{Y}', \boldsymbol{c})$, where $\boldsymbol{c}$ is a textual prompt describing a novel concept and $\boldsymbol{Y}' = \{\boldsymbol{y}'_i\}_{i=1}^N$ are reference images. A common approach is textual inversion (Gal et al., 2022), which introduces a learnable embedding vector $\boldsymbol{v}$ that replaces part of the text prompt $\boldsymbol{c}(\boldsymbol{v})$. The embedding is learned by minimizing the denoising objective, $\boldsymbol{v}^* = \arg\min_{\boldsymbol{v}} \mathbb{E}_{\boldsymbol{x},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{x}_t; \boldsymbol{c}(\boldsymbol{v}), t)\|_2^2\right]$, where t denotes the diffusion timestep.

In multi-view customization, each reference image is paired with its camera pose, $\boldsymbol{Y} = \{(\boldsymbol{y}_i, \pi_i)\}_{i=1}^N$. The goal is to model the conditional distribution

$$p(\boldsymbol{x}_{0:M} \mid \boldsymbol{Y}, \boldsymbol{c}, \{\phi_m\}_{m=0}^M), \tag{1}$$

where $\boldsymbol{x}_{0:M} = \{\boldsymbol{x}_m\}_{m=0}^M$ denotes a set of generated images under target camera poses $\{\phi_m\}$. For brevity, we denote the set of multi-view outputs as $\boldsymbol{x}$ in the following sections. This formulation enables explicit camera pose control in addition to identity preservation and text alignment, thereby enhancing controllability, consistency, and realism of the generated results.

3.2 CONDITIONING CAMERA POSE IN DIFFUSION MODELS

To effectively learn the subject's geometry from reference data, we adopt the pose-conditioned transformer block from CustomDiffusion360 (Kumari et al., 2024), replacing the original spatial transformer in the diffusion models. The transformer block is defined as $F_{pose}(\boldsymbol{z}_0, \{(\boldsymbol{z}_i, \pi_i)\}_{i=1}^N, \boldsymbol{c}, \phi)$, where $\boldsymbol{z}_0$ is the main-branch feature map and $\{(\boldsymbol{z}_i, \pi_i)\}$ are reference features with corresponding poses.

The two branches play complementary roles:

- **Main branch.** Generates target-view features for decoding into the final image. Its feature map is refined via self-attention s and cross-attention g modules conditioned on $\boldsymbol{c}$: $\boldsymbol{X}_x := g(s(\boldsymbol{z}_0), \boldsymbol{c})$.
- **Multi-view branch.** Aggregates reference-view features $\{\boldsymbol{X}_i\}$, computed as $\boldsymbol{X}_i := f(g(s(\boldsymbol{z}_i), \boldsymbol{c}))$. FeatureNeRF synthesizes a pose-aligned feature map $\boldsymbol{X}_y$ by combining $\{\boldsymbol{X}_i\}$ with camera poses $\{\pi_i\}$ via epipolar geometry (Yu et al., 2021) and volume rendering (Mildenhall et al., 2021):

$$\boldsymbol{X}_y := \text{FeatureNeRF}(\{(\boldsymbol{X}_i, \pi_i)\}_{i=1}^N, \boldsymbol{c}, \phi).$$

These feature maps are concatenated and projected into the backbone's feature space, as shown in Fig. 2a.

3.3 BACKBONE FOR DYNAMIC VIEW CHANGE

A pose-conditioned transformer block F_{pose} generally produces consistent multi-view images about the subject, but novel surroundings or clothings are often become inconsistent across views. To address this, we repurpose video generation into multi-view generation based on AnimateDiff (Guo et al., 2023), inherently suited for handling viewpoint transitions. Our video denoising model D_θ is defined as:

$$D_\theta : (\tilde{\boldsymbol{x}}_{1:N}; \boldsymbol{Y}, \boldsymbol{c}, \phi_{1:N}) \mapsto \hat{\boldsymbol{x}}_{1:N}, \tag{2}$$

mapping noisy inputs $\tilde{\boldsymbol{x}}_{1:N}$ to clean frames $\hat{\boldsymbol{x}}_{1:N}$, conditioned on camera poses $\phi_{1:N}$.

AnimateDiff's 1D temporal attention limits its interactions to identical spatial positions, hindering effective modeling of viewpoint-induced displacements. We extend it with dense 3D spatio-temporal attention (STT) for richer context modeling. To preserve stability and pretrained knowledge, we gradually expand the spatial attention field of STT during training (Fig. 2b). The detailed design choices are discussed in Sec. A.

With this backbone, we fine-tune our customized model by incorporating textual inversion and a pose-conditioned transformer block, optimizing with a standard denoising and additional FeatureNeRF losses (please see Sec. B for the details).

3.4 INFERENCE-TIME MULTI-VIEW CONSISTENCY UNDER LIMITED DATA

Depth-aware feature rendering. Although our video backbone (Sec. 3.3) produces coherent surroundings, it does not explicitly enforce geometric consistency under camera motion. To address this, we propose *depth-aware feature rendering*, which explicitly imposes geometric constraints conditioned on novel prompts during inference. Unlike previous depth-conditioned multi-view generation methods (Ren et al., 2025; Yu et al., 2024), which rely on large-scale training data, our method effectively addresses the lack of geometric supervision for novel prompt-driven content.

First, the *anchor feature mesh* $\mathcal{M}_a$ is defined using an anchor frame $\hat{\boldsymbol{x}}_a$ selected from $\hat{\boldsymbol{x}}_{1:N}$, denoted as $\mathcal{M}_a = (\boldsymbol{P}_a, \boldsymbol{F}_a, \mathcal{T}_a)$, where the anchor frame's feature map $\boldsymbol{F}_a$ is directly used as texture of mesh.[1]. The vertices $\boldsymbol{P}_a \in \mathcal{R}^{H \times W \times 3}$ are derived from the depth map D, estimated by an off-the-shelf depth

[1] $\boldsymbol{F}_a$ is the feature map taken immediately before the spatial transformer in the second up-block (Fig. 2c), a feature level previously demonstrated to be effective for diffusion-based feature modification (Go et al., 2024).

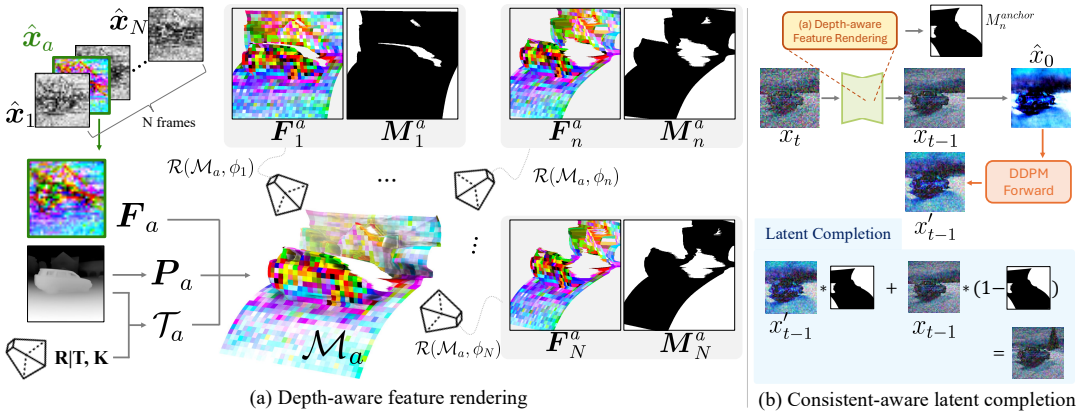

Figure 3: (a) Anchor feature mesh $\mathcal{M}_a$, consists of a texture $\boldsymbol{F}_a$, vertices $\mathbf{P}_a$, and triangles $\mathcal{T}_a$, is constructed using the feature and depth maps, and camera pose of the anchor frame. The $\mathcal{M}_a$ is used to render the projected feature maps for the other camera poses. (b) Completion via latent perturbation for new visible areas.

estimator (Bhat et al., 2023) applied to $\hat{\boldsymbol{x}}_a$. To align the estimated depth $\hat{D}$ with FeatureNeRF's geometric scale, we normalize $\hat{D}$ and shift it by the median depth d_{med} of the anchor view: $D \leftarrow \text{norm}(\hat{D}) + d_{\text{med}}$. The depth map D is resized to the feature resolution (H_F, W_F) of $\boldsymbol{F}_a$. Using rotation $R \in \mathbb{R}^{3\times3}$, translation $T \in \mathbb{R}^3$, and intrinsic matrix $K \in \mathbb{R}^{3\times3}$ of the camera parameters associated with $\hat{\boldsymbol{x}}_a$, the 3D points are computed as $\boldsymbol{P} = R(DK^{-1}[u, v, 1]^\top) + T$, where $[u, v]$ denotes a feature-space coordinate. Dense mesh triangles $\mathcal{T}_a$ are defined on the pixel grid using $\hat{D}$, while pruning the regions that become newly visible from other viewpoints, yielding discontinuous mesh boundaries (see Fig. 3a, $\mathcal{M}_a$).

Second, we render $\mathcal{M}_a$ for a given camera pose ϕ_n, producing the rendered feature map $\boldsymbol{F}_n^a$ and visibility masks $\boldsymbol{M}_n^a$. Notice that the rendering is performed in the feature-space of $\boldsymbol{F}_a$:

$$\boldsymbol{F}_n^a, \boldsymbol{M}_n^a = \mathcal{R}(\mathcal{M}_a, \phi_n), \quad 1 \le n \le N,\ n \ne a, \tag{3}$$

where $\mathcal{R}$ denotes a differentiable mesh renderer.

Finally, during the first 35 steps of the 50-step DDIM sampling process, we update each feature map by replacing masked regions with rendered anchor features:

$$\hat{\boldsymbol{F}}_n = \boldsymbol{M}_n^a \odot \boldsymbol{F}_n^a + (1 - \boldsymbol{M}_n^a) \odot \boldsymbol{F}_n, \quad 1 \le n \le N,\ n \ne a, \tag{4}$$

then, we substitute the combined feature map $\hat{\boldsymbol{F}}$ for $\boldsymbol{F}$ before the spatial transformer in the second up-block.

Consistent-aware latent completion. Regions where $(1 - \boldsymbol{M}_n^a)$ is nonzero correspond to newly visible areas that requires content generation not present in the anchor frame. To address this, we introduce *consistent-aware latent completion*, which leverages stochastic perturbations to synthesize these 'disoccluded' regions (see Fig. 3b). Specifically, given an intermediate noisy latent x_t in the denoising process, we predict an initial latent x_0 that is semantically meaningful yet incomplete. We then reintroduce noise into x_0 via the forward diffusion process, reverting to the original timestep t and yielding a perturbed latent x'_t. The disoccluded regions in the original latent x_t are selectively replaced with those from x'_t, enforcing spatial coherence across frames through the temporal consistency of the video backbone. This procedure is iteratively conducted from timestep T down to an early timestep τ (close to T), allowing semantic flexibility and coherent synthesis of novel details in newly exposed regions. Further implementation details, including anchor mesh construction and inference pseudo-code, are provided in Sec. B.

Figure 4: **Qualitative results.** The light blue boxes indicate the multi-view training dataset for the target concept, while the light pink boxes illustrate the inference phase, where results are conditioned on new text and target camera poses.

4 Experiment

4.1 Experimental setup

Dataset. We train our video diffusion backbone using a subset (430K samples) of the WebVid10M dataset (Bain et al., 2021). For customization experiments, we use concepts selected from the Common Objects in 3D (CO3Dv2) dataset (Reizenstein et al., 2021), following the setup in CustomDiffusion360 (Kumari et al., 2024). Specifically, we select four categories—car, chair and motorcycle—with three concepts per category. For evaluation, we randomly sample camera trajectories from the CO3Dv2 test set as target camera poses.

Competitors. As our task is novel, we compare our proposed method against various applicable baseline approaches: (1) *Custom img + Img-MVgen:* This method generates multi-view images by inputting a single customized image into the image-conditioned multi-view generation model, SEVA (Zhou et al., 2025). The single input image is taken from the first frame of the output produced by our model, conditioned on the

Method	MV Generation		Customization		Inference Cost	
	Camera Pose Accuracy (↑)	Multi-view Consistency (↓)	Identity Preservation (↓)	Text Alignment (↑)	Time (s)	GPU (GB)
Custom Img + Img-MV gen	0.675 ± 0.12	0.214 ± 0.15	0.504 ± 0.12	0.676 ± 0.11	96.18	6.73
Txt-MV gen with DB	0.283 ± 0.25	0.116 ± 0.09	0.557 ± 0.12	0.723 ± 0.10	27.20	5.42
CustomDiffusion360	0.000 ± 0.00	0.190 ± 0.11	0.417 ± 0.12	0.806 ± 0.10	74.97	4.99
MVCustom (Ours)	0.735 ± 0.10	0.121 ± 0.10	0.448 ± 0.11	0.744 ± 0.10	130.92	19.29

Table 2: **Quantitative comparison on multi-view generation, customization, and inference cost.** We highlight the best score in light red and the second-best in yellow.

target text and camera pose. (2) *Txt-MVgen with DB:* A text-conditioned camera-motion-controllable model, CameraCtrl (He et al., 2024), customized with the conventional DreamBooth-LoRA (Ryu, 2023) approach. (3) *CustomDiffusion360:* An existing object viewpoint-controllable customization method (Kumari et al., 2024). Further comparisons and detailed discussions regarding additional competitors' capabilities and limitations are provided in Sec. C.

Evaluation metrics. We evaluate our method using four metrics: camera pose accuracy, multi-view consistency, text alignment, and identity preservation. Camera pose accuracy is measured as the average inter-frame relative rotation accuracy (range: [0, 1]), computed via COLMAP (Schonberger & Frahm, 2016). If COLMAP fails to reconstruct camera poses, we assign the minimal accuracy score (0). Multi-view consistency is quantified by visual similarity (Fu et al., 2023) across views, computed over all view pairs. Identity preservation is measured via DreamSim similarity (Fu et al., 2023) between generated outputs and reference images. Text alignment is evaluated using CLIP similarity scores between textual prompts and generated images. Further details and additional evaluations are provided in Sec. C.

4.2 Results

As shown quantitatively in Tbl. 2 and qualitatively in Fig. 4, MVCustom is the only approach that simultaneously achieves high multi-view consistency and accurate customization fidelity. More comprehensive video comparisons can be found in the supplementary material ("mvcustom.html").

Multi-view consistency with perspective alignment. Accurately reflecting target camera poses is crucial for multi-view customization. As shown in Tbl. 2 (camera pose accuracy) and qualitative examples (Fig. 4), MVCustom faithfully generates multi-view images aligned with specified viewpoints. In contrast, *Txt-MV gen with DB* fails to reflect rotation-aware trajectories despite explicit conditioning, as clearly observed in the chair example of Fig. 4, and confirmed by poor pose accuracy (Tbl. 2). This indicates that the strong camera controllability in Txt-MV generation does not directly translate into multi-view customization through conventional fine-tuning (see Sec. D). Similarly, *Img-MV gen* methods rely on a single reference image, limiting subject appearance and geometry, and causing unnatural subject–surrounding relationships in distant views (e.g., the motorcycle in Fig. 4). Although *CustomDiffusion360* maintains subject consistency, arbitrary surroundings across viewpoints yield poor holistic multi-view consistency, leading to COLMAP reconstruction failure and zero pose accuracy (Tbl. 2). By leveraging our video backbone and inference strategies, MVCustom substantially improves holistic multi-view consistency and perspective alignment, outperforming all baselines.

As shown in Tbl. 2, MVCustom requires higher computational resources primarily due to the external depth estimator (increasing GPU memory) and the feature replacement step (increasing inference time), unlike other competitors relying solely on denoising. Nevertheless, explicitly enforcing geometric consistency at inference is critical given the constraint of extremely limited training data. Thus, we argue that our significant

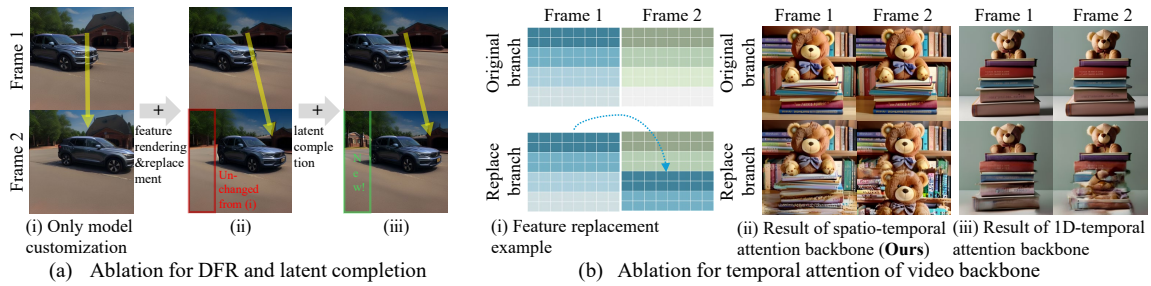

Figure 5: **Results of ablation studies.** (a) Stepwise effect of applying depth-aware feature rendering (DFR) and consistent-aware latent completion under x-translation camera pose. (b) Impact of temporal attention on feature replacement. (i) Feature replacement vertically copies the feature map from frame 1 to frame 2. Our method successfully enforces spatial flow, whereas 1D temporal attention fails to capture the intended translation.

improvements in multi-view consistency, geometric accuracy, and customization fidelity clearly justify this computational trade-off.

ID preservation with text alignment The *Custom img + Img-MV gen* baseline fails to preserve subject identity and the textual description of surroundings, particularly as viewpoints move further from the input image (as shown qualitatively in Fig. 4). *Txt-MV gen with DB* also fails to retain the reference subject's appearance and geometry, leading to poor identity preservation. In contrast, both *CustomDiffusion360* and our MVCustom method successfully preserve the reference subject and effectively reflect diverse textual prompts across all views, demonstrating superior customization fidelity.

4.3 Ablation Study

Depth-aware feature rendering & Consistent-aware latent Ccmpletion. Customization fine-tuning alone yields static surroundings despite varying subject poses (Fig. 5a-i). Our novel depth-aware feature rendering enforces geometric consistency, enabling accurate spatial shifts (e.g., building position) according to camera movements (Fig. 5a-ii). However, newly revealed regions reuse previous content, reducing realism. Thus, we propose latent completion, leveraging the generative power of our diffusion backbone to naturally synthesize previously unseen, context-appropriate details (Fig. 5c). Unlike conventional multi-view methods requiring extensive datasets, our method explicitly addresses data limitations in customization, significantly enhancing multi-view coherence and realism; see Sec. D for additional completion results demonstrating visual diversity.

Spatio-temporal attention. We evaluate dense spatio-temporal attention's effectiveness for spatial consistency. As illustrated in Fig. 5b-i, we vertically shift and insert the first frame's features into subsequent frames, expecting clear semantic translations. While original AnimateDiff with 1D temporal attention fails to preserve spatial coherence due to limited pixel interactions (Fig. 5b-ii), our proposed spatio-temporal attention successfully maintains spatial consistency and semantic flow (Fig. 5b-iii). Thus, integrated spatio-temporal attention is crucial for accurately modeling large view displacements and explicitly enforcing spatial constraints, especially when employing feature replacement (Sec. 3.4).

5 Conclusion

In this work, we introduced the novel task of *multi-view customization*, integrating explicit camera viewpoint control, subject customization, and spatial consistency for both subjects and their surroundings. To address

this task, we proposed *MVCustom*, a diffusion-based framework leveraging dense spatio-temporal attention for robust multi-view synthesis. Additionally, we introduced two inference-stage strategies—*depth-aware feature rendering* and *consistent-aware latent completion*—to explicitly enforce geometric consistency and faithfully generate disoccluded regions. Extensive comparisons show that MVCustom is the only approach that effectively integrates accurate multi-view generation and high-fidelity customization. We believe this framework provides a foundation for future work on controllable and customizable multi-view generation.

Limitations and future work Our framework currently cannot alter the intrinsic object pose based on text prompts during inference (e.g., changing from sitting to standing). This limitation arises because FeatureNeRF learns a fixed canonical pose from reference images, and its radiance field does not take text prompts as input conditions. Consequently, the object's intrinsic pose remains tied to this canonical representation. Experimentally, we found that injecting the rendered feature map X_y via cross attention conditioned on textual prompts does not overcome this issue. Similar limitations related to intrinsic pose control are noted in prior work (Song et al., 2024). Future approaches might involve optimizing a dynamic neural field conditioned on textual prompts built upon a frozen static field from FeatureNeRF, using techniques such as score distillation sampling, or hypernetwork-based methods. We leave these directions for future exploration.

Additionally, another limitation arises from inaccuracies in the depth maps used in our depth-aware feature rendering. When the external depth estimator produces incorrect geometry, especially for reflective or textureless surfaces, our method directly constructs feature meshes using these inaccuracies. This limitation originates from the external depth estimator rather than our framework itself. Similar issues affect other depth-conditioned methods (Yang et al., 2025; Liu et al., 2025; Hou & Chen, 2024) due to their inherent dependence on accurate depth maps. Recent models (Yang et al., 2024; Min et al., 2025) have significantly improved depth estimation accuracy for reflective and textureless surfaces, suggesting potential mitigation of this issue. Fig. 6 demonstrates that accurate depth estimation produces realistic background geometry across multiple views: correctly estimating the depth of a textureless wall ensures the building naturally rotates with the viewpoint change. Conversely, incorrect estimation perceiving the wall as distant background results in unrealistic backgrounds across views. In conclusion, we expect that ongoing advancements in depth estimation techniques will soon overcome this limitation, enabling our framework to produce even more realistic and consistent multi-view results.

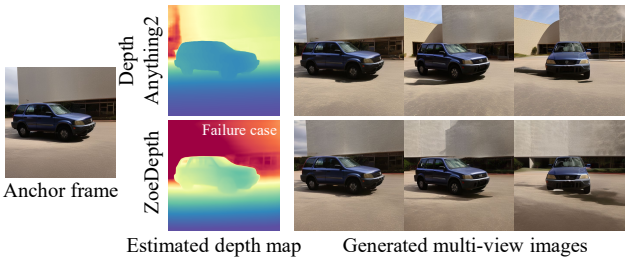

Figure 6: Comparison of background perspective alignment in generated images depending on the quality of estimated depth.

REFERENCES

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023.

Morris Alper, David Novotny, Filippos Kokkinos, Hadar Averbuch-Elor, and Tom Monnier. Wildcat3d: Appearance-aware multi-view diffusion in the wild. *arXiv preprint arXiv:2506.13030*, 2025.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Sooyeon Go, Kyungmook Choi, Minjung Shin, and Youngjung Uh. Eye-for-an-eye: Appearance transfer with semantic correspondence in diffusion models. *arXiv preprint arXiv:2406.07008*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5043–5052, 2024.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Nupur Kumari, Grace Su, Richard Zhang, Taesung Park, Eli Shechtman, and Jun-Yan Zhu. Customizing text-to-image diffusion with camera viewpoint control. *arXiv preprint arXiv:2404.12333*, 2024.

Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. *arXiv preprint arXiv:2410.05591*, 2024.

Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct consistency optimization for compositional text-to-image personalization. *arXiv preprint arXiv:2402.12004*, 2024.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024a.

Lingxiao Li, Kaixiong Gong, Weihong Li, Xili Dai, Tao Chen, Xiaojun Yuan, and Xiangyu Yue. Bifr\" ost: 3d-aware image compositing with language instructions. *arXiv preprint arXiv:2410.19079*, 2024b.

Fangfu Liu, Hanyang Wang, Weiliang Chen, Haowen Sun, and Yueqi Duan. Make-your-3d: Fast and consistent subject-driven 3d content generation. *arXiv preprint arXiv:2403.09625*, 2024.

Lijuan Liu, Wenfa Li, Dongbo Zhang, Shuo Wang, and Shaohui Jiao. Idcnet: Guided video diffusion for metric-consistent rgbd scene generation with precise camera control. *arXiv preprint arXiv:2508.04147*, 2025.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, 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.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

Junhong Min, Jimin Kim, Cheol-Hui Min, Minwook Kim, Youngpil Jeon, and Minyong Choi. Depthfocus: Controllable depth estimation for see-through scenes. *arXiv preprint arXiv:2511.16993*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6121–6132, 2025.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. *Low-rank adaptation for fast text-to-image diffusion fine-tuning*, 3, 2023.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Minjung Shin, Yunji Seo, Jeongmin Bae, Young Sun Choi, Hyunsu Kim, Hyeran Byun, and Youngjung Uh. Ballgan: 3d-aware image synthesis with a spherical background. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7268–7279, 2023.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8048–8058, 2024.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023.

Lehan Yang, Lu Qi, Xiangtai Li, Sheng Li, Varun Jampani, and Ming-Hsuan Yang. Unified dense prediction of video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28963–28973, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4578–4587, 2021.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784*, 2023.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.