RIGI: Rectifying Image-to-3D Generation Inconsistency via Uncertainty-aware Learning

Jiacheng Wang, Zhedong Zheng, *Senior Member, IEEE*, Wei Xu†, Ping Liu†, *Senior Member, IEEE*

***Abstract*—Image-to-3D generation aims to predict a geometrically and perceptually plausible 3D model from a single 2D image. Conventional approaches typically follow a cascaded pipeline: initially generating multi-view projections from the single input image through view synthesis, followed by optimizing 3D geometry and appearance strictly using these projections. However, such deterministic optimization neglects epistemic uncertainty from imperfectly generated data, particularly due to limited observations and inconsistent content. To address this issue, we propose an uncertainty-aware optimization framework that explicitly models and mitigates epistemic uncertainty, leading to more robust and reliable 3D generation. For epistemic uncertainty arising from incomplete viewpoint coverage, we employ a progressive sampling strategy that sinusoidally varies camera elevations and progressively integrates diverse viewpoints into training, enhancing viewpoint coverage and stabilizing optimization. For epistemic uncertainty caused by the deterministic optimization on the noisy and inconsistent generated multi-view frames, we estimate an uncertainty map from the discrepancies between two independently optimized Gaussian models. This map is incorporated into uncertainty-aware regularization, adaptively adjusting loss weights to suppress unreliable supervision. Furthermore, we provide a theoretical analysis of uncertainty-aware optimization by deriving a probabilistic upper bound on the expected generation error, providing insights into its effectiveness. Extensive experiments demonstrate that our method significantly reduces artifacts and inconsistencies, leading to higher-quality 3D generation. More visual results are available at our website.**

***Index Terms*—Image to 3D Generation, Uncertainty-aware Learning, 3D Gaussian Splatting.**

I. Introduction

IMAGE-to-3D generation is to synthesize 3D objects with both geometric structures and realistic textures from a single-view image, significantly reducing manual modeling costs and accelerating 3D creation. This capability is particularly valuable in industries requiring large-scale 3D asset generation, including video game development, film production, and virtual/augmented reality. Despite recent advances [1]–[6], generating high-fidelity 3D assets remains a major challenge due to the inherent complexities of spatial structure, occlusions, and texture consistency.

J. Wang and W. Xu are with the Hubei Key Laboratory of Smart Internet Technology, School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China, e-mail: jiacheng@hust.edu.cn, xuwei@hust.edu.cn.

Z. Zheng is with FST and ICI, University of Macau, Macau 999078, China, email: zdzheng12@gmail.com.

P. Liu is with CSE, University of Nevada, Reno, NV 89557, USA, e-mail: pino.pingliu@gmail.com.

† denotes co-corresponding author.

Early approaches [7]–[11], given a single RGB image, could successfully craft a 3D model of specific object categories. To further improve scalability, some methods [12]–[16] have introduced image-based 3D generative models with diverse 3D supervision. Large reconstruction models (LRMs) [17]–[19] have also been developed to efficiently map image features into 3D triplane space. To further improve geometry quality, modern models [20]–[24] mainly use structured latent representations and 3D-aware transformers to fully capture the geometry and texture of 3D assets. However, these models typically require high-quality, large-scale 3D datasets, which are expensive and difficult to acquire. Recent research has thus explored leveraging large-scale 2D diffusion models [25]–[27] for 3D synthesis, including methods such as DreamFusion [1] and Zero123 [2]. Although these approaches have shown generalization capabilities, they inherently lack 3D awareness and lead to geometric inconsistencies.

To improve multi-view consistency, recent methods [28]–[35] enhance diffusion models with global self-attention, while others [36]–[42] leverage video diffusion to improve spatio-temporal coherence. Although these approaches produce dense and high-resolution frames for 3D optimization, they often overlook epistemic uncertainty—variations in the reconstructed 3D assets that persist even when training on the same pseudo-labels. We attribute these variations to two primary sources: limited observations and inconsistent content. Limited observations stem from predefined camera poses with restricted coverage, leading to under-reconstruction and missing details in unobserved regions, as illustrated in Figure 1(a). In contrast, inconsistent content arises from deterministic optimization over pseudo labels containing geometric and textural discrepancies across overlapping views, which results in artifacts and distortions, as shown in Figure 1(b). These issues are further exacerbated by the noisy and incomplete nature of generated pseudo-labels, which impose ambiguous or conflicting supervision in regions with insufficient coverage or inconsistency. Together, they manifest as high epistemic uncertainty in the final 3D reconstruction, motivating the need for an uncertainty-aware optimization framework that explicitly accounts for these failure modes.

In this paper, we introduce an uncertainty-aware optimization framework, structured as a two-stage pipeline to establish a strong baseline for evaluating uncertainty-aware learning. Specifically, we first employ SV3D [39] to generate multi-view frames as pseudo-labels, and then apply 3D Gaussian Splatting [43] for efficient optimization and high-

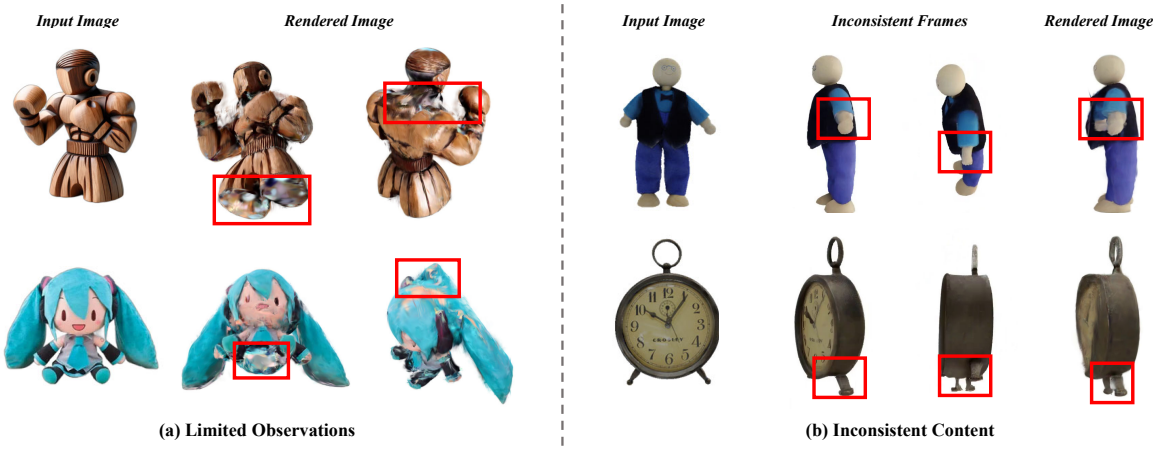


Fig. 1. **Illustration of two sources of epistemic uncertainty.** We categorize the sources of epistemic uncertainty in 3D asset optimization into two main types, with red bounding boxes highlighting regions of suboptimal generation and degraded geometry. (a) *Limited observations*: caused by incomplete viewpoint coverage, leading to under-reconstruction, especially in top and bottom views. (b) *Inconsistent content*: caused by geometric and textural inconsistencies across overlapping views, resulting in artifacts and distortions in the final 3D reconstructions.

quality rendering. To address epistemic uncertainty arising from limited and noisy pseudo-labels, we propose two targeted strategies. First, to mitigate uncertainty due to insufficient observation, we adopt a progressive sampling strategy that generates multi-view frames with sinusoidal elevation variations and progressively incorporates pseudo-labels during training. We sample camera elevations from a sine function to ensure a balanced distribution of viewpoints, especially improving coverage of top and bottom views. Moreover, we prioritize the integration of high-confidence pseudo-labels across multiple feed-forward passes, which helps stabilize training and reduce geometric distortions. Second, to address epistemic uncertainty caused by pseudo-label inconsistency, we introduce an uncertainty-aware learning mechanism. It estimates per-pixel uncertainty from the discrepancies between two independently optimized Gaussian models and integrates this information into the optimization process. By adaptively adjusting pixel-wise loss weights, the model downweights unreliable supervision, thereby improving stability and enhancing 3D generation quality. Notably, our framework operates at the optimization stage and does not depend on a specific multi-view generator, offering potential applicability to other multi-view-based image-to-3D pipelines. Extensive experiments demonstrate that our approach reduces artifacts, improves geometric and texture consistency, and achieves superior overall reconstruction performance.

In summary, our main contributions are as follows:

- We propose an uncertainty-aware optimization framework for image-to-3D generation, explicitly addressing epistemic uncertainty caused by limited observations and pseudo-label inconsistencies. We introduce a progressive sampling strategy to increase viewpoint diversity and an uncertainty regularization method based on discrepancies between two Gaussian models. This unified design improves the robustness and accuracy of 3D reconstruction.
- We conduct extensive experiments on the Google Scanned Objects (GSO) dataset [44], including ablation studies and user evaluations. Our results show consistent improvements over state-of-the-art baselines in both quantitative metrics and qualitative assessments, demonstrating enhanced geometric and texture consistency, as well as effective suppression of boundary artifacts.
- We further provide a theoretical analysis of our approach, deriving a probabilistic upper bound on the reconstruction error, which offers insight into how uncertainty modeling improves optimization stability.

II. Related Work

In this section, we provide an overview of related works on image-to-3D generation and uncertainty-aware learning. Section II-A discusses various approaches to image-to-3D generation, including 3D generative models, 2D-to-3D lifting methods, and large-scale reconstruction frameworks. Section II-B explores different types of uncertainty and recent advancements in uncertainty-aware learning, with a particular emphasis on its applications in 3D domain.

A. Image-to-3D Generation

Image-to-3D generation aims to synthesize high-fidelity 3D assets from a single 2D image, a challenging task due to the need for reliable modeling of unseen views. Early approaches primarily focus on single-view 3D reconstruction [7]–[11], typically relying on explicit 3D representations such as meshes or voxel grids. More recently, the field has shifted towards image-based 3D generative models, which leverage diverse 3D representations [12]–[16], enabling the synthesis of more expressive and structurally complex 3D assets.

A significant limitation of traditional 3D generation methods is their reliance on high-quality paired 3D data, which restricts scalability and generalization. Inspired by the success of large-scale 2D diffusion models [25]–[27] in image and video generation, recent efforts have explored 2D-to-3D lifting approaches. For instance, DreamFusion [1] introduces score distillation sampling (SDS), which optimizes 3D structures by leveraging the prior knowledge of pre-trained 2D diffusion models, showing strong zero-shot text-to-3D generation capabilities. Furthermore, Yi *et al.* [45] propose a progressive learning strategy to gradually increase the resolution to

facilitate the optimization, while Zero123 [2] fine-tunes a text-to-image model using 3D data, enabling novel view synthesis by conditioning on relative camera poses. Building upon this foundation, subsequent works [3], [4], [6] have sought to improve multi-view consistency, generating more coherent and visually plausible 3D assets.

While optimization-based methods such as DreamFusion [1] achieve impressive results, their high computational cost limits practical deployment. To address this, feed-forward approaches utilize large reconstruction models (LRMs) for efficient 3D asset generation. Early LRM-based methods [17]–[19] adopt transformer-based architectures to map image features into 3D triplane space, followed by volume rendering for reconstruction. More recent works [46]–[52] have significantly enhanced geometry and texture fidelity by integrating multi-view diffusion models [28]–[33], [53]. To improve geometry quality degraded by the inconsistency of 2D multi-view diffusion models, modern 3D generative models [20]–[24] mainly adopt end-to-end native 3D generation paradigms. These models directly learn the inherent distribution of 3D data, and typically leverage structured latent representation, 3D-aware transformers to capture the geometric structure and texture details of 3D assets comprehensively. Nevertheless, they require a large amount of high-quality 3D training data, which incurs high collection costs. With the scarcity of high-quality 3D data, their ability to learn detailed geometric features is limited. Meanwhile, a large portion of the training data rendered from 3D synthetic assets will cause a domain gap with real images in test scenarios, leading to poor performance in practical applications [23].

Our proposed method belongs to the 2D-to-3D lifting paradigm, a category that typically harnesses temporal consistency from video diffusion models to generate dense multi-view frames—these frames are then used as pseudo-labels to guide the 3D optimization process [36]–[40]. This paradigm requires relatively little 3D training data by capitalizing on the strong generative capacity of pre-trained 2D diffusion models, yet it is prone to spatial inconsistency issues. Despite the improvements brought by video diffusion models, recent methods [36]–[40] have enhanced coherence across viewpoints but often restrict viewpoint diversity and still suffer from inconsistencies in complex scenes. From a probabilistic perspective, the reliance on limited and noisy pseudo-labels introduces epistemic uncertainty during optimization, leading to ambiguous supervision and unstable reconstruction. This observation underscores the need for a more robust optimization framework that can adapt to the uncertainty inherent in multi-view generation and effectively mitigate its impact on 3D reconstruction quality.

B. Uncertainty-aware Learning in 3D Reconstruction

With the advancement of deep learning, increasing attention has been given to improving model reliability and interpretability. Estimating model uncertainty not only enhances interpretability but also provides a quantitative measure of confidence in model predictions. Early works [54] categorize uncertainty into two primary types: *epistemic uncertainty* and *aleatoric uncertainty*. Epistemic uncertainty, also known as model uncertainty, reflects the variability in model parameters when trained on the same dataset. Bayesian networks [55]–[59], Monte Carlo dropout [60], [61], and Gaussian noise-based approaches [62], [63] have been explored to quantify this uncertainty by modeling variance in weight distributions. Other methods [64]–[67] introduce auxiliary branches to explicitly model uncertainty, albeit at the cost of increased training complexity. Aleatoric uncertainty, on the other hand, arises from noise in observations, including both input data and annotations. Methods addressing aleatoric uncertainty [68]–[71] often employ dynamic uncertainty-aware loss functions to stabilize training and mitigate the impact of noisy inputs.

Uncertainty estimation has been extensively studied in 3D reconstruction, particularly in Neural Radiance Fields (NeRF) [72] and 3D Gaussian Splatting (3DGS) [43]. While uncertainty in NeRF models stems from variations in camera parameters, lighting conditions [73], [74], occlusions, and sparse viewpoints [75]–[77], similar challenges exist in 3DGS due to its reliance on multi-view consistency. Recent works, such as [78]–[81], have integrated uncertainty-aware learning into training, adaptively adjusting pixel contributions to suppress noise in uncertain regions.

While uncertainty estimation has been extensively studied in NeRF and 3DGS, it remains underexplored in multi-view diffusion-based Image-to-3D generation. To bridge this gap, we propose an uncertainty-aware optimization framework to analyze and mitigate epistemic uncertainty in the 3D asset optimization stage, thereby improving 3D object quality. Specifically, we analyze the sources of epistemic uncertainty and introduce a progressive sampling strategy along with an uncertainty-aware loss function to mitigate its effects, enhancing optimization stability and producing visually compelling 3D assets. To further substantiate our approach, we provide both a theoretical analysis that characterizes its optimization behavior and empirical results that demonstrate consistent improvements in reconstruction quality across diverse scenarios.

III. Method

As illustrated in Figure 2, our framework generates high-quality 3D assets from a reference image through a two-stage pipeline. First, a multi-view video diffusion model synthesizes dense, high-fidelity frames as pseudo-labels for subsequent 3D optimization (*cf.*, Section III-A). Second, uncertainty-aware learning enhances robustness by explicitly addressing epistemic uncertainty arising from two sources: limited observations due to fixed camera poses, and inconsistent content due to geometric and textural conflicts across generated frames. Specifically, we introduce a progressive sampling strategy to incorporate pseudo-labels from diverse viewpoints, effectively alleviating uncertainty from limited observations (*cf.*, Section III-B). Moreover, an uncertainty-aware regularization adaptively adjusts pixel-wise supervision to mitigate uncertainty stemming from inconsistent content (*cf.*, Sections III-C and III-D). Finally, we provide theoretical insights by deriving a probabilistic bound on the expected reconstruction error (*cf.*, Section III-E).

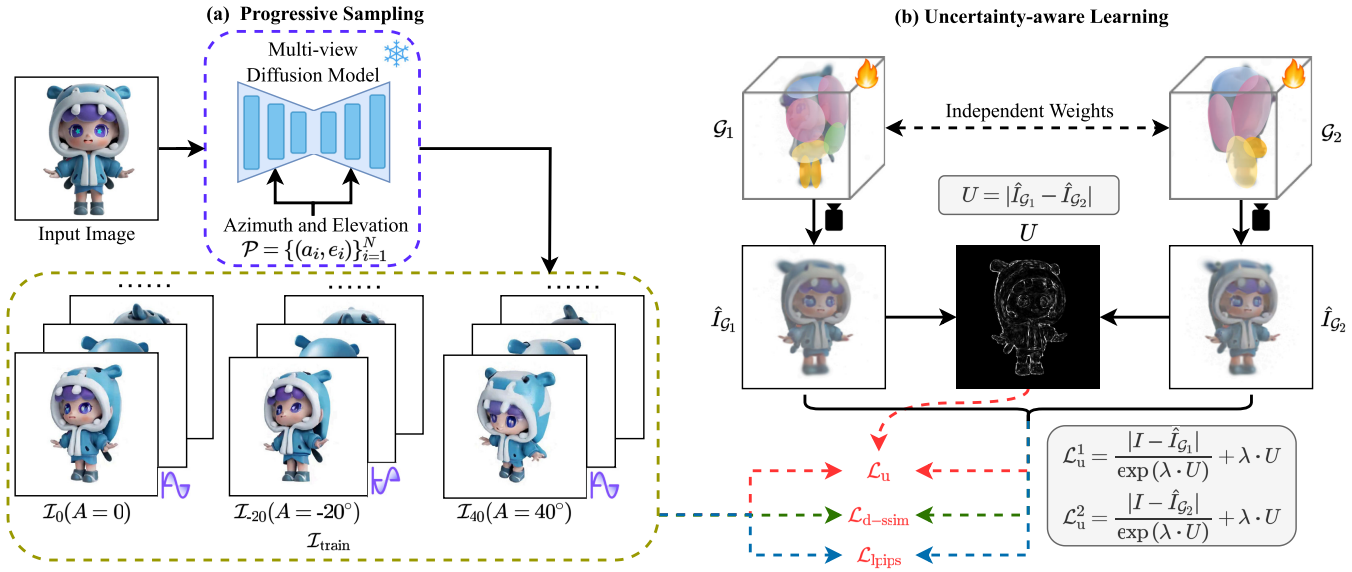


Fig. 2. **Overview of pipeline.** Our method consists of two main components: (a) **Progressive Sampling**, where we first employ a multi-view diffusion model (*e.g.*, SV3D [39]) to generate multi-view frames from diverse viewpoints, progressively integrating these pseudo labels ($\mathcal{I}_{\text{train}}$) into the training process. This strategy mitigates epistemic uncertainty caused by limited observations, enhancing viewpoint coverage. (b) **Uncertainty-aware Learning**, where we independently optimize two Gaussian models ($\mathcal{G}_1, \mathcal{G}_2$) and estimate an uncertainty map (U) from their discrepancies. This uncertainty map captures epistemic uncertainty arising from ambiguous interpretations of inconsistent pseudo labels. Finally, we employ uncertainty-aware regularization ($\mathcal{L}_{\text{u}}$) to adaptively suppress unreliable supervision signals, effectively resolving inconsistencies and producing high-quality, visually coherent 3D assets.

A. Preliminary

3D Gaussian Splatting. 3D Gaussian Splatting (3DGS) [43] is an explicit point-based 3D representation, where each learnable Gaussian point is parameterized by its center position $\mu_i \in \mathbb{R}^3$, scaling $s_i \in \mathbb{R}^3$, rotation $r_i \in \mathbb{R}^4$, color $c_i \in \mathbb{R}^3$, spherical harmonic (SH) coefficients $h_i \in \mathbb{R}^{3\times(k+1)^2}$ up to order k, and opacity $\sigma_i \in \mathbb{R}$. Each Gaussian is modeled as:

$$G_i(x) = e^{-\frac{1}{2}(x-\mu_i)^\top \boldsymbol{\Sigma_i}^{-1}(x-\mu_i)}, \quad \Sigma_i = R_i S_i S_i^T R_i^T, \quad (1)$$

where x represents a 3D position, $\boldsymbol{\Sigma}$ is the covariance matrix, and S_i, R_i are the scaling and rotation matrices derived from s_i and r_i, respectively.

To render a 2D image from a given camera pose, the pixel color is computed via α-blending of the sorted Gaussian points:

$$C = \sum_{i=1}^{N} c_i \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j), \quad (2)$$

where c_i and α_i denote the color and projected opacity of each Gaussian.

The traditional 3DGS model is trained using a reconstruction loss on the 2D projection, which can be formulated as:

$$\mathcal{L}_{\text{rec}} = (1 - \lambda_{\text{s}})\mathcal{L}_1 + \lambda_{\text{s}}\mathcal{L}_{\text{D-SSIM}}, \quad (3)$$

where λ_s is a balancing weight. Starting from a sparse point cloud, the 3DGS model undergoes densification, where high-gradient Gaussians are duplicated and split, followed by pruning to remove redundant points. We adopt 3DGS as our 3D representation due to its efficient training and high-quality rendering capabilities.

Multi-view Diffusion Model. Multi-view diffusion models [27], [82] extend pre-trained 2D image diffusion models [25], [83], [84] to generate spatially and temporally consistent video sequences by jointly denoising multiple frames. A representative method is Stable Video Diffusion (SVD) [27], which consists of an encoder, a denoising U-Net ϵ_θ, and a decoder. Given a conditional image c and an initial noisy sequence x_T, the denoising U-Net estimates the noise component at timestep t. A noise scheduler [85] iteratively refines the frames, updating them as:

$$x_{t-1} = \Phi\left(\epsilon_\theta(x_t; t, c), t, x_t\right), \quad (4)$$

where Φ denotes the noise scheduler. After T denoising steps, a high-quality video sequence of N frames is obtained.

To enhance multi-view consistency, SV3D [39] extends SVD by conditioning the denoising U-Net on camera poses, enabling viewpoint control in image-to-3D generation:

$$x_{t-1} = \Phi\left(\epsilon_\theta(x_t; t, c, a, e), t, x_t\right), \quad (5)$$

where a and e denote the azimuth and elevation angles, respectively. Given an input image, SV3D can generate N frames from diverse viewpoints, supporting both static and dynamic camera trajectories. In this study, we do not pursue a better multi-view diffusion model, but focus on how to employ the off-the-shelf SV3D to generate multi-view frames as pseudo labels for 3D asset optimization, leveraging its strong performance in geometric and texture consistency.

B. Progressive Sampling Strategy

Given an input 2D image, image-to-3D generation aims to produce a 3D model that preserves the geometry and texture of

the reference while maintaining high visual quality from arbitrary viewpoints. In typical two-stage image-to-3D generation methods [36]–[40], [86], video frames with fixed elevations are first generated and subsequently employed as pseudo labels to supervise 3D asset optimization. However, limited viewpoint diversity often leads to high epistemic uncertainty, where the optimized 3D model overfits the provided pseudo labels, resulting in under-reconstruction for novel viewpoints such as top and bottom views. While the model aligns well with the pseudo labels under training viewpoints, it suffers from under-reconstruction and artifacts under novel viewpoints, increasing epistemic uncertainty. To mitigate this, we introduce a sinusoidal elevation sampling strategy combined with progressive pseudo-label integration, enhancing viewpoint coverage while ensuring stable optimization.

Sinusoidal Elevation Sampling. To improve viewpoint diversity, we adopt sinusoidal elevation sampling in which the elevation angle e varies sinusoidally rather than remaining fixed, while the azimuth angle a is uniformly distributed over a full 360° range. Specifically, we define the camera poses $\mathcal{P}$ as:

$$\mathcal{P} = \{(a_i, e_i)\}_{i=1}^{N} = \left\{ \left(\frac{2\pi i}{N}, A \sin\left(\frac{2\pi i}{N} \right) \right) \right\}_{i=1}^{N}, \quad (6)$$

where A denotes the amplitude controlling the maximum deviation, and N denotes the number of frames. This strategy ensures comprehensive azimuth coverage while significantly improving coverage of top and bottom viewpoints, leading to smoother viewpoint transitions and enhanced temporal consistency in generated multi-view frames.

Progressive Learning. Sinusoidal elevation sampling improves viewpoint diversity by generating multi-view frames with varying elevation angles, providing richer supervision for following optimization process. To further enhance supervision, we generate multiple frames with a fixed initial noise latent across different runs, effectively expanding the pseudo-label set. Specifically, given that $\mathcal{I}$ denotes the set of generated pseudo frames, sampling elevation amplitudes A at 0°, -20°, and 40°, we effectively triple the number of pseudo labels, yielding $\{\mathcal{I}_0, \mathcal{I}_{-20}, \mathcal{I}_{40}\}$. Considering that larger elevation differences further reduce frame overlap, we progressively integrate pseudo labels into training, incorporating those with greater elevation variations as training time t increases:

$$\mathcal{I}_{train} = \begin{cases} \{\mathcal{I}_0\}, & 0 \le t < t_1, \\ \{\mathcal{I}_0, \mathcal{I}_{-20}\}, & t_1 \le t < t_2, \\ \{\mathcal{I}_0, \mathcal{I}_{-20}, \mathcal{I}_{40}\}, & t_2 \le t < 1, \end{cases} \quad (7)$$

where t_1 and t_2 represent the step ratios at which different pseudo labels are incorporated into the training process, with values of 0.5 and 0.8, respectively. This progressive learning approach initially trains with a subset of frames to ensure proper initialization and stabilize the training, then gradually incorporates more diverse viewpoints to refine textures and improve coverage.

C. Uncertainty Estimation

While our progressive sampling strategy improves viewpoint diversity, inconsistencies in the generated pseudo labels inevitably persist, introducing epistemic uncertainty into the optimization process. These uncertainties originate from ambiguous interpretations of inconsistent and noisy pseudo labels, resulting in artifacts within the reconstructed 3D objects. To effectively estimate this epistemic uncertainty, we leverage the discrepancies arising from two independently optimized Gaussian models, exploiting the inherent stochasticity in their training process. Intuitively, regions with inconsistent pseudo labels present conflicting optimization objectives, resulting in greater variability between these independently trained models. Conversely, regions with consistent labels exhibit less variation. By modeling these discrepancies, we explicitly quantify pseudo-label inconsistency, facilitating stable and accurate optimization for high-quality 3D asset generation.

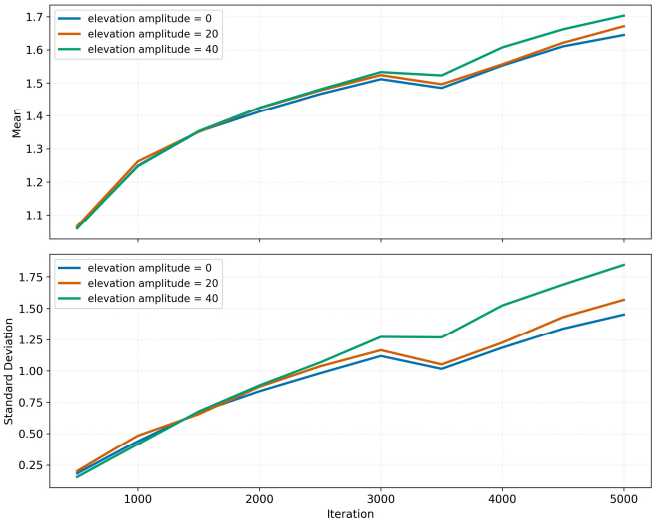


Fig. 3. **Statistical results of the variability between independently trained models.** This figure presents the mean and std of uncertainty maps of training viewpoints under different elevation orbits, calculated every 500 training steps.

Formally, given a set of multi-view pseudo labels $\mathcal{I}_{train} = \{I_i\}_{i=1}^{N}$ corresponding to camera poses $\mathcal{P} = \{P_i\}_{i=1}^{N}$, we simultaneously optimize two Gaussian models, $\mathcal{G}_1$ and $\mathcal{G}_2$, where N denotes the number of frames. At each optimization step, we randomly sample a camera pose $P_i \in \mathcal{P}$, and render both Gaussian models from the corresponding viewpoint to obtain the rendered images $\hat{I_{\mathcal{G}_1}}$ and $\hat{I_{\mathcal{G}_2}}$. The uncertainty map U is then computed as the absolute difference between these rendered images:

$$U = |\hat{I}_{\mathcal{G}_1} - \hat{I}_{\mathcal{G}_2}|. \quad (8)$$

Notably, each Gaussian model is randomly initialized, ensuring observable differences between $\mathcal{G}_1$ and $\mathcal{G}_2$ throughout the optimization. This allows the models to capture pseudo-label uncertainty effectively, guiding the following process.

Why Not a Learnable Uncertainty Model? A possible approach for 3DGS is to assign a learnable variance property to each Gaussian point, enabling the rendering of an uncertainty map through α-blending. However, directly optimizing variance can destabilize training, as some Gaussian points develop excessively large or small variances, leading to suboptimal results [87]. Moreover, variance-based methods often struggle to generalize across different viewpoints, as the

estimated uncertainty is highly sensitive to local optimization dynamics and dataset biases. This makes the learned variance unreliable, particularly in regions with sparse supervision, where the optimization tends to either suppress uncertainty too aggressively or amplify it unnecessarily.

Instead, we estimate uncertainty by computing the discrepancies between two independently optimized Gaussian models, leveraging the inherent randomness in optimization. Unlike variance-based methods [88]–[90], this approach does not require additional learnable parameters, avoiding instability caused by ill-conditioned variance updates. Furthermore, it provides a structurally meaningful uncertainty estimate, as the differences between two models naturally highlight regions of label inconsistency and under-constrained optimization.

To quantify the variability between independently trained models, we conducted supplementary experiments calculating the mean and std of uncertainty maps of training viewpoints under different elevation orbits every 500 steps, as shown in Figure 3. Statistical results show that variability increases with training steps, and larger elevation amplitudes yield greater mean and std of variability—numerical evidence that pseudo-label training enhances variability between two independently optimized Gaussian models in training viewpoints.

Our experiments in Sec. IV-C show that two models suffice for robust uncertainty estimation, capturing epistemic uncertainty without introducing additional learnable parameters or optimization sensitivities. This estimated uncertainty map is then used for uncertainty-aware regularization, which significantly reduces artifacts and floating structures in the final 3D assets. Compared to variance-based methods, our approach improves training stability, enhances generalization across viewpoints, and ensures a more reliable uncertainty estimation without requiring heuristic variance constraints.

D. Uncertainty Regularization

As discussed in Sec. III-C, our uncertainty estimation method captures inconsistencies in pseudo labels, directly incorporating these estimates into the optimization process is essential to fully mitigate their impact. Inconsistent regions introduce conflicting optimization directions, leading to artifacts and floating structures in the generated 3D assets. If these inconsistencies are not properly handled, 3D Gaussian Splatting usually densify redundant points to fit these conflicting pseudo-labels, exacerbating artifacts and floating structures when viewed from alternative perspectives.

To address this, we incorporate our estimated uncertainty into the pixel-wise reconstruction loss, adaptively adjusting the optimization intensity based on the reliability of pseudo labels:

$$\mathcal{L}_{\mathrm{u}}^{1} = \mathbb{E}\left[\frac{|I - \hat{I}_{\mathcal{G}_1}|}{\exp\left(\lambda \cdot U\right)} + \lambda \cdot U\right], \qquad \mathcal{L}_{\mathrm{u}}^{2} = \mathbb{E}\left[\frac{|I - \hat{I}_{\mathcal{G}_2}|}{\exp\left(\lambda \cdot U\right)} + \lambda \cdot U\right], \tag{9}$$

where U represents the estimated uncertainty between the two Gaussian models, and λ controls the uncertainty regularization. Finally, the uncertainty regularization loss is defined as $\mathcal{L}_{\mathrm{u}} = \mathcal{L}_{\mathrm{u}}^{1} + \mathcal{L}_{\mathrm{u}}^{2}$. This formulation adaptively adjusts optimization intensity, assigning lower weights to regions with high uncertainty while regularizing the uncertainty map itself to prevent excessive disparity between the two models. Following prior work [54], [68], we stabilize the optimization by ensuring $\exp(U) > 0$, mitigating numerical instability from division by zero.

To further enhance the impact of uncertainty regularization, we empirically set $\lambda = 5$ to amplify variation in loss weights. In the uncertainty-aware regularization Equation 9, the first term adaptively reweights each pixel's contribution based on uncertainty, reducing the influence of unreliable pseudo labels. The second term regularizes the uncertainty map, preventing excessive uncertainty across all viewpoints. In the limiting case where the uncertainty map is zero, the loss formulation reduces to a standard L1 loss, ensuring uniform optimization across all regions. To further improve visual quality, we incorporate LPIPS loss [91] and D-SSIM Loss [92] into the reconstruction objective. The final loss function is formulated as:

$$\mathcal{L}_{\mathrm{total}} = (1 - \lambda_{\mathrm{s}})\mathcal{L}_{\mathrm{u}} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{d\text{-}ssim}} + \lambda_{\mathrm{l}}\mathcal{L}_{\mathrm{lpips}}, \tag{10}$$

where λ_{s} and λ_{l} are empirically set to 0.2 and 0.5, respectively.

How does uncertainty regularization impact the 3D asset optimization? Pseudo-label inconsistencies often arise in overlapping regions where different viewpoints provide conflicting supervision signals. Without proper regularization, Gaussian models attempt to reconcile these conflicts by densifying redundant points to satisfy pseudo labels from certain viewpoints, while neglecting others. This imbalance can result in artifacts and geometric distortions, particularly in regions with sparse pseudo-label coverage. By integrating uncertainty into the loss function, our method adaptively adjusts the optimization intensity based on pseudo-label reliability. Higher uncertainty values are assigned to regions with inconsistent supervision, reducing their direct influence on optimization. As a result, the model prioritizes well-constrained regions with lower uncertainty while maintaining flexibility in under-constrained areas. This prevents excessive Gaussian densification, ensuring smoother transitions between viewpoints and reducing inconsistencies in the synthesized 3D asset.

E. Bounding the Optimization Error

To analyze the impact of uncertainty-aware optimization, we derive a probabilistic bound on the expected reconstruction error using Hoeffding's inequality [93]. Let I_{gt}^{i} denote the projection pixel value of the ground-truth 3D object, I^{i} the pseudo label generated by the multi-view diffusion model, and $\hat{I}^{i}$ the rendered pixel value. All pixel values are assumed to be bounded within $[0, 1]$, and the uncertainty estimate satisfies $U_M^i \in [0, 1]$, where M represents the number of independently optimized Gaussian models used for uncertainty estimation.

1) Error Decomposition: Applying the *triangle inequality*, we decompose the expected reconstruction error:

$$\mathbb{E}\left[|I_{\mathrm{gt}}^{i} - \hat{I}^{i}|\right] \leq \mathbb{E}\left[|I_{\mathrm{gt}}^{i} - I^{i}|\right] + \mathbb{E}\left[|I^{i} - \hat{I}^{i}|\right]. \tag{11}$$

The first term represents the error introduced by noisy pseudo labels, while the second term quantifies how well the 3D model fits the pseudo labels.

2) Uncertainty-Aware Loss and Expectation: Our uncertainty-aware loss is formulated as:

$$\mathcal{L}_1^{\mathrm{u}} = \sum_i \frac{|I^i - \hat{I}^i|}{\exp\left(\lambda U_M^i\right)} + \lambda U_M^i, \tag{12}$$

where the uncertainty estimate U_M^i is derived from M Gaussian models:

$$U_M^i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{I}_m^i - \bar{I}^i)^2}, \quad \text{where} \quad \bar{I}^i = \frac{1}{M}\sum_{m=1}^{M}\hat{I}_m^i. \tag{13}$$

This formulation represents the *dispersion* of the Gaussian models, capturing epistemic uncertainty.

To bound the expected optimization error, we define the first term in Eq. 12 as a variable: $X_i = \frac{|I^i - \hat{I}^i|}{\exp\left(\lambda U_M^i\right)}$, which represents the uncertainty-weighted reconstruction error at pixel i. Since pixel values are bounded in $[0, 1]$, $\lambda = 5$, and $\exp\left(\lambda U_M^i\right) \in [1, e^5]$, we have $X_i \in [0, 1]$.

Given the uncertainty estimate is bounded by $U_M^i \in [0, 1]$, we have:

$$\exp(\lambda U_M^i) \leq e^{\lambda}. \tag{14}$$

Therefore, we derive an upper bound for the expectation:

$$\mathbb{E}[|I^i - \hat{I}^i|] = \mathbb{E}\left[X_i \cdot \exp(\lambda U_M^i)\right] \leq e^{\lambda}\mathbb{E}[X_i]. \tag{15}$$

3) Applying Hoeffding's Inequality: We now leverage Hoeffding's inequality to bound $\mathbb{E}[X_i]$. For n independently sampled pixels[1], and with $X_i \in [0, 1]$, Hoeffding's inequality states:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mathbb{E}[X_i]\right| \geq t\right) \leq 2\exp(-2nt^2). \tag{16}$$

For a confidence interval of at least $1 - \delta$, we set:

$$2\exp(-2nt^2) = \delta, \tag{17}$$

which gives us the following solution for t:

$$t = \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{18}$$

Thus, with probability at least $1 - \delta$, we obtain the upper bound:

$$\mathbb{E}[X_i] \leq \frac{1}{n}\sum_{i=1}^{n}X_i + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{19}$$

4) Final Bound on the Reconstruction Error: By integrating the above results in Eq. 11, Eq. 15 and Eq. 19, we derive a probabilistic upper bound for the original reconstruction error:

$$\mathbb{E}[|I_{\mathrm{gt}}^i - \hat{I}^i|] \leq \mathbb{E}[|I_{\mathrm{gt}}^i - I^i|] + e^{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n}X_i + \sqrt{\frac{\log(2/\delta)}{2n}}\right), \tag{20}$$

which provides a theoretical guarantee for our uncertainty-aware reconstruction optimization framework. Since the first term is usually fixed, it depends on the quality of pseudo label I_i. The upper bound is, therefore, mainly determined by X_i, which is the primary part of our optimization objective. By minimizing the loss function, we effectively tighten the upper bound on the discrepancy between our generated object and the ground-truth.

5) Impact of Increasing M: Given the definition in Eq. 13, we have:

$$\mathbb{E}[U_M^i] = \frac{1}{\sqrt{M}}\mathbb{E}[U_2^i], \tag{21}$$

which implies that increasing M reduces uncertainty estimation noise at a rate of $O(1/\sqrt{M})$. Substituting this into our uncertainty-weighted loss, we have:

$$\exp(-\lambda U_M^i) \approx \exp(-\lambda U_2^i/\sqrt{M}). \tag{22}$$

Thus, the weight applied to the loss function does *not* decrease linearly with M but rather at a diminishing rate.

Impact of Gaussian Number M. The derived probabilistic bound demonstrates how increasing M impacts uncertainty estimation. As M increases, uncertainty estimation becomes more stable. However, the reduction in uncertainty follows a *diminishing return pattern* of $O(1/\sqrt{M})$. This shows that using more than two Gaussian models does not significantly improve optimization but increases computational cost. Thus, selecting $M = 2$ provides an optimal trade-off, ensuring robust uncertainty estimation while maintaining efficiency.

IV. Experiment

A. Experimental Settings

Dataset and Evaluation Metrics. To evaluate the visual quality of the generated assets, we select 25 objects from the GSO dataset [44], manually choosing front-facing input images. We render 36 ground truth images with uniformly sampled azimuth angles and randomly sampled elevation angles, ensuring coverage of both top and bottom perspectives of the 3D assets. Reconstruction accuracy is assessed using PSNR, SSIM, and LPIPS, which measure pixel-wise similarity and perceptual quality.

Implementation Details. For multi-view frames generation, we employ *sv3d_p* [39], which generates frames from various viewpoints along a sinusoidal orbit. We then sample multiple frames with azimuth angles uniformly distributed across $360°$, and elevation angles defined by sinusoidal amplitudes of $0°$, $-20°$, and $40°$, resulting in 63 frames in total. Moreover, we integrate Perturbed-Attention Guidance (PAG) [94] into the multi-view video diffusion model, improving texture and geometry, particularly in rear-view perspectives.

For 3D asset optimization, we follow the standard 3DGS [43] setup with minor modifications. The spherical harmonics (SH) degree is set to 0, and total optimization iterations are reduced to $5,000$. To mitigate redundant white Gaussian points in inconsistent regions, we apply a random background color technique, effectively reducing white artifacts. During optimization, we progressively increase the render ratio, beginning at 0.25 and scaling up to 0.5 at 20% of the total iterations, and scaling up to 1.0 at 50%. Additionally,

[1] In practice, pixel samples are locally correlated, and the independence assumption does not strictly hold. We mitigate this issue via random pixel/patch sampling during optimization, which reduces spatial correlation. Such a simplification is commonly adopted in stochastic optimization analysis to enable tractable theoretical bounds.

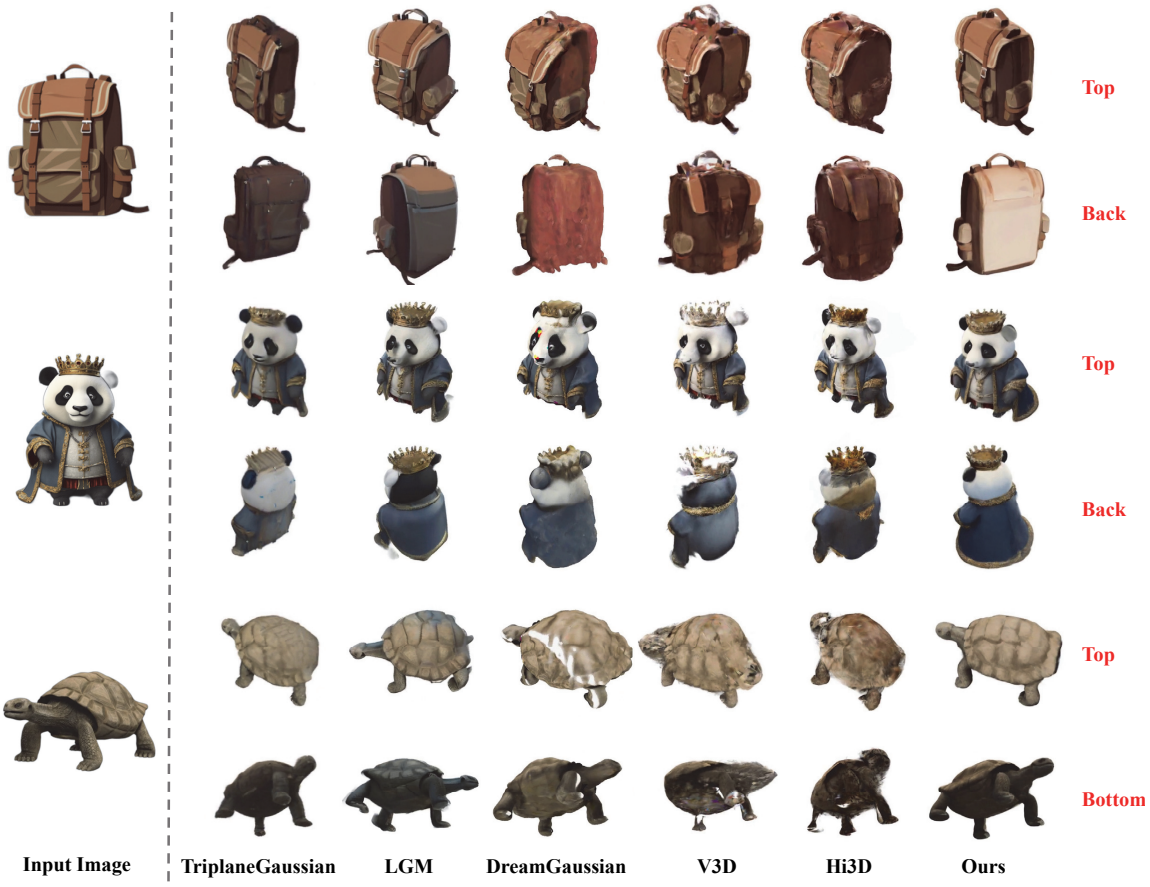


Fig. 4. **Qualitative comparison.** We compare our method with prior state-of-the-art image-to-3D approaches, including TriplaneGaussian [95], LGM [47], DreamGaussian [5], V3D [38], and Hi3D [40]. Our approach produces high-quality geometric and texture details, maintaining consistency even from top and bottom perspectives.

we progressively incorporate frames with different elevations at 50% and 80% of the total iterations.

Competitive Counterparts. We select five image-to-3D generation methods based on 3D Gaussian Splatting [43] for comparison: (1) DreamGaussian [5] is an optimization-based approach that refines 3D assets under Zero123 [2] supervision; (2) TriplaneGaussian [95] is an inference-only method introducing a hybrid triplane-gaussian representation to achieve fast and high-quality 3D reconstruction; (3) LGM [47] is another inference-only method that reconstructs Gaussian models from generated multi-view images; (4) V3D [38] is a multi-view video diffusion model that generates dense frames, which serve as pseudo labels for 3D asset reconstruction; (5) Hi3D [40] employs a two-stage generation framework that enhances texture details by producing high-resolution multi-view frames.

B. Experimental Results

Qualitative Comparison. As shown in Figure 4, we provide qualitative comparisons across approaches, including optimization-based, inference-only, and two-stage methods. TriplaneGaussian [95] efficiently generates 3D assets but produces lower-resolution outputs with limited texture detail. LGM [47] employs an asymmetric U-Net to produce high-resolution 3D objects; however, inconsistencies in the input multi-view images often lead to artifacts and floats. DreamGaussian [5] leverages SDS Loss for 3D object optimization, but it frequently generates coarse textures on the back, causing a noticeable disconnect between front and back views. V3D [38] and Hi3D [40] employ multi-view video diffusion models to generate dense, high-quality frames, producing 3D objects with detailed textures. However, since optimization is performed from a limited set of fixed viewpoints, these methods tend to overfit to the generated frames, leading to underdeveloped geometry and texture details, particularly in top and bottom perspectives. Our approach samples multi-view frames from diverse viewpoints and incorporates uncertainty-aware learning to mitigate inconsistencies, resulting in more visually coherent and high-fidelity 3D generation.

Quantitative Comparison. We select 25 objects from the GSO [44] dataset and use SSIM, PSNR, and LPIPS to evaluate the visual quality of the generated 3D objects. As shown in Table I, we achieve superior or comparable results, demonstrating the effectiveness of our approach in generating high-quality and visually impressive 3D assets. Specifically, our method performs well on SSIM, indicating strong structural consistency and effectively mitigating noise, artifacts, and floats in inconsistent regions. Additionally, we observe a slight improvement in LPIPS, suggesting enhanced perceptual quality in the generated 3D assets. While our method excels in

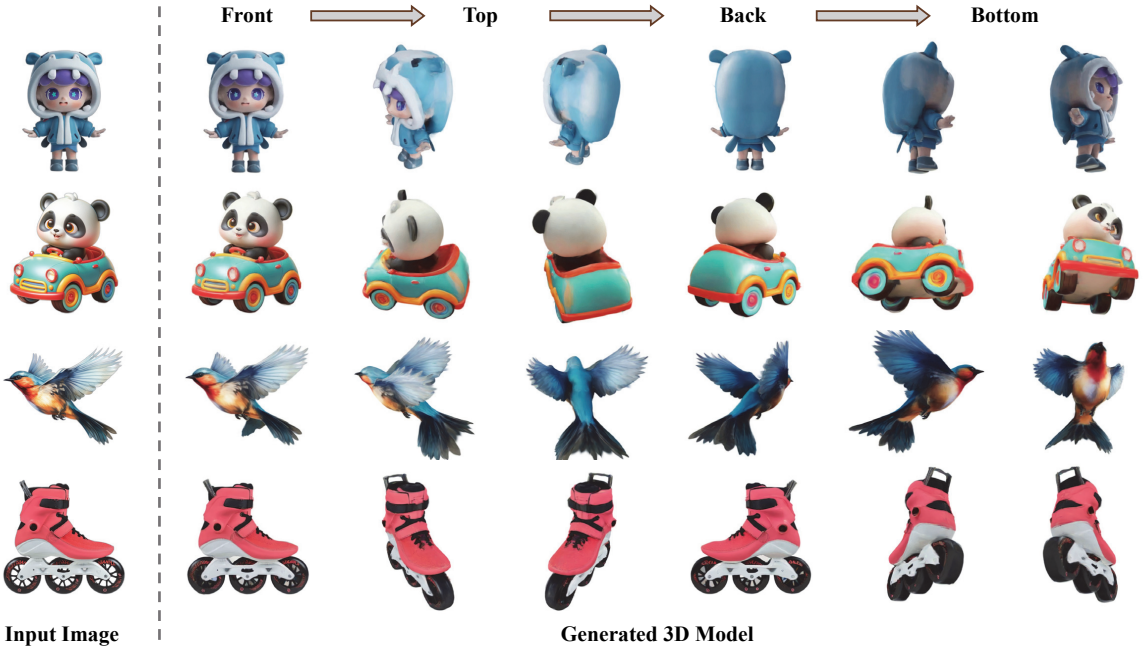


Fig. 5. **Visual results from different perspectives.** Given an input image, our method mitigates inconsistencies between generated dense frames during 3D asset optimization, reducing edge artifacts and floats while producing high-fidelity 3D objects. We render six images uniformly across an azimuth range of 0 to 360°, with elevations following a sine function with 30° amplitude, effectively capturing **front, top, back, and bottom** perspectives, which are crucial to real-world applications yet often overlooked by most existing methods.

TABLE I
QUANTITATIVE COMPARISON. OUR METHOD ACHIEVES SUPERIOR OR COMPARABLE RESULTS, DEMONSTRATING ITS EFFECTIVENESS IN GENERATING HIGH-QUALITY 3D ASSETS.

Methods	PSNR↑	SSIM↑	LPIPS↓	Preference↑
DreamGaussian [5]	17.162	0.8252	0.2039	25.14%
TriplaneGaussian [95]	14.0062	0.8161	0.2531	19.81%
LGM [47]	14.5874	0.8083	0.2488	46.95%
V3D [38]	17.1847	0.8085	0.2055	19.24%
Hi3D [40]	**17.2559**	0.8217	0.2014	21.91%
Ours	16.9684	**0.8346**	**0.2004**	**66.95%**

TABLE II
QUANTITATIVE ANALYSIS OF PROGRESSIVE SAMPLING. BY PROGRESSIVELY INCORPORATING MULTIPLE GENERATED FRAMES WITH SINUSOIDAL ELEVATIONS, OUR METHOD ENHANCES 3D GENERATION QUALITY.

Setting	PSNR↑	SSIM↑	LPIPS↓
Constant Elevations	16.1616	0.8156	0.2196
+ Sinusoidal Elevations	16.4778	0.8221	**0.2122**
+ Progressive Learning	**16.4826**	**0.8268**	0.2126

SSIM and LPIPS, achieving strong structural consistency and perceptual quality, it slightly trails Hi3D [40] in PSNR. This discrepancy stems from our uncertainty regularization, which adaptively adjusts pixel-level supervision, prioritizing structural smoothness in inconsistent regions rather than enforcing strict pixel-wise alignment. As a result, our method optimizes for overall geometric and perceptual coherence, which leads to marginally lower scores on pixel-wise metrics that favor exact structural matching.

User Study. To evaluate visual quality, we curate a set of 30 samples and conducted a user study with 35 participants, as summarized in Table I. Each participant is tasked with selecting the top two results that best matched the input image and exhibited the highest visual quality. The total preference score across all participants is normalized to sum to 200%, enabling a fair comparison across methods. As shown in the results, our method is selected more frequently, highlighting its ability to consistently generate visually superior 3D assets that align more closely with user expectations. This findings further validate the effectiveness of our approach in delivering high-quality visual results.

Visualization. To provide a more comprehensive visualization of our results, we render the 3D models from multiple camera perspectives, as shown in Figure 5. Specifically, azimuth angles are uniformly sampled across a full 360-degree range, while elevation angles are designed to follow a sinusoidal function with an amplitude of 30°. This setup captures a full range of perspectives, including front, back, sides, top, and bottom, offering a complete and balanced representation of the 3D models. Thanks to our progressive sampling strategy and uncertainty-aware learning, we not only achieve high-quality results from various camera perspectives but also effectively reduce floating artifacts and inconsistencies in generated frames.

C. Ablation Analysis

Impact of Progressive Sampling. As presented in Table II and Figure 6, we incrementally incorporate different components to evaluate the effectiveness of progressive sampling in optimizing 3D assets. Expanding viewpoint coverage

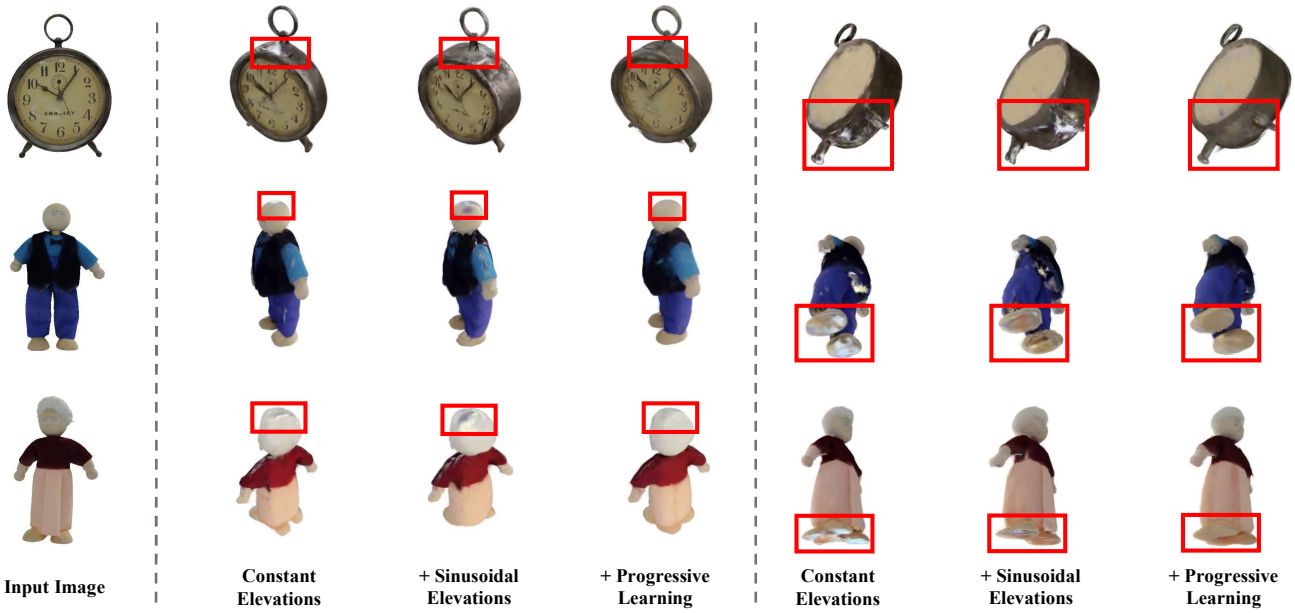


Fig. 6. **Ablation analysis of the progressive sampling strategy.** By incrementally incorporating progressive sampling techniques, our approach achieves enhances visual quality across different viewpoints while alleviating the under-reconstruction issue in unobserved regions.

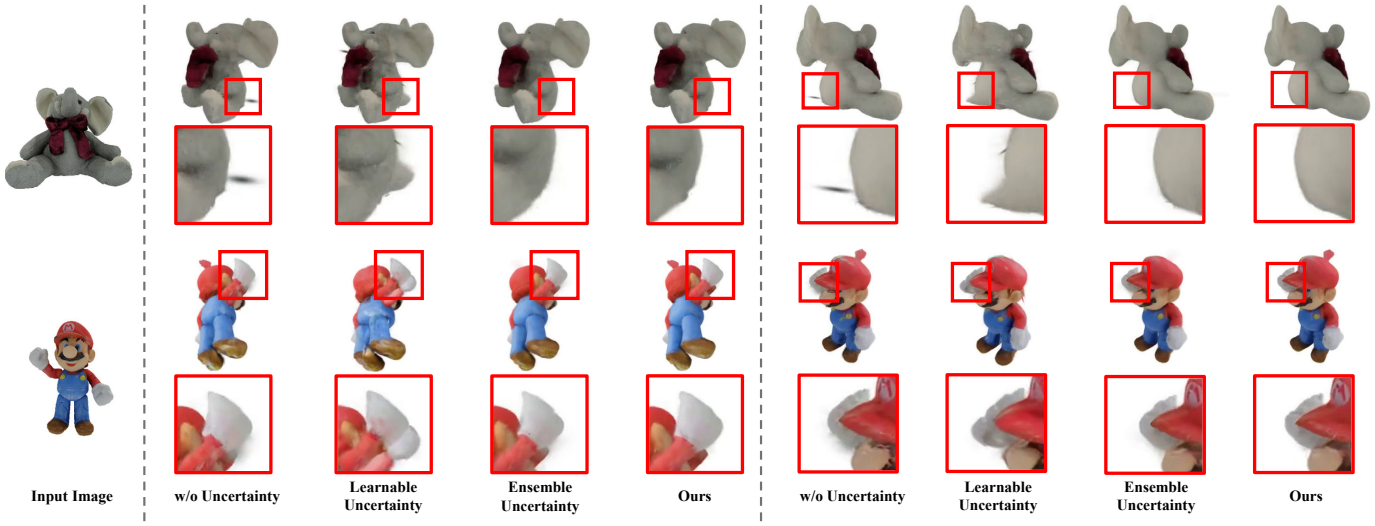


Fig. 7. **Ablation analysis of the uncertainty estimation design.** Our approach models uncertainty through the absolute difference between two concurrently optimized Gaussian models, ensuring both stability and efficiency.

results in qualitative improvements, with fewer artifacts in top and bottom perspectives, partially alleviating the under-reconstruction issue. Additionally, through progressive learning that gradually incorporates pseudo labels sampled from diverse elevation amplitudes, we achieve high-quality visual results across various viewpoints, effectively mitigating the epistemic uncertainty from limited multi-view supervision in unobserved regions. As shown in Table II, the incremental integration of sinusoidal elevations and progressive learning leads to improved quantitative results, further validating their effectiveness in 3D object optimization. By initially optimizing with a small set of frames at diverse elevations for accurate initialization and progressively incorporating additional frames, our progressive sampling strategy refines texture details while preventing geometric distortions.

Design of Uncertainty Estimation. We explore different approaches for uncertainty estimation, as shown in Table III and Figure 7, with additional analysis on the number of Gaussian models (M) and time complexity (Time (s)) to evaluate both performance and efficiency. A straightforward approach is to assign a learnable variance property into 3DGS [43], allowing uncertainty to be rendered through α-blending. However, introducing a learnable variance property and directly regressing uncertainty in this manner often lead to training instability, resulting in degraded performance. An alternative is the ensemble approach, which simultaneously optimizes multiple Gaussian models, averaging their predictions for the final rendered image while using variance as a

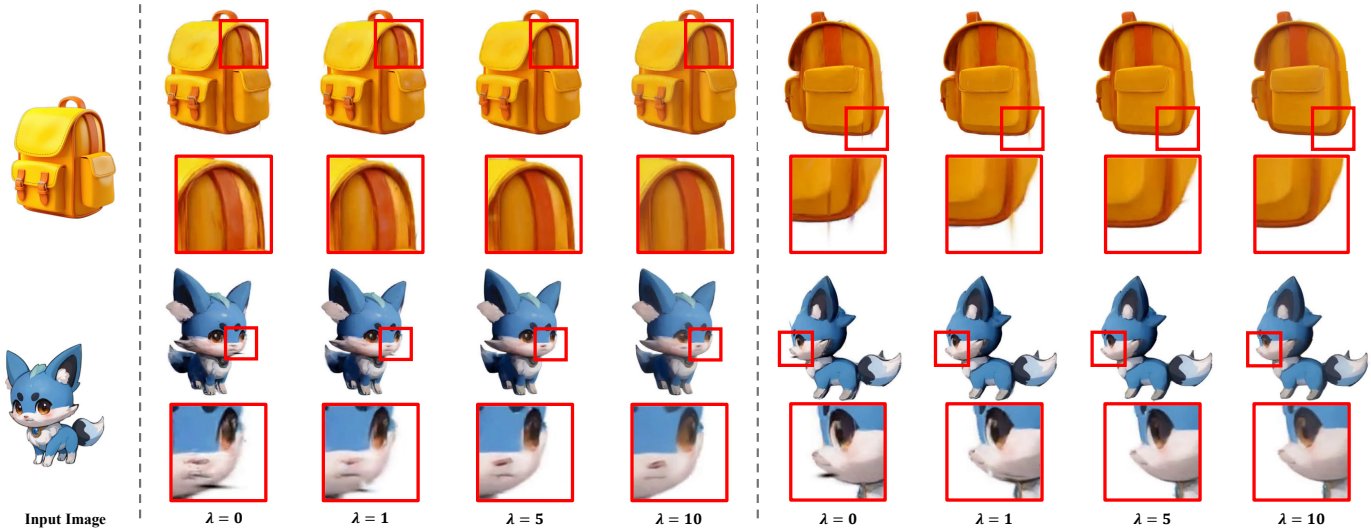


Fig. 8. **Ablation analysis of the uncertainty weight λ.** Increasing λ reduces edge noise and artifacts but results in smoother and blurrier generated results. In our experiments, $\lambda = 5$ achieves the best trade-off between artifact suppression and detail preservation.

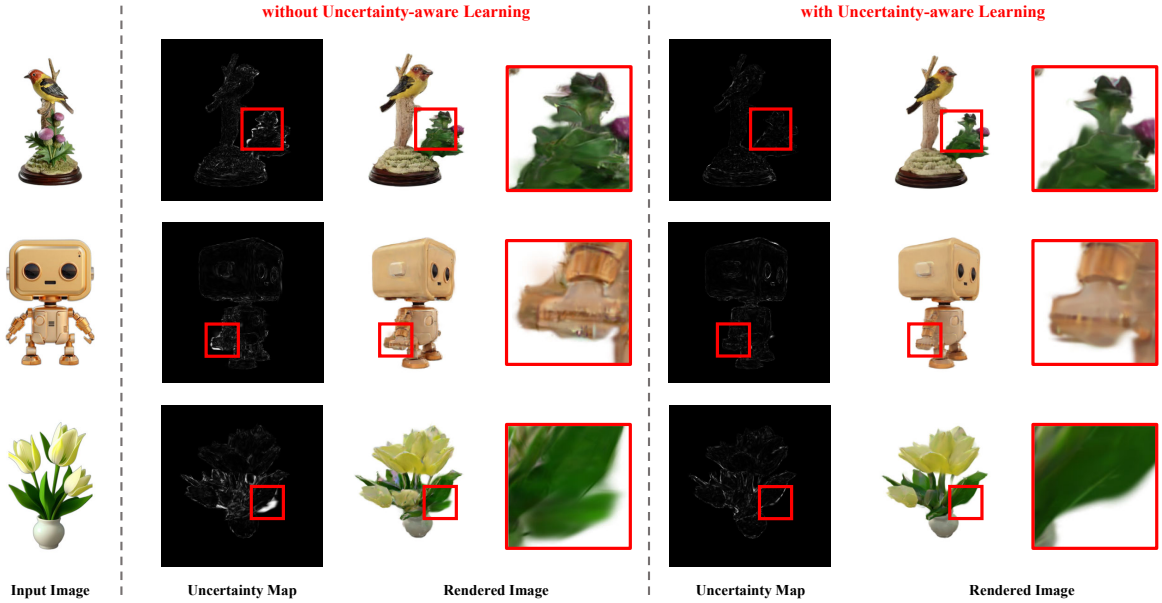


Fig. 9. **Visualization of the uncertainty map.** The estimated uncertainty map highlights higher values in edge and conflicts regions, effectively capturing the inconsistencies present in the pseudo labels. With uncertainty-aware learning, we significantly reducing artifacts and floats in inconsistent regions.

measure of uncertainty. This approach effectively alleviated noise, artifacts, and floats at the edges of generated 3D assets, resulting in smoother output. Our method can be seen as a simplified ensemble approach, where *two* Gaussian models suffice to accurately capture uncertainty. With the increase of M, the training time of both methods increases, while the quantitative performance metrics also gradually improve. Under the same M value, the training time consumption of our proposed method is comparable to that of the ensemble method, but our method achieves slightly better quantitative performance. Experimental results show that absolute differences between these two models effectively model uncertainty while achieving comparable benefits with greater efficiency. This observation is consistent with our theoretical analysis in Sec. III-E, where Eq. (22) predicts that increasing M yields diminishing returns at a rate of $O(1/\sqrt{M})$. Table III empirically confirms this: from $M = 2$ to $M = 5$, performance gains are marginal while training time increases significantly.

Impact of the Uncertainty Regularization Weight λ. We conduct an ablation study to assess the impact of the uncertainty regularization weight λ in Table IV. As λ increases, the impact of uncertainty regularization on the optimization process of 3D assets becomes more pronounced, along with better quantitative results. This indicates that uncertainty regularization effectively enhances the generation quality of 3D assets, with higher weights leading to improved results. When λ is set to a low value, the pixel-wise weights in uncertainty regularization become uniform, failing to dif-

TABLE III
QUANTITATIVE ANALYSIS OF UNCERTAINTY ESTIMATION DESIGN. WE EXPLORE VARIOUS APPROACHES, INCLUDING LEARNABLE AND ENSEMBLE-BASED METHODS. OUR METHOD ACHIEVES COMPARABLE PERFORMANCE TO THE ENSEMBLE APPROACH WHILE OUTPERFORMING THE LEARNABLE APPROACH, DEMONSTRATING THAT TWO GAUSSIAN MODELS ARE BOTH SUFFICIENT AND EFFICIENT.

Setting	PSNR↑	SSIM↑	LPIPS↓	Time (s)↓
w/o Uncertainty	16.4826	0.8268	0.2126	**177.37**
Learnable	16.4052	0.8201	0.2127	178.52
Ensemble (M=2)	16.7651	0.8269	0.2066	194.60
Ensemble (M=3)	16.8107	0.8287	0.2058	210.67
Ensemble (M=5)	17.0168	0.8339	0.2015	243.29
Ours (M=2)	16.9684	0.8346	**0.2004**	195.47
Ours (M=3)	16.9582	0.8357	0.2024	213.50
Ours (M=5)	**17.0746**	**0.8373**	0.2033	254.02

TABLE IV
QUANTITATIVE ANALYSIS OF UNCERTAINTY REGULARIZATION WEIGHT. AS THE VALUE OF λ INCREASES, THE DEGREE OF UNCERTAINTY REGULARIZATION INTENSIFIES, LEADING TO IMPROVED QUANTITATIVE RESULTS.

Setting	PSNR↑	SSIM↑	LPIPS↓
λ=0	16.4826	0.8268	0.2126
λ=1	16.8259	0.8281	0.2047
λ=5	16.9684	0.8346	**0.2004**
λ=10	**17.0635**	**0.8373**	0.2006

TABLE V
QUANTITATIVE ANALYSIS OF CAMERA ORBITS. DIFFERENT CAMERA ORBITS HAVE VARYING IMPACTS ON THE GENERATION QUALITY, AND THE OPTIMAL ORBIT IS SELECTED BASED ON COMPREHENSIVE EVALUATION.

Method	PSNR↑	SSIM↑	LPIPS↓
Sinusoidal (0, -10, 20)	16.5856	0.8285	0.2109
Ours-Sinusoidal (0, -20, 40)	16.9684	0.8346	**0.2004**
Sinusoidal (0, -30, 60)	16.8286	0.8339	0.2061
Constant (0, -20, 40)	16.1462	0.8318	0.2139
Sinusoidal (0, -10, 20, -20, 40)	16.9932	0.8352	0.2024
Sinusoidal (0, -10, 20, -30, 60)	16.7460	0.8348	0.2062
Sinusoidal (0, -20, 40, -30, 60)	**17.0611**	**0.8368**	0.2013

TABLE VI
QUANTITATIVE ANALYSIS OF STEP RATIOS. THE STEP RATIOS HAVE LITTLE IMPACT ON THE MODEL PERFORMANCE, AND THE DEFAULT SETTINGS CAN SUPPORT STABLE CONVERGENCE.

t_1	t_2	PSNR↑	SSIM↑	LPIPS↓
0.0	0.0	16.8100	0.8318	0.2037
0.4	0.7	16.8644	0.8329	0.2032
0.4	0.8	16.8833	0.8330	0.2030
0.4	0.9	16.8862	0.8328	0.2034
0.5	0.7	16.8805	0.8332	0.2030
0.5	0.8	16.9684	0.8346	0.2004
0.5	0.9	16.9020	0.8330	0.2030
0.6	0.7	16.8787	0.8330	0.2030
0.6	0.8	16.8974	0.8332	0.2030
0.6	0.9	16.9088	0.8331	0.2029

ferentiate between inconsistent and consistent regions. This results in insufficient mitigation of artifacts and floats arising from over-reconstruction in inconsistent regions. In contrast, a high value for λ substantially reduces supervision in high-uncertainty regions, effectively alleviating artifacts and floats.

TABLE VII
QUANTITATIVE ANALYSIS OF DIFFERENT BASELINE MODELS WITH AND WITHOUT OUR METHOD. OUR METHOD CAN EFFECTIVELY IMPROVE THE PERFORMANCE OF DIFFERENT MULTI-VIEW DIFFUSION MODELS, ESPECIALLY FOR SV3D.

Method	PSNR↑	SSIM↑	LPIPS↓
V3D [38]	17.1847	0.8085	0.2055
V3D [38] + Ours	17.1437	0.8169	0.2017
Hi3D [40]	17.2559	0.8217	0.2014
Hi3D [40] + Ours	**17.2917**	0.8281	0.2009
SV3D [39]	16.1616	0.8156	0.2196
SV3D [39] + Ours	16.9684	**0.8346**	**0.2004**

However, as shown in Figure 8, excessively large λ values lead to under-reconstruction, causing inconsistent areas to appear overly smooth and blurry. Overall, we determine that $\lambda = 5$ achieves the best trade-off, effectively reducing artifacts while preserving texture details and preventing excessive smoothing.

Impact of Camera Orbits and Step Ratios. We conduct ablation studies to evaluate the influences of camera orbits and step ratios (t_1, t_2) of progressive sampling strategy on 3D generation quality, with the results presented in Tables V and VI, respectively. For camera orbits, the study covers variations in elevation amplitudes, comparisons between sinusoidal and fixed orbits, and the effects of orbit numbers. As illustrated by the quantitative results, camera orbit configurations significantly affect generation quality: both small and large elevation amplitudes yield poorer performance, likely due to insufficient viewing angle coverage and increased generation inconsistencies, respectively. Fixed orbits underperform sinusoidal orbits with the same elevation range, since SV3D is trained on sinusoidal orbits and fixed orbits have slightly worse continuity in viewpoints coverage for video diffusion models. Furthermore, increasing the number of orbits contributes to improved generation quality but incurs additional computational costs during the multi-view image generation stage. Overall, we conclude that the sinusoidal orbit with elevations of 0, -20, 40 achieves the optimal trade-off between generation quality and computational efficiency. In contrast, step ratios (t_1, t_2) have little influence on model performance, as the PSNR, SSIM, and LPIPS metrics are very close. This indicates that the default settings of (t_1, t_2) are sufficient to support stable model convergence, eliminating the need for additional parameter adjustment for different object categories.

Generalization Verification of Our Method. To verify the effectiveness of our method on different multi-view video diffusion models, we conduct an ablation study on V3D [38] and Hi3D [40], with the experimental results presented in Table VII. Since these models exhibit low generation consistency on camera orbits with non-zero elevation, we follow their default settings, where they generate 16 frames and 18 frames respectively; these frames were then divided into three parts at intervals and trained with our progressive learning and uncertainty regularization strategies. As illustrated by the quantitative results, both V3D and Hi3D combined with our method achieve a certain improvement in SSIM, which indicates that our strategy has a positive effect on enhancing the structural consistency of generated 3D assets across different

models. However, both the performance improvement and the final results are limited compared to those of SV3D [39], which is why we use SV3D as our default multi-view model.

Overall, our method demonstrates good generalization ability, as it can stably improve the performance of different multi-view video diffusion models. We note that our framework is not restricted to video diffusion pipelines. For single-pose-conditioned models (*e.g.*, Zero123++ [31]), multiple views can be independently generated by varying camera poses, after which our progressive integration and uncertainty-aware optimization apply in the same manner. In fact, the lack of temporal coherence in independently generated views is likely to introduce greater pseudo-label inconsistency, making uncertainty-aware optimization even more beneficial in such settings.

Visualization of the Uncertainty Map. As shown in Figure 9, we visualize of the estimated uncertainty map, normalized using min-max scaling for clarity. Without uncertainty-aware learning, high uncertainty values are observed along object boundaries and inconsistent regions, indicating that the two Gaussian models, despite being jointly optimized on the same data, fail to fully converge due to inherent pseudo-label uncertainty. Moreover, in the corresponding rendered images, regions with high uncertainty often correlate with degraded visual quality, manifesting as unexpected floating artifacts and conflicting geometry. This suggests that the uncertainty map effectively captures discrepancies between the two Gaussian models, serving as a valuable indicator for identifying conflict regions during optimization. By integrating uncertainty regularization, we mitigate the adverse effects of these inconsistencies, significantly reducing artifacts and floating elements, thereby improving overall visual fidelity.

V. Conclusion

In this work, we have presented an uncertainty-aware optimization framework for image-to-3D generation, explicitly targeting the epistemic uncertainty that emerges from ambiguous interpretations and conflicting optimization objectives caused by imperfect pseudo labels. Specifically, we have identified and addressed two primary sources of epistemic uncertainty: limited observations due to restricted viewpoints, and inconsistent content arising from noisy multi-view frames. To mitigate uncertainty from limited observations, we introduce a progressive sampling strategy that systematically enriches viewpoint diversity throughout the optimization process. Concurrently, our uncertainty-aware learning approach adaptively estimates inconsistencies in pseudo labels, effectively suppressing artifacts and enhancing the robustness of pixel-wise supervision. Furthermore, we provide a theoretical analysis by deriving a probabilistic upper bound on the expected reconstruction error, offering insights into the effectiveness of uncertainty-aware optimization. Extensive experiments validate that our approach improves structural consistency and perceptual quality, leading to smoother and more accurate 3D generation. Our approach thus serves as a robust solution for mitigating pseudo-label inconsistencies, providing a strong foundation for broader applicability to diverse image-to-3D tasks involving multi-view supervision.

References

[1] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *ICLR*, 2023.

[2] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *ICCV*, 2023.

[3] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," in *NeurIPS*, 2024.

[4] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov, and B. Ghanem, "Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors," in *ICLR*, 2024.

[5] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," in *ICLR*, 2024.

[6] S. Yang, Y. Wang, H. Li, J. Meng, X. Meng, and J. Zhang, "Fourier123: One image to high-quality 3d object generation with hybrid fourier score distillation," *arXiv:2405.20669*, 2024.

[7] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *CVPR*, 2019.

[8] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *ECCV*, 2018.

[9] A. Trevithick and B. Yang, "Grf: Learning a general radiance field for 3d representation and rendering," in *ICCV*, 2021.

[10] S. Duggal and D. Pathak, "Topologically-aware deformation fields for single-view 3d reconstruction," in *CVPR*, 2022.

[11] X. Zhang, Z. Zheng, D. Gao, B. Zhang, Y. Yang, and T.-S. Chua, "Multi-view consistent generative adversarial networks for compositional 3d-aware image synthesis," *International Journal of Computer Vision*, vol. 131, no. 8, 2023.

[12] H. Jun and A. Nichol, "Shap-e: Generating conditional 3d implicit functions," *arXiv:2305.02463*, 2023.

[13] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv:2212.08751*, 2022.

[14] A. Gupta, W. Xiong, Y. Nie, I. Jones, and B. Oğuz, "3dgen: Triplane latent diffusion for textured mesh generation," *arXiv:2303.05371*, 2023.

[15] Y. Zhou, D. Ye, H. Zhang, X. Xu, H. Sun, Y. Xu, X. Liu, and Y. Zhou, "Recurrent diffusion for 3d point cloud generation from a single image," *IEEE Transactions on Image Processing*, 2025.

[16] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen *et al.*, "Rodin: A generative model for sculpting 3d digital avatars using diffusion," in *CVPR*, 2023.

[17] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, "LRM: large reconstruction model for single image to 3d," in *ICLR*, 2024.

[18] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao, "Triposr: Fast 3d object reconstruction from a single image," *arXiv:2403.02151*, 2024.

[19] M. Boss, Z. Huang, A. Vasishta, and V. Jampani, "Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement," *arXiv:2408.00653*, 2024.

[20] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, "Structured 3d latents for scalable and versatile 3d generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 2025.

[21] S. Wu, Y. Lin, Y. Zeng, F. Zhang, J. Xu, P. Torr, X. Cao, and Y. Yao, "Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer," in *NeurIPS*, 2024.

[22] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, Y. Feng, M. Yang, S. Zhang, X. Yang, H. Shi, S. Liu, J. Wu, Y. Lian, F. Yang, R. Tang, Z. He, X. Wang, J. Liu, X. Zuo, Z. Chen, B. Lei, H. Weng, J. Xu, Y. Zhu, X. Liu, L. Xu, C. Hu, T. Huang, L. Wang, J. Zhang, M. Chen, L. Dong, Y. Jia, Y. Cai, J. Yu, Y. Tang, H. Zhang, Z. Ye, P. He, R. Wu, C. Zhang, Y. Tan, J. Xiao, Y. Tao, J. Zhu, J. Xue, K. Liu, C. Zhao, X. Wu, Z. Hu, L. Qin, J. Peng, Z. Li, M. Chen, X. Zhang, L. Niu, P. Wang, Y. Wang, H. Kuang, Z. Fan, X. Zheng, W. Zhuang, Y. He, T. Liu, Y. Yang, D. Wang, Y. Liu, J. Jiang, J. Huang, and C. Guo, "Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation," *arXiv:2501.12202*, 2025.

[23] C. Ye, Y. Wu, Z. Lu, J. Chang, X. Guo, J. Zhou, H. Zhao, and X. Han, "Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging," *arXiv:2503.22236*, 2025.

[24] W. Li, X. Zhang, Z. Sun, D. Qi, H. Li, W. Cheng, W. Cai, S. Wu, J. Liu, Z. Wang, X. Chen, F. Tian, J. Pan, Z. Li, G. Yu, X. Zhang, D. Jiang, and P. Tan, "Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets," *arXiv:2505.07747*, 2025.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[26] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*, 2022.

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

[28] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," in *ICLR*, 2024.

[29] P. Wang and Y. Shi, "Imagedream: Image-prompt multi-view diffusion for 3d generation," *arXiv:2312.02201*, 2023.

[30] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," in *ICLR*, 2024.

[31] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," *arXiv:2310.15110*, 2023.

[32] S. Woo, B. Park, H. Go, J.-Y. Kim, and C. Kim, "Harmonyview: Harmonizing consistency and diversity in one-image-to-3d," in *CVPR*, 2024.

[33] P. Li, Y. Liu, X. Long, F. Zhang, C. Lin, M. Li, X. Qi, S. Zhang, W. Xue, W. Luo, P. Tan, W. Wang, Q. Liu, and Y. Guo, "Era3d: High-resolution multiview diffusion using efficient row-wise attention," in *NeurIPS*, 2024.

[34] C. Cao, C. Yu, S. Liu, F. Wang, X. Xue, and Y. Fu, "Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model," in *CVPR*, 2025.

[35] Q. Zhang, S. Zhai, M. A. B. Martin, K. Miao, A. Toshev, J. Susskind, and J. Gu, "World-consistent video diffusion with explicit 3d modeling," in *CVPR*, 2025.

[36] L. Melas-Kyriazi, I. Laina, C. Rupprecht, N. Neverova, A. Vedaldi, O. Gafni, and F. Kokkinos, "IM-3D: iterative multiview diffusion and reconstruction for high-quality 3d generation," in *ICML*, 2024.

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

[38] Z. Chen, Y. Wang, F. Wang, Z. Wang, and H. Liu, "V3d: Video diffusion models are effective 3d generators," *arXiv:2403.06738*, 2024.

[39] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani, "Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion," in *ECCV*, 2024.

[40] H. Yang, Y. Chen, Y. Pan, T. Yao, Z. Chen, C. Ngo, and T. Mei, "Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models," in *ACM MM*, 2024.

[41] X. Ren, T. Shen, J. Huang, H. Ling, Y. Lu, M. Nimier-David, T. Müller, A. Keller, S. Fidler, and J. Gao, "Gen3c: 3d-informed world-consistent video generation with precise camera control," in *CVPR*, 2025.

[42] J. Zhou, H. Gao, V. Voleti, A. Vasishta, C.-H. Yao, M. Boss, P. Torr, C. Rupprecht, and V. Jampani, "Stable virtual camera: Generative view synthesis with diffusion models," in *ICCV*, 2025.

[43] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, 2023.

[44] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *ICRA*, 2022.

[45] H. Yi, Z. Zheng, X. Xu, and T.-s. Chua, "Progressive text-to-3d generation for automatic 3d prototyping," *arXiv:2309.14600*, 2023.

[46] J. Li, H. Tan, K. Zhang, Z. Xu, F. Luan, Y. Xu, Y. Hong, K. Sunkavalli, G. Shakhnarovich, and S. Bi, "Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model," in *ICLR*, 2024.

[47] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu, "Lgm: Large multi-view gaussian model for high-resolution 3d content creation," in *ECCV*, 2024.

[48] Z. Wang, Y. Wang, Y. Chen, C. Xiang, S. Chen, D. Yu, C. Li, H. Su, and J. Zhu, "CRM: single image to 3d textured mesh with convolutional reconstruction model," in *ECCV*, vol. 15089, 2024.

[49] Y. Xu, Z. Shi, Y. Wang, H. Chen, C. Yang, S. Peng, Y. Shen, and G. Wetzstein, "GRM: large gaussian reconstruction model for efficient 3d reconstruction and generation," in *ECCV*, 2024.

[50] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," *arXiv:2404.07191*, 2024.

[51] H. Kong, X. Yang, and X. Wang, "Generative sparse-view gaussian splatting," in *CVPR*, 2025.

[52] H. Xu, S. Peng, F. Wang, H. Blum, D. Barath, A. Geiger, and M. Pollefeys, "Depthsplat: Connecting gaussian splatting and depth," in *CVPR*, 2025.

[53] Z. Huang, M. Boss, A. Vasishta, J. M. Rehg, and V. Jampani, "Spar3d: Stable point-aware reconstruction of 3d objects from single images," in *CVPR*, 2025.

[54] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NeurIPS*, 2017.

[55] S. Parsons, "Bayesian networks and decision graphs," *Springer*, vol. 23, no. 4, 2008.

[56] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[57] Z. Lu, D. Li, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Uncertainty-aware source-free domain adaptive semantic segmentation," *IEEE Transactions on Image Processing*, 2023.

[58] B. N. Patro, M. Lunayach, and V. P. Namboodiri, "Uncertainty class activation map (u-cam) using gradient certainty method," *IEEE Transactions on Image Processing*, 2021.

[59] Y. Zhang, J. Zhang, W. Hamidouche, and O. Deforges, "Predictive uncertainty estimation for camouflaged object detection," *IEEE Transactions on Image Processing*, 2023.

[60] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.

[61] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *NeurIPS*, 2015.

[62] T. Yu, D. Li, Y. Yang, T. M. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *ICCV*, 2019.

[63] Y. Chen, Z. Zheng, W. Ji, L. Qu, and T.-S. Chua, "Composed image retrieval with text feedback via multi-grained uncertainty regularization," in *ICLR*, 2024.

[64] M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, and J. Kleinberg, "Direct uncertainty prediction for medical second opinions," in *ICML*, 2019.

[65] J. Nandy, W. Hsu, and M. L. Lee, "Towards maximizing the representation gap between in-domain & out-of-distribution examples," in *NeurIPS*, 2020.

[66] J. Lee and G. AlRegib, "Gradients as a measure of uncertainty in neural networks," in *ICIP*, 2020.

[67] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, 2021.

[68] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *International Journal of Computer Vision*, 2021.

[69] M. Litrico, A. Del Bue, and P. Morerio, "Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation," in *CVPR*, 2023.

[70] P. Her, L. Manderle, P. A. Dias, H. Medeiros, and F. Odone, "Uncertainty-aware gaze tracking for assisted living environments," *IEEE Transactions on Image Processing*, 2023.

[71] W. Zhong, C. Xia, D. Zhang, and J. Han, "Uncertainty modeling for gaze estimation," *IEEE Transactions on Image Processing*, 2024.

[72] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, 2021.

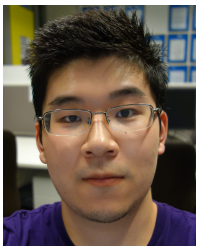
[73] L. Jin, X. Chen, J. Rückin, and M. Popović, "Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering," in *IROS*, 2023.

[74] S. Sabour, S. Vora, D. Duckworth, I. Krasin, D. J. Fleet, and A. Tagliasacchi, "Robustnerf: Ignoring distractors with robust losses," in *CVPR*, 2023.

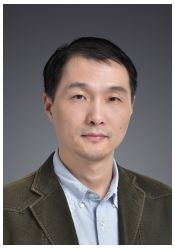
[75] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *CVPR*, 2021.

[76] J. Shen, A. Agudo, F. Moreno-Noguer, and A. Ruiz, "Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification," in *ECCV*, 2022.
[77] J. Shen, A. Ruiz, A. Agudo, and F. Moreno-Noguer, "Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations," in *3DV*, 2021.
[78] W. Sun, Q. Zhang, Y. Zhou, Q. Ye, J. Jiao, and Y. Li, "Uncertainty-guided optimal transport in depth supervised sparse-view 3d gaussian," *arXiv:2405.19657*, 2024.
[79] S. Zhang, B. Ye, X. Chen, Y. Chen, Z. Zhang, C. Peng, Y. Shi, and H. Zhao, "Drone-assisted road gaussian splatting with cross-view uncertainty," *arXiv:2408.15242*, 2024.
[80] Z. Tan, X. Chen, J. Zhang, L. Feng, and D. Hu, "Uncertainty-aware normal-guided gaussian splatting for surface reconstruction from sparse image sequences," *arXiv preprint arXiv:2503.11172*, 2025.
[81] F. Guo, C.-C. Hsu, S. Ding, and C. Zhang, "Uncertainty matters in dynamic gaussian splatting for monocular 4d reconstruction," in *ICLR*, 2026.
[82] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *CVPR*, 2023.
[83] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.
[84] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.
[85] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *NeurIPS*, 2022.
[86] A. Engelhardt, M. Boss, V. Voleti, C.-H. Yao, H. Lensch, and V. Jampani, "Svim3d: Stable video material diffusion for single image 3d generation," in *ICCV*, 2025.
[87] G. Zhang, H.-a. Gao, Z. Jiang, H. Zhao, and Z. Zheng, "Ctrl-u: Robust conditional image generation via uncertainty-aware reward modeling," in *ICLR*, 2025.
[88] W. Jiang, B. Lei, and K. Daniilidis, "Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information," *arXiv:2311.17874*, 2023.
[89] X. Pan, Z. Lai, S. Song, and G. Huang, "Activenerf: Learning where to see with uncertainty estimation," in *ECCV*, 2022.
[90] L. Savant, D. Valsesia, and E. Magli, "Modeling uncertainty for gaussian splatting," *arXiv:2403.18476*, 2024.
[91] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.
[92] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.
[93] W. Hoeffding, *Probability Inequalities for sums of Bounded Random Variables*. Springer New York, 1994.
[94] D. Ahn, H. Cho, J. Min, W. Jang, J. Kim, S. Kim, H. H. Park, K. H. Jin, and S. Kim, "Self-rectifying diffusion sampling with perturbed-attention guidance," in *ECCV*, 2024.
[95] Z.-X. Zou, Z. Yu, Y.-C. Guo, Y. Li, D. Liang, Y.-P. Cao, and S.-H. Zhang, "Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers," in *CVPR*, 2024.

Jiacheng Wang received the B.E. degree in electronic and information engineering, in 2022 from the Huazhong University of Science and Technology, Wuhan, China, where he is currently working toward the M.S. degree with the School of Electronic Information and Communications. His research interests include image generation and manipulation, computer vision, and machine learning.



Zhedong Zheng (Senior Member, IEEE) is an Assistant Professor with the University of Macau. He received the Ph.D. degree from the University of Technology Sydney in 2021 and the B.S. degree from Fudan University in 2016. He was a postdoctoral research fellow at the School of Computing, National University of Singapore. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He served as the senior PC for IJCAI and AAAI, and the area chair for ACM MM'24, ACM MM'25 and ICASSP'25.



Wei Xu (Member, IEEE) received the Ph.D.degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008. He is currently an Associate Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology. His research interests include automatic singing, multimedia, and machine learning.

Ping Liu (Senior Member, IEEE) is an Assistant Professor with the University of Nevada, Reno. He served as a Senior Research Scientist at the Center for Frontier AI Research (CFAR) under A*STAR in Singapore. Prior to his tenure at CFAR, he was affiliated with the Center for Artificial Intelligence at the University of Technology Sydney. Dr. Liu earned his Bachelor's degree from Wuhan University of Technology, his Master's from Huazhong University of Science and Technology, both in Wuhan, China, and his Ph.D. from the Department of Computer Science and Engineering at the University of South Carolina, Columbia, SC, USA. He specializes in computer vision, machine learning, and deep learning methodologies. He served as the senior PC for AAAI, and the area chair for ACM MM 2022-2025, IJCAI 2025.