AlignCVC: Aligning Cross-View Consistency for Single-Image-to-3D Generation

Xinyue Liang*, Zhiyuan Ma*, Lingchen Sun, Yanjun Guo, Lei Zhang†

Department of Computing, The Hong Kong Polytechnic University
Hong Kong, China

{xinyue.liang, zm2354.ma, ling-chen.sun, yanjunn.guo}@connect.polyu.hk, cslzhang@polyu.edu.hk

Abstract

Single-image-to-3D models typically follow a sequential generation and reconstruction workflow. However, intermediate multi-view images synthesized by pre-trained generation models often lack cross-view consistency (CVC), significantly degrading 3D reconstruction performance. While recent methods attempt to refine CVC by feeding reconstruction results back into the multi-view generator, these approaches struggle with noisy and unstable reconstruction outputs that limit effective CVC improvement. We introduce AlignCVC, a novel framework that fundamentally reframes single-image-to-3D generation through distribution alignment rather than relying on strict regression losses. Our key insight is to align both generated and reconstructed multi-view distributions toward the ground-truth multi-view distribution, establishing a principled foundation for improved CVC. Observing that generated images exhibit weak CVC while reconstructed images display strong CVC due to explicit rendering, we propose a soft-hard alignment strategy with distinct objectives for generation and reconstruction models. This approach not only enhances generation quality but also dramatically accelerates inference to as few as 4 steps. As a plug-and-play paradigm, our method, namely AlignCVC, seamlessly integrates various combinations of multiview generation models with 3D reconstruction models. Extensive experiments demonstrate the effectiveness and efficiency of AlignCVC for single-image-to-3D generation.

Code — https://github.com/Liangsanzhu/AlignCVC.git

Introduction

With advancements in 3D datasets and large-scale pre-trained 2D diffusion models, single-image-to-3D generation (Liu et al. 2023a, 2024b,a, 2023b; Long et al. 2024; Tang et al. 2025) has made significant progress in recent years. One approach leverages 3D priors from 3D datasets for 3D generation (Zhang et al. 2024b; Xiang et al. 2024; Zhao et al. 2025). However, these methods struggle with generalization and flexibility due to the limited size and diversity of existing 3D datasets (Deitke et al. 2023, 2024; Collins et al. 2022; Fu et al. 2021), which are smaller and less varied than their 2D counterparts, restricting applicability across domains. Another approach utilizes diffusion models (Rombach et al. 2022; Blattmann et al. 2023), to generate 3D content. Early works rely on Score Distillation (Poole et al. 2022; Wang et al. 2024c; Ma et al. 2025b; Yu et al. 2023) to iteratively optimize a single 3D model for a given image, which is time-consuming. Recent works employ multi-view generation (MVG) models (Liu et al. 2023a; Long et al. 2024; Wang and Shi 2023; Huang et al. 2024b) to synthesize sparse-view images, followed by feed-forward 3D reconstruction (Liu et al. 2023a; Chen et al. 2025; Hong et al. 2023; Tang et al. 2025; Xu et al. 2024) for efficiency.

However, pre-trained MVG models often struggle with cross-view consistency (CVC) when synthesizing sparse-view images, posing significant challenges for 3D reconstruction and undermining the reliability of the final 3D output. Recent works (Melas-Kyriazi et al. 2024; Chen et al. 2024; Zuo et al. 2024; Wen et al. 2024; Xue et al. 2024b; Tang et al. 2024; Xie et al. 2024a; Xue et al. 2024a) attempt to enhance CVC through 3D-aware sampling, which feeds rendered 3D reconstructions back into the generation stage. However, these methods lack effective joint optimization between generation and reconstruction models. While noisy intermediate multi-view images result in low-quality reconstructions, the integration of generation and reconstruction outputs remains fragile, as they are only optimized against fixed GT images, ignoring the inherent diversity of generation distribution under a certain input (Chen et al. 2024; Zuo et al. 2024; Wen et al. 2024; Xue et al. 2024b; Tang et al. 2024; Xue et al. 2024a). Additionally, 3D-aware sampling methods are slow (Wen et al. 2024; Zuo et al. 2024; Xue et al. 2024a,b; Tang et al. 2024), often requiring over 25 recursive steps, limiting their practicality and efficiency.

To address these, we propose relaxing the rigid constraints of input-conditioned generation and reconstruction. Unlike previous methods, where each input corresponds to fixed target multi-views, our approach aligns the generation process with a distribution derived from images rendered from 3D assets. By aligning both generation and reconstruction within a shared distribution space, we minimize discrepancies between them and enable more stable 3D-aware sampling. Specifically, we introduce a soft-hard alignment strategy: soft alignment for generation models to enhance implicit CVC, and hard alignment for reconstruction models

*These authors contributed equally.
†Corresponding author.

Figure 1: Our AlignCVC method jointly post-trains multi-view generation and reconstruction models with **distribution alignment** for 3D-aware sampling, enabling high-fidelity image-to-3D generation with only 4 diffusion steps for efficient inference.

to enforce explicit CVC. This distribution-aligned approach produces cleaner intermediate images and more stable outputs with higher CVC, outperforming 3D-aware sampling methods relying solely on regression losses with GT images.

As illustrated in Fig. 1, our proposed paradigm, namely **AlignCVC**, serves as a plug-and-play solution for **Align**ing the **CVC** for generation-reconstruction based image-to-3D pipelines. Our contributions are as follows:

- We propose a plug-and-play paradigm, AlignCVC, for single-image-to-3D generation that aligns MVG and reconstruction results with the distribution rendered from high-quality 3D assets, enabling more effective 3D-aware sampling to enhance the robustness of CVC.
- We propose a soft-hard distribution alignment strategy, assigning a soft objective to MVG models and a hard objective to reconstruction models.
- Extensive experiments show that our plug-and-play paradigm boosts CVC and enhances image-to-3D generation quality across various MVG and reconstruction models. Additionally, we cut multi-view diffusion steps to just 4, achieving much faster inference speed than other 3D-aware sampling methods.

Related Work

Score Distillation for Generation. Score Distillation Sampling (SDS)(Poole et al. 2022) transfers 2D knowledge into 3D generation by aligning 3D renderings with distributions from pre-trained diffusion models. As a base for recent advancements in 3D generation (Lin et al. 2023; Shi et al. 2023; Wang et al. 2023; Yu et al. 2023; Wang et al. 2024c; Liu et al. 2023a; Ma et al. 2025a,b), SDS leverages pre-trained diffusion models effectively. However, its dependence on per-scene optimization limits inference speed and generalizability, making it unsuitable for real-time 3D generation and driving the development of more efficient feed-forward methods.

Multi-view Generation (MVG). MVG generates multiple images simultaneously to enhance 3D generation speed and consistency. Recent models (Shi et al. 2023; Liu et al. 2023b,a; Huang et al. 2024b; Long et al. 2024; Blattmann et al. 2023) fine-tune pre-trained diffusion models on large 3D datasets (Deitke et al. 2023) for multi-view image generation. Stable Video Diffusion (Blattmann et al. 2023) supports multi-frame outputs, with which SV3D (Voleti et al. 2024) generating orbit-view videos from a single image. Integrating MVG improves both efficiency and quality in 3D pipelines. Feed-forward reconstruction models (Chen et al. 2025; Tang et al. 2025; Zhang et al. 2024a) address this by accelerating sparse-view reconstruction, forming a two-stage paradigm (see Fig. 2 (a)). However, the lack of explicit cross-view consistency (CVC) constraints in MVG-generated views often degrades reconstruction performance.

3D-Aware Sampling. 3D-aware-sampling-based methods (Chen et al. 2024; Wen et al. 2024; Xue et al. 2024a; Zuo et al. 2024; Tang et al. 2024) enhance MVG's CVC by integrating feedback from reconstructed 3D renderings and explicit CVC constraints into the generation process (see Fig. 2 (b)). While improving CVC, these methods face challenges: noisy intermediate images (see Fig. 3) degrade final 3D quality, weak joint optimization between generation and reconstruction limits performance, and multi-step recursion (25–50 iterations) hinders time efficiency.

Method

In this paper, we propose **AlignCVC**, a novel framework that enhances CVC by shifting strict regression losses with distribution alignment against GT, enabling more efficient and higher-quality 3D model generation through 3D-aware sampling, as shown in Fig. 4.

Preliminaries and Motivation

MVG Model. In this paper, we focus on diffusion-based MVG models. Specifically, the MVG takes a single view

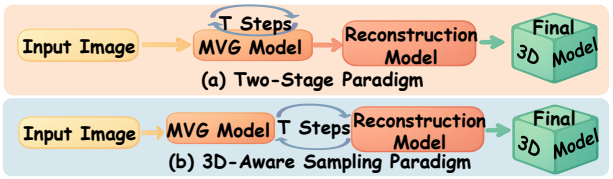

Figure 2: Two typical MVG-based Image-to-3D approaches.

$\boldsymbol{x}_c$ as the input condition and generates N views $\hat{\boldsymbol{X}}_0^\pi = \{\boldsymbol{x}^{\pi_1}, \boldsymbol{x}^{\pi_2}, \ldots, \boldsymbol{x}^{\pi_N}\}$ corresponding to camera poses $\pi = \{\pi_1, \ldots, \pi_N\}$ by initializing from Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$ and progressively denoising in K steps $t_K > t_{K-1} > \cdots > t_1$. The denoising results at the last timestep t_1 are taken as the final generation results $\hat{\boldsymbol{X}}_0^\pi$.

Reconstruction Model. Given the generated multi-views $\hat{\boldsymbol{X}}_0^\pi$, the reconstruction model produces a 3D output, differentiably rendered as $\tilde{\boldsymbol{X}}_0^\pi$ at camera poses π. This two-stage generation-reconstruction framework is shown in Fig. 2 (a). However, since most reconstruction models are trained only on GT multi-views with perfect CVC and rely heavily on CVC, such as aggregating multi-view features in 3D space, they often fail on noisy and inconsistent views from MVG, producing a low-quality 3D model misaligned with $\boldsymbol{x}_c$.

3D-Aware Sampling. To address the above issues, 3D-aware sampling refines the 3D model by reimplementing the MVG model. As shown in Fig.2(b), during progressive denoising, renderings of the reconstructed 3D outputs $\tilde{\boldsymbol{X}}_{k+1}^\pi$ are re-noised at t_k and fed back into the MVG model as the current denoising target. This introduces explicit CVC guidance from reconstruction, enhancing the MVG's CVC and improving 3D quality. Unlike prior 3D-aware sampling methods(Wen et al. 2024; Zuo et al. 2024; Xue et al. 2024a,b; Tang et al. 2024), which produce noisy and inconsistent intermediates (Fig. 3) that amplify 3D inconsistencies, we adopt a robust soft–hard alignment strategy for more stable and consistent 3D generation. This strategy maintains consistent intermediates and progressive refinement (Fig.1, Fig.3). It aligns both generation and reconstruction distributions with the ground-truth (GT) distribution: reconstructed multi-views match a clean MVG distribution, while MVG-generated multi-views align with the reconstruction model's high-CVC and generalizable domain. Alternating between these closely related distributions establishes a stable 3D-aware sampling loop, allowing each model to operate in its optimal domain.

Soft-Aligned MVG

As mentioned above, in each step k, existing 3D-aware sampling methods often produce noisy multi-view predictions $\hat{\boldsymbol{X}}_k^\pi$, which are fed into the reconstruction model (see Fig. 3). The reconstruction model then renders $\tilde{\boldsymbol{X}}_k^\pi$ as feedback for the next denoising step. This recursive feedback mechanism perpetuates noise across iterations, limiting the MVG model's ability to adapt to the directional shifts of the diffusion process and leading to a gradual degradation in the quality of 3D reconstruction. Consistent and robust guidance on the denoising direction is essential for enhancing the CVC of generated multi-view images. To address this, we propose shifting the MVG's generation target to align $\{\hat{\boldsymbol{X}}_k^\pi\}_{k=1}^K$ with a broader GT-consistent distribution, rather than being restricted to fixed GT images in the dataset. This adjustment enables the MVG model to produce clearer images in the early stages and more effectively handle reconstruction results $\{\tilde{\boldsymbol{X}}_k^\pi\}_{k=1}^K$, while maintaining flexibility and consistency with the GT distribution.

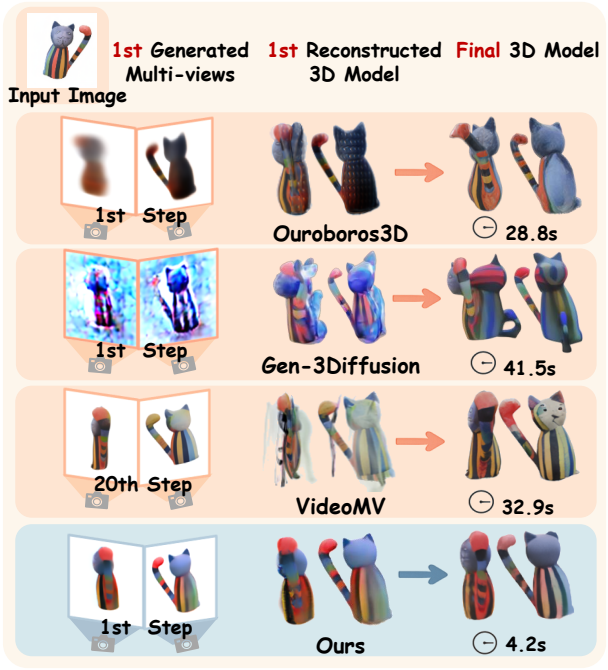

Figure 3: **The impact of CVC in 3D-aware sampling**. For Ouroboros3D, VideoMV , and Gen-3Diffusion, noise and lack of CVC affect 3D reconstructions. VideoMV applies feedback starting at the 20th step, so its results are shown after its first feedback. Our model, integrating Wonder3D and GeoLRM, delivers better results with high time efficiency.

However, directly enforcing alignment with the GT distribution imposes overly strict constraints on individual views because the MVG model inherently learns CVC implicitly within its neural representations. Furthermore, using hard alignment methods, such as adversarial generative training, makes it easier for the discriminator to identify inconsistencies between the generated views and the GT distribution with CVC. This often results in mode collapse during training, as demonstrated in our experiments, further degrading the quality of generated multi-views.

Inspired by advances in Score Distillation (Poole et al. 2022; Wang et al. 2024c; Yu et al. 2023; Ma et al. 2025b), we distill our MVG model using a pre-trained teacher that approximates the GT distribution. By leveraging the teacher's proximity to the GT and their shared distribution overlap, our model achieves **soft alignment** with the GT, which alleviates mode collapse caused by hard alignment and nat-

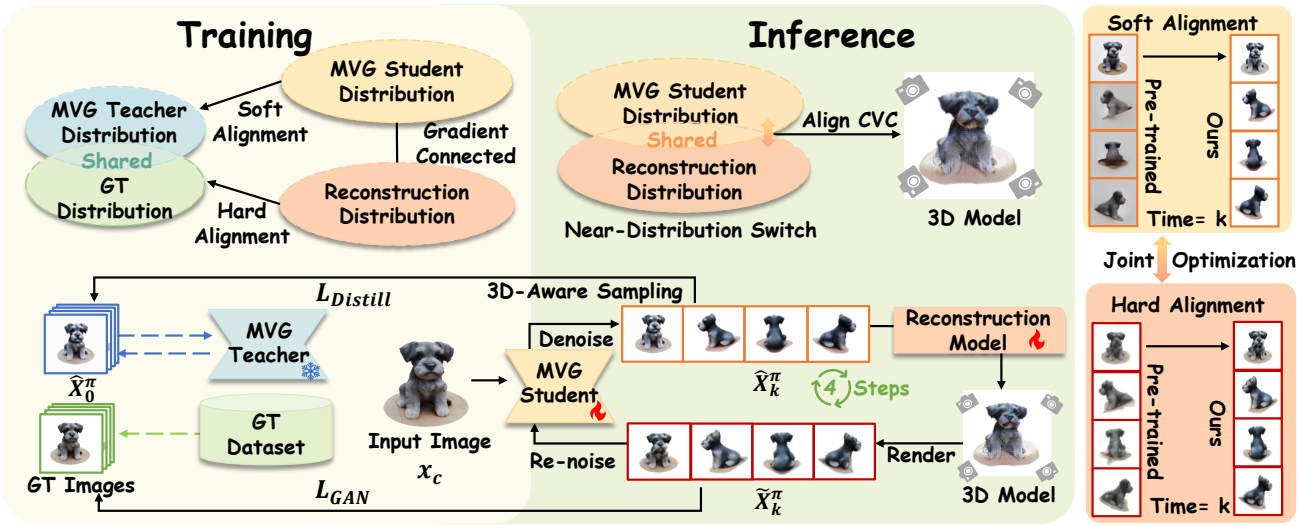

Figure 4: **The framework of AlignCVC.** During training, the multi-view generation (MVG) student model generates multi-view images $\hat{\boldsymbol{X}}_k^\pi$ from input image $\boldsymbol{x}_c$ at camera poses π. A pre-trained MVG teacher model aligns $\hat{\boldsymbol{X}}_k^\pi$ with the GT distribution via a soft-alignment method. We then obtain the 3D model from the reconstruction model and adversarially supervise its renderings $\tilde{\boldsymbol{X}}_k^\pi$ to the GT distribution in a hard-aligned manner. In the inference phase, we reconstruct an intermediate 3D model with the generated multi-view images $\hat{\boldsymbol{X}}_k^\pi$ at each timestep, where the renderings $\tilde{\boldsymbol{X}}_k^\pi$ are then re-noised for the next denoising timestep with 3D-aware sampling. This recursive sampling, repeated for 4 steps, produces the final 3D model.

urally aligns with diffusion training dynamics. Specifically, Score Distillation is employed to align the $\{\hat{\boldsymbol{X}}_k^\pi\}_{k=1}^K$ generated by the trainable MVG student with those predicted by the teacher. This alignment reduces noise in intermediate denoising stages, improving the stability of the generation process. Furthermore, aligning with the GT distribution, rather than strictly to the target GT images, enhances the model's generalization ability, particularly in adapting sampling guidance during 3D-aware sampling.

Among recent distillation methods (Ma et al. 2025b; Wang et al. 2024c; Poole et al. 2022; Yu et al. 2023; Wang et al. 2023; Liang et al. 2023; Wang et al. 2024a), we adopt Asynchronous Score Distillation (ASD) (Ma et al. 2025b) as the distribution alignment objective due to its ability to produce stable gradients when training deep generative models. In this framework, the MVG teacher predicts the noise residual in $\hat{\boldsymbol{X}}_{k,t}^\pi$, which represents the $\hat{\boldsymbol{X}}_k^\pi$ generated by the MVG student diffused with Gaussian noise ϵ at timestep $t \in \{1, \ldots, 1000\}$ (Ho, Jain, and Abbeel 2020). Let θ and ϕ denote the parameters of the MVG student and the MVG teacher, respectively. In ASD, given the input image $\boldsymbol{x}_c$ and camera parameters π as conditions, and the k-step multi-view predictions $\hat{\boldsymbol{X}}_k^\pi$, the training gradient with respect to θ can be formulated as follows:

$$\nabla_\theta \mathcal{L}_{\text{Distill}}(\hat{\boldsymbol{X}}_k^\pi) \triangleq \mathbb{E}_{t,\Delta t,\epsilon}\Big[\omega(t)\Big(\epsilon_\phi(\hat{X}_{k,t}; t, x_c, \pi) - \epsilon_\phi(\hat{X}_{k,t+\Delta t}; t+\Delta t, x_c, \pi)\Big)\frac{\partial \hat{X}_k}{\partial \theta}\Big], \tag{1}$$

where $\omega(t)$ is a timestep-dependent weighting factor, and Δt is a timestep shift proposed in ASD. With the soft-aligned MVG, we achieve high-quality denoised results even in early diffusion steps, producing clearer multi-views with higher CVC. To this objective, we use pre-trained MVG teachers (Long et al. 2024; Voleti et al. 2024) trained with multi-view renderings from the 3D object dataset (Deitke et al. 2023) to ensure consistency between the distributions of generated and GT images. Instead of training from scratch, the student is initialized with the pre-trained teacher and fine-tuned by adding LoRA (Hu et al. 2021) layers.

With the soft-alignment strategy, MVG student achieves clear multi-view images even at the first denoising step, ensuring outputs align closely with the GT and enhancing the generalizability of sampling guidance. Inspired by SD-Turbo (Sauer et al. 2024), we limit the inference steps to as few as $K = 4$, striking a trade-off between improving generation speed, avoiding excessive latent averaging (Huang et al. 2024a), and enabling effective 3D-aware sampling.

Hard-Aligned Reconstruction

Supervising the reconstruction model directly with GT multi-view images fails to resolve the distributional gap caused by inconsistent MVG-generated inputs, which often lack CVC. Our experiment also shows that such supervision severely degrades the model's reconstruction ability, as the inconsistent inputs hinder effective multi-view aggregation. Forcing the model to strictly match GT views under these conditions leads to blurry 3D reconstructions.

To address this, a distribution-based training objective for plausible reconstruction is necessary to ensure CVC across stages. Aligning with the GT distribution ensures the reconstructed model achieves consistency in both geometry and texture. While Score Distillation could theoretically align

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP Sim. ↑	FID ↓	CVC ↑	Time (s) ↓
LGM† (Tang et al. 2025)	18.010	0.893	0.145	0.859	113.945	4.316	**2.96**
Trellis (Xiang et al. 2024)	15.106	0.850	0.209	0.834	115.413	-	12.69
SyncDreamer (Liu et al. 2023b)	18.423	0.885	0.153	0.787	270.902	**5.668**	81.12
VideoMV (Zuo et al. 2024)	18.042	0.802	0.151	0.825	127.095	5.363	32.93
Ouroboros3D (Wen et al. 2024)	**21.069**	**0.902**	**0.117**	**0.887**	**102.379**	5.503	28.82
Gen-3Diffusion (Xue et al. 2024a)	**21.561**	**0.904**	**0.115**	**0.880**	**110.796**	5.023	41.52
Wonder3D (Long et al. 2024)+LGM	16.893	0.717	0.220	0.849	189.967	4.702	12.12
AlignCVC (Ours)	**21.977**	**0.912**	**0.104**	**0.899**	**101.506**	**5.808**	**8.87**
Wonder3D+GeoLRM (Zhang et al. 2024a)	18.421	0.828	0.142	0.852	150.949	4.714	12.21
AlignCVC (Ours)	19.693	0.860	0.128	0.881	128.379	**5.619**	**4.17**
Wonder3D+LaRa (Chen et al. 2025)	16.490	0.719	0.250	0.823	220.770	4.671	5.01
AlignCVC (Ours)	19.210	0.832	0.159	0.844	158.699	5.010	4.76
SV3D (Voleti et al. 2024)+LGM	18.963	0.837	0.148	0.867	162.327	4.974	58.59
AlignCVC (Ours)	20.483	0.892	0.120	0.876	117.220	5.522	17.48
SV3D+GeoLRM	17.872	0.816	0.142	0.836	150.977	4.873	55.17
AlignCVC (Ours)	19.322	0.874	0.131	0.853	125.381	5.094	17.04
SV3D+LaRa	16.232	0.701	0.277	0.772	218.200	4.989	49.68
AlignCVC (Ours)	18.270	0.847	0.142	0.841	127.942	4.995	16.87

Table 1: **Comparison on single-image-to-3D generation.** The top three results for each metric are highlighted in **red**, **blue**, and **green**, respectively. LGM† refers to the two-stage generation approach proposed in the original paper, which employs the pre-trained ImageDream (Wang and Shi 2023) model as the MVG, followed by the LGM reconstruction model.

the distribution of $\{\tilde{\boldsymbol{X}}_k^\pi\}_{k=1}^K$, the rendering process inherently enforces CVC, making a **hard alignment** strategy more effective. Specifically, adversarial generative learning aligns the reconstruction outputs directly with the GT distribution derived from renderings. This explicit supervision guarantees high-quality 3D models with robust CVC.

We use the following adversarial objective to align the reconstruction results $\tilde{\boldsymbol{X}}_k^\pi$ with the target distribution, providing a higher upper bound for alignment and enhancing robustness beyond strict regression losses:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{X_{gt}}[\log D(X_{gt}^\pi)] + \mathbb{E}_{\tilde{X}_k}[\log(1 - D(\tilde{X}_k^\pi))], \quad (2)$$

where X_{gt}^π is the GT multi-views. We assist it with a reconstruction loss for better performance:

$$\mathcal{L}_{\text{Recon}} = \|X_{gt}^\pi - \tilde{X}_k^\pi\|_2^2. \quad (3)$$

In addition, since the reconstruction process uses multi-view images generated by the MVG as input, preserving the gradient connection between the two models facilitates more coordinated training. As a result, the objectives not only guide the MVG student to generate $\{\hat{\boldsymbol{X}}_k^\pi\}_{k=1}^K$, improving the alignment of $\{\tilde{\boldsymbol{X}}_k^\pi\}_{k=1}^K$ with the GT distribution, but also directly supervise the final 3D model for quality. This process further enhances the reconstruction model through Score Distillation from the MVG.

Experiments

Experimental Setup

Training Settings. We train our AlignCVC framework on the Gobjaverse dataset (Qiu et al. 2024), which comprises multi-view renderings of 280K 3D objects sourced from Objaverse (Deitke et al. 2023). The resolution of the final results is set to 512×512 pixels for both training and testing. Our proposed method is trained for 300,000 iterations on an H20 GPU, utilizing the Adan (Xie et al. 2024b) optimizer with a learning rate of 2×10^{-5}. All computations are performed in 32-bit precision and implemented using the ThreeStudio (Guo et al. 2023) framework.

Testing Settings. Following prior works, we evaluate our method with GSO (Downs et al. 2022) dataset. Tests are conducted on four orthogonal azimuths, and the evaluation metrics include PSNR, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018), CLIP Similarity (Radford et al. 2021), and FID (Kynkäänniemi et al. 2022). While these metrics primarily assess reconstruction quality against a fixed set of GT images, they inherently bias against models capable of generating novel yet plausible 3D-consistent content deviating from predefined GT. To address this limitation, we further evaluate the CVC of MVG outputs using the model introduced by DUSt3R (Wang et al. 2024b), which predicts corresponding point maps between adjacent images. The CVC score is computed as the average confidence of the top 100 matched points based on their confidence scores.

Baseline MVG and Reconstruction Models. We prioritize MVG models that ensure camera control and generate high-fidelity multi-views for downstream tasks. Wonder3D (Long et al. 2024) and SV3D (Voleti et al. 2024) are selected as MVG teachers for their ability to adjust camera parameters. Both are fine-tuned on the Gobjaverse dataset to better control elevation and restrict generated frames to orthogonal views. For reconstruction, we adopt LGM (Tang et al. 2025) for its end-to-end prediction of 3DGS parameters from multi-views, along with GeoLRM (Zhang et al. 2024a) and LaRa (Chen et al. 2025), which project 2D features into 3D volumes to construct 3D models.

Results

Quantitative Comparison. The quantitative results of competing methods are shown in Table 1. For two-stage image-

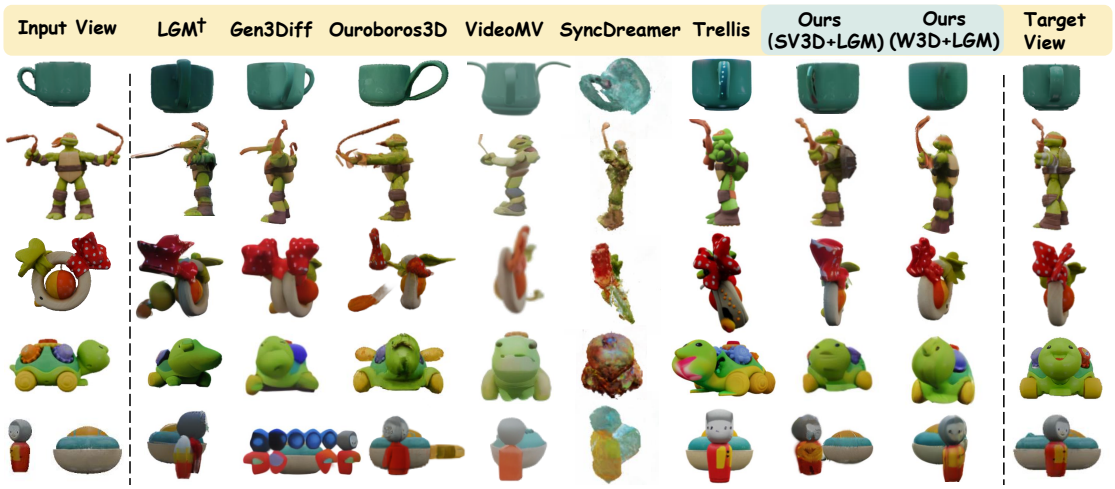

Figure 5: Comparison results on image-to-3D generation. Gen3Diff is short for Gen-3Diffusion.

to-3D generation, we combine pre-trained generation and reconstruction models, where multi-view images are first generated and then reconstructed into a 3D model. AlignCVC, as a plug-and-play paradigm, outperforms these combinations across all 3D metrics. Distribution alignment further enhances CVC performance, improving 3D model quality. Table 1 highlights the importance of pairing compatible generation and reconstruction models. For 3D generation, Wonder3D, optimized for orthogonal images, surpasses SV3D with better camera control. For reconstruction, LGM outperforms GeoLRM and LaRa by directly predicting 3DGS parameters, avoiding errors from projecting incomplete 2D images into 3D. By leveraging neural networks to aggregate multi-view features, LGM better adapts to distribution alignment, achieving significant gains.

Compared with existing methods in the upper panel of Table 1, AlignCVC integrating Wonder3D and LGM surpasses all SOTA methods. While Gen-3Diffusion demonstrates competitive generation quality, it requires significantly more time. On the other hand, the two-stage method LGM† is the fastest but suffers from poor reconstruction metrics and CVC, prioritizing speed over quality. In contrast, AlignCVC achieves both high quality and fast inference speed, outperforming others in efficiency when using Wonder3D and LGM as baseline models. In terms of the CVC metric, excluding Trellis, which generates 3D models without relying on MVG, LGM achieves the lowest scores due to the absence of 3D-aware sampling. SyncDreamer achieves the highest score by leveraging 3D features from multi-view images during sampling. Incorporating AlignCVC significantly enhances CVC performance, validating the effectiveness of our approach.

Qualitative Comparison. Fig. 5 provides visual examples highlighting the superior CVC performance of our method compared to competing approaches, which exhibit inconsistencies and artifacts. The two-stage method, LGM†, struggles with cross-view inconsistencies, leading to incorrect geometry and misaligned rotations. 3D-aware sampling methods like Gen-3Diffusion and Ouroboros3D suffer from noise propagation, producing redundant artifacts. Similarly, VideoMV and SyncDreamer, constrained by low resolution, display significant blurriness. Trellis, a 3D-native generator, preserves geometry well but has limited generalization to unseen data, resulting in deviations in color, semantics, and pose, ultimately lowering reconstruction metrics. Further results and discussions are in the supplementary materials.

Ablation Study

In the ablation studies, we employ Wonder3D as the MVG model and LGM as the reconstruction model.

The necessity of tuning both models. We evaluate the **Two-stage** paradigm with pre-trained models as the baseline. We further train the reconstruction model to adapt to inconsistent multi-view images (serving as data augmentation) in the two-stage process, but this leads to blurrier reconstruction outputs. Additionally, directly applying 3D-aware sampling also fails to yield significant improvements. As shown in Fig. 6, these settings produce results inconsistent with the input, while 3D-aware sampling even degrades metrics like PSNR and FID. This performance drop stems from the misalignment between the intermediate multi-views generated by the pre-trained MVG and the distributions of both the MVG teacher and ground truth (GT), emphasizing the need for distribution alignment. Experiments with **Fixed MVG** and **Fixed Recon.** further demonstrate that aligning either the reconstruction or MVG model significantly improves 3D-aware sampling performance.

The necessity of distribution alignment. With 3D-aware sampling and tuning of both MVG and reconstruction models, we investigate the impact of different distribution alignment objectives. With only the regression loss $\mathcal{L}_{\text{Recon}}$, as shown in **w/o Distribution Align** in Table 2, the improvement is limited. Building on this, applying alignment to either the MVG or the reconstruction model leads to better performance, as demonstrated in **w/o Recon. Align** and **w/o MVG Align**. However, when distribution alignment for the MVG model is removed, as shown in **w/o MVG Align**, the

Settings	Loss $\mathcal{L}_{\text{Distill}}$	Loss $\mathcal{L}_{\text{GAN}}$	Loss $\mathcal{L}_{\text{Recon}}$	PSNR ↑	SSIM ↑	LPIPS ↓	CVC ↑	FID ↓
Two-stage (Pre-trained Models)	✗	✗	✗	16.89	0.72	0.22	4.70	189.97
Two-stage (Train Recon.)	✗	✗	✓	18.24 (+1.35)	0.66	0.16 (-0.06)	4.70	194.44
+ 3D-Aware Sampling								
Pre-trained Models	✗	✗	✗	15.92	0.78 (+0.06)	0.19 (-0.03)	4.49	249.05
Fixed MVG	✗	✓	✓	17.40 (+0.51)	0.78 (+0.06)	0.19 (-0.03)	4.71 (+0.01)	175.12 (-14.85)
Fixed Recon.	✓	✓	✓	19.05 (+2.16)	0.86 (+0.14)	0.12 (-0.10)	5.12 (+0.42)	142.16 (-47.81)
+ Joint Training of Both MVG & Recon.								
Ours	✓	✓	✓	**21.98** (+5.09)	**0.91** (+0.19)	**0.10** (-0.12)	**5.81** (+1.11)	**101.51** (-88.46)
w/o $\mathcal{L}_{\text{Recon}}$	✓	✓	✗	21.04 (+4.15)	0.89 (+0.17)	0.12 (-0.10)	5.67 (+0.97)	128.09 (-61.88)
w/o MVG Align	✗	✓	✓	18.28 (+1.39)	0.84 (+0.12)	0.14 (-0.08)	5.02 (+0.32)	150.52 (-39.45)
w/o Recon. Align	✓	✗	✓	20.84 (+3.95)	0.88 (+0.16)	0.12 (-0.10)	5.31 (+0.61)	124.38 (-75.59)
w/o Distribution Align	✗	✗	✓	18.16 (+1.27)	0.83 (+0.11)	0.16 (-0.06)	4.51	187.05 (-2.92)
Impact of Alignment Strategy								
Hard-aligned MVG	$\mathcal{L}_{\text{GAN}}^{\text{MVG}}$ ✓		✓	17.31 (+0.42)	0.82 (+0.10)	0.16 (-0.06)	4.36	187.05 (-2.92)
Soft-aligned Recon.	✓	$\mathcal{L}_{\text{Distill}}^{\text{Recon.}}$	✓	16.69	0.80 (+0.08)	0.18 (-0.04)	3.97	191.17
Impact of Replacing Distillation Loss with Diffusion Loss								
$\mathcal{L}_{\text{Distill}}^{\text{MVG}} \rightarrow \mathcal{L}_{\text{Diff}}^{\text{MVG}}$	$\mathcal{L}_{\text{Diff}}^{\text{MVG}}$ ✓		✓	16.39	0.69	0.15 (-0.07)	4.22	171.63 (-18.34)
+ w/o Recon. Align	$\mathcal{L}_{\text{Diff}}^{\text{MVG}}$ ✗		✓	16.82	0.71	0.15 (-0.07)	4.38	201.81

Table 2: **Ablation studies.** Improvements over the pre-trained two-stage baseline (first row) are indicated in red.

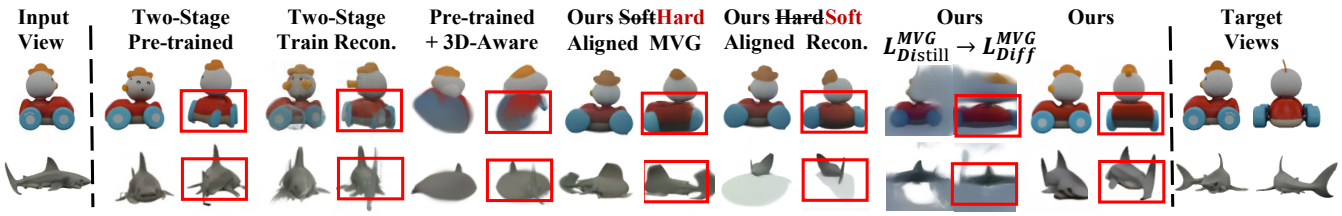

Figure 6: Ablation results of image-to-3D generation, evaluating the effect of different alignment strategies.

potential improvement is significantly hindered, indicating that MVG plays a more critical role in the sampling loop. These results underscore the importance of effective MVG alignment objectives, which are essential for achieving optimal performance in 3D-aware sampling.

The necessity of soft alignment for MVG model. We use Score Distillation as the training objective for the MVG model, aligning its intermediate outputs with the distribution modeled by the MVG teacher. While this teacher distribution is derived from but not identical to the GT, directly aligning with the GT is another potential approach. However, combining hard alignment in the reconstruction model with adversarial training for the MVG model results in training collapse and performance degradation, as shown in **Hard-aligned MVG** in Table 2 and Fig. 6. This is because the multi-view discriminator, relying on the CVC, easily distinguishes between generated and GT images, leading to unbalanced adversarial training and hindering the MVG model from effectively learning. We also replace the distillation loss with diffusion loss (Rombach et al. 2022), as shown in **Ours** $\mathcal{L}_{\text{Distill}}^{\text{MVG}} \rightarrow \mathcal{L}_{\text{Diff}}^{\text{MVG}}$ in Table 2 and Fig. 6, following the MVG training objective in Gen-3Diffusion. However, diffusion loss introduces significant noise to the rendered results, as observed in some Gen-3Diffusion cases, severely degrading the final 3D model quality. In this setting, reconstruction model alignment has minimal impact.

The necessity of hard alignment for reconstruction model. Score Distillation can lead to deteriorated performance, as shown in the rows of **Soft-aligned Recon.** in Table 2, and the inconsistent output, such as the shark in Fig. 6. It reveals that with explicit CVC, the reconstructed multi-views are qualified for directly aligning with GT distributions. Ablation with **w/o** $\mathcal{L}_{\text{Recon}}$ also shows that using the regression loss provides further improvement, because the reconstruction model is initially trained with this objective.

Conclusion

We presented AlignCVC, a plug-and-play framework to enhance cross-view consistency (CVC) of single-image-to-3D generation by jointly post-training multi-view generation (MVG) and reconstruction models. AlignCVC addresses distribution misalignment, the key bottleneck in 3D-aware sampling, by aligning generated and reconstructed multi-view distributions with the GT distribution. Recognizing that CVC is implicit in MVG models but explicit in reconstruction models, we adopt a soft-hard alignment strategy: Score Distillation for MVG (soft) and adversarial training for reconstruction (hard). Experiments show AlignCVC improves performance across various MVG-reconstruction combinations, reducing 3D-aware sampling to 4 steps.

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Chen, A.; Xu, H.; Esposito, S.; Tang, S.; and Geiger, A. 2025. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision*, 338–355. Springer.

Chen, H.; Shen, B.; Liu, Y.; Shi, R.; Zhou, L.; Lin, C. Z.; Gu, J.; Su, H.; Wetzstein, G.; and Guibas, L. 2024. 3D-Adapter: Geometry-Consistent Multi-View Diffusion for High-Quality 3D Generation. *arXiv preprint arXiv:2410.18974*.

Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Vicente, T. F. Y.; Dideriksen, T.; Arora, H.; et al. 2022. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21126–21136.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Downs, L.; Francis, A.; Koenig, N.; Kinman, B.; Hickman, R.; Reymann, K.; McHugh, T. B.; and Vanhoucke, V. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, 2553–2560. IEEE.

Fu, H.; Jia, R.; Gao, L.; Gong, M.; Zhao, B.; Maybank, S.; and Tao, D. 2021. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129: 3313–3337.

Guo, Y.-C.; Liu, Y.-T.; Shao, R.; Laforte, C.; Voleti, V.; Luo, G.; Chen, C.-H.; Zou, Z.-X.; Wang, C.; Cao, Y.-P.; and Zhang, S.-H. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Z.; Guo, Y.-C.; Wang, H.; Yi, R.; Ma, L.; Cao, Y.-P.; and Sheng, L. 2024a. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*.

Huang, Z.; Wen, H.; Dong, J.; Wang, Y.; Li, Y.; Chen, X.; Cao, Y.-P.; Liang, D.; Qiao, Y.; Dai, B.; et al. 2024b. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9784–9794.

Kynkäänniemi, T.; Karras, T.; Aittala, M.; Aila, T.; and Lehtinen, J. 2022. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*.

Liang, Y.; Yang, X.; Lin, J.; Li, H.; Xu, X.; and Chen, Y. 2023. LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching. arXiv:2311.11284.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Liu, M.; Shi, R.; Chen, L.; Zhang, Z.; Xu, C.; Wei, X.; Chen, H.; Zeng, C.; Gu, J.; and Su, H. 2024a. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10072–10083.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2024b. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023a. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023b. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9970–9980.

Ma, Z.; Liang, X.; Wu, R.; Zhu, X.; Lei, Z.; and Zhang, L. 2025a. Progressive Rendering Distillation: Adapting Stable Diffusion for Instant Text-to-Mesh Generation without 3D Data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11036–11050.

Ma, Z.; Wei, Y.; Zhang, Y.; Zhu, X.; Lei, Z.; and Zhang, L. 2025b. Scaledreamer: Scalable text-to-3d synthesis with asynchronous score distillation. In *European Conference on Computer Vision*, 1–19. Springer.

Melas-Kyriazi, L.; Laina, I.; Rupprecht, C.; Neverova, N.; Vedaldi, A.; Gafni, O.; and Kokkinos, F. 2024. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qiu, L.; Chen, G.; Gu, X.; Zuo, Q.; Xu, M.; Wu, Y.; Yuan, W.; Dong, Z.; Bo, L.; and Han, X. 2024. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9914–9925.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2024. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 87–103. Springer.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2025. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, 1–18. Springer.

Tang, Z.; Zhang, J.; Cheng, X.; Yu, W.; Feng, C.; Pang, Y.; Lin, B.; and Yuan, L. 2024. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. *arXiv preprint arXiv:2407.19548*.

Voleti, V.; Yao, C.-H.; Boss, M.; Letts, A.; Pankratz, D.; Tochilkin, D.; Laforte, C.; Rombach, R.; and Jampani, V. 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, 439–457. Springer.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12619–12629.

Wang, P.; and Shi, Y. 2023. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*.

Wang, P.; Xu, D.; Fan, Z.; Wang, D.; Mohan, S.; Iandola, F.; Ranjan, R.; Li, Y.; Liu, Q.; Wang, Z.; and Chandra, V. 2024a. Taming Mode Collapse in Score Distillation for Text-to-3D Generation. *arXiv preprint: 2401.00909*.

Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024b. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024c. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Wen, H.; Huang, Z.; Wang, Y.; Chen, X.; Qiao, Y.; and Sheng, L. 2024. Ouroboros3D: Image-to-3D Generation via 3D-aware Recursive Diffusion. *arXiv preprint arXiv:2406.03184*.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. *arXiv preprint arXiv:2412.01506*.

Xie, D.; Li, J.; Tan, H.; Sun, X.; Shu, Z.; Zhou, Y.; Bi, S.; Pirk, S.; and Kaufman, A. E. 2024a. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6369–6379.

Xie, X.; Zhou, P.; Li, H.; Lin, Z.; and Yan, S. 2024b. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, Y.; Shi, Z.; Yifan, W.; Chen, H.; Yang, C.; Peng, S.; Shen, Y.; and Wetzstein, G. 2024. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*.

Xue, Y.; Xie, X.; Marin, R.; and Pons-Moll, G. 2024a. Gen-3Diffusion: Realistic Image-to-3D Generation via 2D & 3D Diffusion Synergy. *arXiv preprint arXiv:2412.06698*.

Xue, Y.; Xie, X.; Marin, R.; and Pons-Moll, G. 2024b. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. *arXiv preprint arXiv:2406.08475*.

Yu, X.; Guo, Y.-C.; Li, Y.; Liang, D.; Zhang, S.-H.; and Qi, X. 2023. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*.

Zhang, C.; Song, H.; Wei, Y.; Chen, Y.; Lu, J.; and Tang, Y. 2024a. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation. *arXiv preprint arXiv:2406.15333*.

Zhang, L.; Wang, Z.; Zhang, Q.; Qiu, Q.; Pang, A.; Jiang, H.; Yang, W.; Xu, L.; and Yu, J. 2024b. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics (TOG)*, 43(4): 1–20.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, Z.; Lai, Z.; Lin, Q.; Zhao, Y.; Liu, H.; Yang, S.; Feng, Y.; Yang, M.; Zhang, S.; Yang, X.; et al. 2025. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. *arXiv preprint arXiv:2501.12202*.

Zuo, Q.; Gu, X.; Qiu, L.; Dong, Y.; Zhao, Z.; Yuan, W.; Peng, R.; Zhu, S.; Dong, Z.; Bo, L.; et al. 2024. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*.