Hallo3D: Multi-Modal Hallucination Detection and Mitigation for Consistent 3D Content Generation

Hongbo Wang[1,2] **Jie Cao**[1,2] **Jin Liu**[1,3] **Xiaoqiang Zhou**[1,4] **Huaibo Huang**[1,2]* **Ran He**[1,2]

[1]MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]School of Information Science and Technology, ShanghaiTech University, Shanghai, China
[4]University of Science and Technology of China, Hefei, China
wanghongbo2024@ia.ac.cn, jie.cao@cripac.ia.ac.cn, liujin2@shanghaitech.edu.cn
xq525@mail.ustc.edu.cn, huaibo.huang@cripac.ia.ac.cn, rhe@nlpr.ia.ac.cn

Abstract

Recent advances in pretrained 2D diffusion models have significantly improved visual prior guidance for 3D content generation. However, this process often lacks geometric constraints, leading to spatial perception hallucinations and multi-view inconsistencies. To address this, we introduce **Hallo3D**, a tuning-free method for 3D content generation that leverages the geometric perception capabilities of large multi-modal models to detect and mitigate these hallucinations. Our approach follows a generation-detection-correction paradigm, using multi-modal inconsistencies as query information to guide the detection of hallucinations and formulate enhanced negative prompts that ensure consistent renderings. Additionally, we propose a denoising strategy that employs attention mechanisms to maintain consistency in color and texture across multiple views during visual guidance. Our method is data-independent, easily integrates with existing 3D content generation frameworks, and supports both text-driven and image-driven approaches. Extensive experiments demonstrate that our method significantly improves the consistency and quality of generated 3D content, particularly in mitigating hallucinations common with 2D pretrained models.

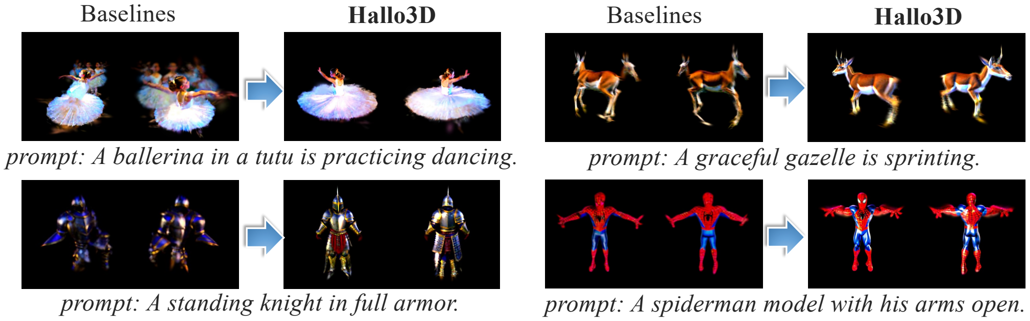
1 Introduction

Recent studies on 3D content generation have made significant progress, emerging as a central research focus in computer vision and computer graphics. The approaches for 3D content generation can be categorized into two primary categories: those based on 2D priors and those based on 3D priors. The strategies utilizing 2D priors typically learn 3D representations by approximating the probability distribution of 2D rendered images relative to a pre-trained diffusion model. This approximation is achieved during the visual guidance phase through a sophisticated optimization technique known as Score Distillation Sampling (SDS) [40].

However, methods based on 2D priors often suffer from overfitting to specific viewpoints of rendered images, resulting in generated 3D content that deviates from the expected distribution [2]. This overfitting leads to spatial perception inaccuracies, such as the Janus problem, where the generated objects display implausible duplications of features like faces or limbs, as depicted in Fig.1. An intuitive would be to learn priors from high-quality 3D data [19, 36, 50, 65, 62, 29, 18]. However, the limited availability and the often sparse supervision of 3D data pose significant challenges to maintaining view consistency and enhancing the generalizability of the generated content [23, 21]. Moreover, due to 3D

[*] Huaibo Huang is the corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



prompt: A ballerina in a tutu is practicing dancing.

prompt: A graceful gazelle is sprinting.

prompt: A standing knight in full armor.

prompt: A spiderman model with his arms open.

Figure 1: 3D Content Generation Results between Hallo3D (ours) and Baseline Model. Hallo3D can effectively solve the "Janus" problem and improve the multi-view consistency of the 3D generation.

generation tasks spanning a diverse array of domains, the scalability of data-driven models remains markedly constrained, limiting their applicability across a comprehensive spectrum of potential uses.

To mitigate the hallucination problem and ensure view-consistent generation, we leverage the large multi-modal models to infer and adjust the geometric structures of the generated content. These models can recognize spatial relationships and evaluate the structural consistency of visual contexts by interpreting 3D elements such as lighting and proportion from 2D renderings. Building on this observation, we propose a novel generation-detection-correction paradigm. In this paradigm, we utilize multi-modal models to refine rendered images, ensuring visually coherent results. Our strategy improves cross-view consistency without relying on the prompt provided for diffusion guidance, thereby bridging text-driven and image-driven methods.

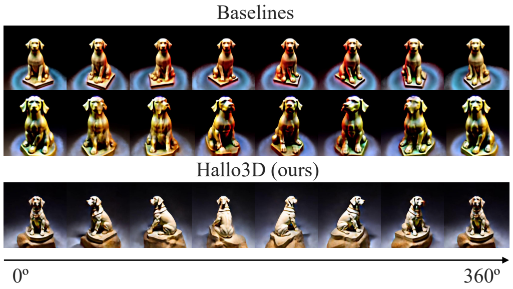


Figure 2: **Illustration of the Janus problem.** The first two rows show overfitting with repeated frontal views, while the third row, using Hallo3D, achieves more consistent results. This highlights the issue and clarifies the expected outcome.

In light of the findings above, we present **Hallo3D**, a novel, tuning-free approach that significantly enhances the multi-view consistency of 3D generation and is applicable across various generation methods. Our approach comprises three core techniques. Multi-Modal Hallucination Detection: This technique detects concrete hallucinations in renderings by leveraging large multi-modal models to represent inconsistency query information. Prompt-Enhanced Reconsistency: Utilizing the detection results from Multi-Modal Hallucination Detection as an enhanced negative prompt to precisely eliminate inconsistent artifacts in the renderings. Multi-view Appearance Alignment: This technique ensures uniform color and texture in renderings from different viewpoints by controlling attention in the diffusion de-noising process. By integrating these techniques, Hallo3D effectively addresses the challenges of maintaining consistency in 3D generation, providing a robust solution applicable to various generative methods. Experimental results demonstrate that our method exhibits a significant advantage in multi-view consistency compared to baseline models and can be robustly applied to various 3D generation tasks.

Our contributions can be summarized as follows:

- We propose Hallo3D, a novel tuning-free method that significantly enhances the multi-view consistency of 3D content generation and can be widely applied across various 3D generation paradigms, achieving outstanding experimental results.
- We demonstrate that large multi-modal models, unconstrained by geometry, can infer geometric structures and be utilized to detect and mitigate hallucinations in 3D generation.
- We introduce an optimization strategy that aligns the structures and surfaces of 3D content across views and addresses artifacts and hallucination through enhanced prompts.

2 Related Work

Text-to-3D Generation. The evolution of diffusion models has markedly enhanced text-to-3D generation. DreamFusion [40] initiated text-guided 3D modeling by using visual priors from 2D diffusion models to train 3D architectures, incorporating MipNeRF 360 [3] and Imagen [47]. While NeRF-based methods [33, 32, 48, 66, 59, 22, 12, 24, 10] handle complex lighting well, they are slower due to continuous parameter updates. In contrast, 3D Gaussian Splatting (3DGS) methods [20, 56, 5, 63, 39] have improved rendering speeds, showing their efficacy in complex scenarios.

Image-to-3D Generation. Images from specific views of a 3D model demonstrate improved visual consistency for 3D generation tasks. Image-based methods [53, 11, 30, 54, 1, 41] typically surpass text-based approaches by leveraging viewpoint-specific ground truth. 3D-aware image generation techniques [61, 7] utilize neural networks to enhance rendering beyond the primary viewpoint, although training data scarcity [23, 21] remains a challenge. Recent strides in integrating 3D visual data into 2D diffusion models [19, 36] have notably enhanced image-based 3D generation, reducing perceptual errors and improving generative quality.

Methods for Enhancing Multi-view Consistency. The primary challenge in enhancing 3D generation consistency is addressing hallucinations from 2D pre-trained diffusion models. A common strategy involves integrating additional 3D information into the diffusion process via fine-tuning [50, 65, 62, 29, 18]. This includes training models to handle consistent 3D subjects [46, 44], transparent backgrounds [64], and diverse viewpoints [49]. Recent advancements have also adjusted prompt embeddings to enhance viewpoint accuracy [17, 2], although this remains limited for non-orthogonal views. Alternatives include using geometric methods [25] or treating 3D generation analogously to video generation [57], though these tend to be framework-specific. In contrast, our method effectively optimizes arbitrary viewpoints, making it versatile across different 3D frameworks.

3 Methodology

3.1 Preliminaries

Diffusion Models. The diffusion model [15] has a fozrward diffusion process with diffusion steps from 0 to T, which degrades the original sample $\mathbf{x}_0$ into pure noise $\mathbf{x}_T$,

$$\mathbf{x}_t = \sqrt{\alpha_\mathbf{t}}\mathbf{x}_0 + \sqrt{1-\alpha_\mathbf{t}}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{1}$$

where t is the noise injection level, and $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_T) \in \mathbb{R}^T_{\geq 0}$ are hyper-parameters to determine noise scales at T diffusion steps, and the reverse diffusion process is used during inference to generate $\mathbf{x}_0$ from $\mathbf{x}_t$. In text-guided diffusion models [52], the model is conditioned on text prompts P, which are converted into text embeddings via a text encoder such as CLIP [43]. The diffusion model $\boldsymbol{\epsilon}_\phi$ is trained using the MSE loss between the predicted noise $\hat{\epsilon}_\phi$ and the actual noise $\boldsymbol{\epsilon}$,

$$L(\phi) = \mathbb{E}_{t\sim U(1,T),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, P)\|_2^2,$$

where $U(1,T)$ represents a uniform distribution over the set $\{1, \cdots, T\}$, and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. To enhance the alignment between text and images, Classifier-free guidance (CFG) [16] guides the generation of samples using

$$\hat{\boldsymbol{\epsilon}}_\phi(\mathbf{x}_t, t, P, \emptyset) = \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, \emptyset) + s\left(\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, P) - \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t, \emptyset)\right), \tag{2}$$

where $\emptyset$ is a special null text prompt representing the unconditional case, and $s > 0$ is the guidance scale. Increasing the guidance scale improves text-image alignment but reduces diversity. In practice, the $\emptyset$ text prompt is replaced with a negative prompt P^- consisting of negative descriptions [9] to avoid undesired content in the generated samples.

Score Distillation Sampling. Score Distillation Sampling (SDS) employs vision priors from pre-trained 2D diffusion models to supervise 3D models, establishing it as a foundational learning method in the domain of 3D generation, initially proposed by DreamFusion [40]. A 3D representation model, parameterized by θ, and the pre-trained diffusion model $\boldsymbol{\epsilon}_\phi$, together enable the rendering of an image $\mathbf{x} = g(\theta, \mathbf{c})$ from the 3D content, where $g(\cdot)$ is a differentiable generator to render $\mathbf{x}$ and $\mathbf{c}$ is the camera pose. To ensure that $\mathbf{x}$ consistently exhibits high quality from any view and to align the probability of $\mathbf{x}$ with $p(\phi)$, SDS introduces a score estimation function $\hat{\epsilon}_\phi(\mathbf{x}_t; t, P)$. This function

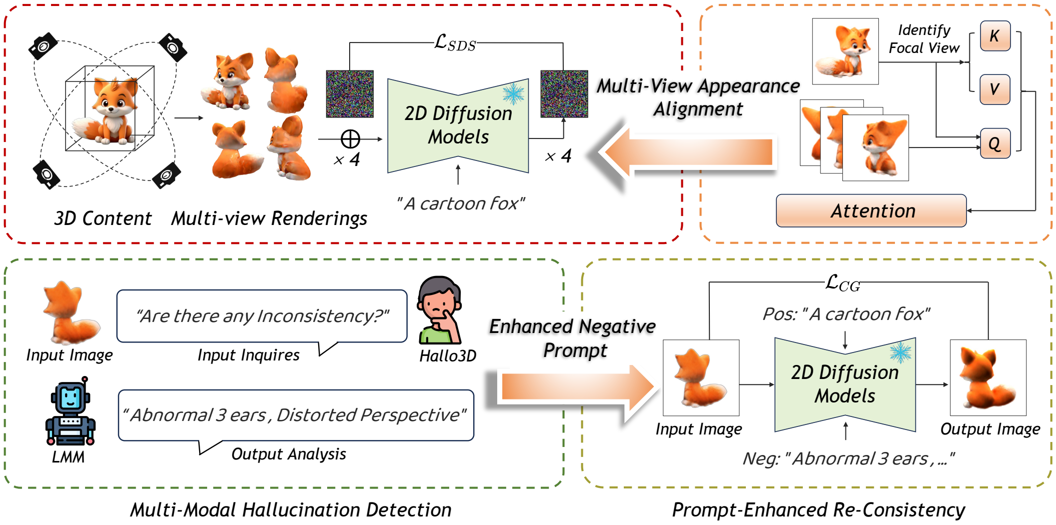


Figure 3: **Illustration of our pipeline.** We jointly optimize our model using $\mathcal{L}_{\mathrm{SDS}}$ and $\mathcal{L}_{\mathrm{CG}}$. For $\mathcal{L}_{\mathrm{SDS}}$, we identify a focal view from multi-view renderings based on the camera pose, utilizing it as the keys (K) and values (V) to align all the four images using attention. This process harmonizes the appearance and feeds the output into the 2D Diffusion on the left, which plays a crucial role in refining the noise prediction. For $\mathcal{L}_{\mathrm{CG}}$, we query hallucinations and inconsistencies in the rendering using an LMM and apply the results, outputted as enhanced negative prompt, to the following image optimization process to re-consistent a high-quality image. We calculate the $\mathcal{L}_{\mathrm{CG}}$ based on the differences between the two images, thereby enhancing the consistency of the 3D content.

predicts the noise $\hat{\epsilon}_\phi$ based on the text condition P and the noisy image $\mathbf{x}_t$, into which Gaussian noise ϵ has been injected. Furthermore, by calculating the discrepancy between $\hat{\epsilon}_\phi$ and ϵ, the score function identifies the gradient direction for updating parameter θ, thereby enhancing the training of the 3D model. The specific computation of the gradient is as follows:

$$\nabla_\theta \mathcal{L}_{\mathrm{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi(\mathbf{x}_t; P, t) - \epsilon\right) \frac{\partial \mathbf{x}}{\partial \theta} \right], \tag{3}$$

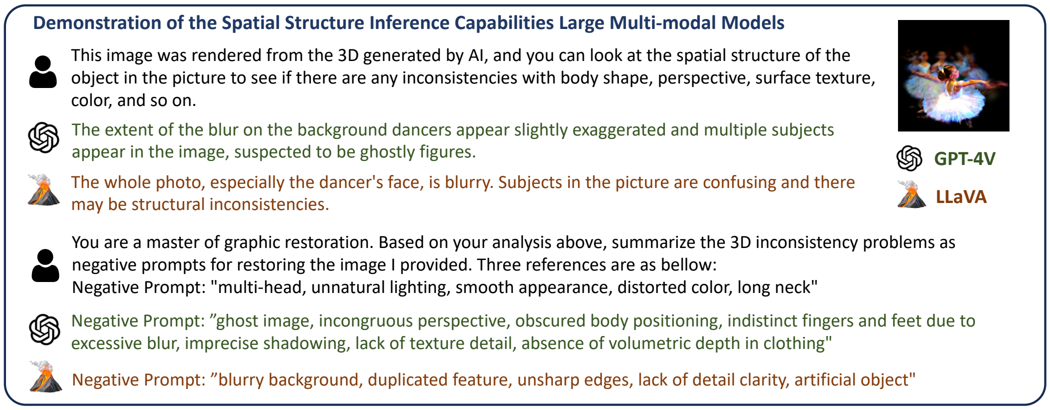
where $w(t)$ is a weighting function.

3.2 Multi-view Appearance Alignment

Building upon our understanding of the SDS as discussed in Sec. 3.1, we further explored the impact of SDS on the consistency of appearances in 3D generation. We observed that SDS processes images from only one view at a time, which contradicts our intuition that enhancing 3D multi-view consistency should involve the simultaneous processing of multiple viewpoint images. Experimental results demonstrate that this approach led to a lack of interaction between images from different views during the training process, resulting in the loss of some surface information during the noise prediction process, as shown in Fig. 7 in the ablation study.

To circumvent the limitations of SDS, which typically favors image generation from a single view, we propose a "Multi-view Appearance Alignment" strategy. This approach introduces a consistent denoising method $\tilde{\epsilon}_\phi(\cdot)$ that incorporates an attention mechanism $\mathrm{AAttn}(\cdot)$, enabling the rendering of multiple images from random views and providing a broader perspective compared to techniques focused primarily on single-view image generation [40, 58].

Specifically, inspired by recent advancements in diffusion models [27, 14, 4, 55, 35], which suggest that query features within attention spaces primarily shape image structure and layout, while key and value features influence texture, our method leverages this insight. As illustrated in the top right corner of Fig. 3, we select a focal view i based on the camera pose, using the image from this viewpoint to provide the key and value features in the attention module. These are used to compute query features across all views, ensuring alignment of appearances. The attention is defined by the



Demonstration of the Spatial Structure Inference Capabilities Large Multi-modal Models

This image was rendered from the 3D generated by AI, and you can look at the spatial structure of the object in the picture to see if there are any inconsistencies with body shape, perspective, surface texture, color, and so on.

The extent of the blur on the background dancers appear slightly exaggerated and multiple subjects appear in the image, suspected to be ghostly figures.

The whole photo, especially the dancer's face, is blurry. Subjects in the picture are confusing and there may be structural inconsistencies.

You are a master of graphic restoration. Based on your analysis above, summarize the 3D inconsistency problems as negative prompts for restoring the image I provided. Three references are as bellow:
Negative Prompt: "multi-head, unnatural lighting, smooth appearance, distorted color, long neck"

Negative Prompt: ”ghost image, incongruous perspective, obscured body positioning, indistinct fingers and feet due to excessive blur, imprecise shadowing, lack of texture detail, absence of volumetric depth in clothing"

Negative Prompt: ”blurry background, duplicated feature, unsharp edges, lack of detail clarity, artificial object"

Figure 4: **A multi-modal case study** for evaluating the capabilities of LMMs in 3D generation tasks. The first round of dialogue demonstrates that LMMs can infer structural consistency from 3D rendered images, while the second round shows that LMMs can respond in specific formats, allowing us to subsequently identify the negative prompts output using regular expressions.

following formula:

$$\mathrm{AAttn}(Q, K_i, V_i) = \mathrm{Softmax}\left(\frac{QK_i^T}{\sqrt{d}}\right) \cdot V_i, \tag{4}$$

where $\mathrm{AAttn}(\cdot)$ is the appearance attention, with K_i and V_i as the key and value features corresponding to the image rendered from the focal view i, and Q as the query feature from all views. The key and value are derived from the focal view, while each of the four views calculates a distinct query. In the denoising strategy $\tilde{\epsilon}_\phi(\cdot)$, this attention mechanism functions as cross-attention, aligning features from all views with the focal view to ensure consistent appearances. This process occurs within the U-Net network [45] in $\tilde{\epsilon}_\phi(\cdot)$, prior to the cross attention with the prompt.

3.3 Multi-modal Hallucination Detection

As shown in Fig. 4, the rendering image in the top right corner of the figure exhibits significant inconsistencies, due to the limitations of 2D pre-trained models in comprehending spatial concepts. This often leads to hallucinations and overfitting to specific viewpoints. However, we believe that Large Multi-modal Models (LMMs) have the capability to reason about and mitigate these hallucinations. To demonstrate this, we designed a two-phase inquiry involving LMMs, specifically using high-performing GPT-4V [38] and LLaVA [26] as examples. The dialogue depicted in the figure indicates that although LMMs were not explicitly trained with geometric constraints, they could identify inconsistencies in the 3D renderings and categorize them as negative prompts. Additionally, LMMs can standardize their output format based on a one-shot reference, making it easier for us to extract negative prompts.

Specifically, in our model, we input one 2D rendered image alongside 3D-aware inquiry prompts, denoted as P_I, into the multi-modal large modal to assist in automatically identifying inconsistencies present during the 3D generation process. To further mitigate hallucinations and correct inconsistencies, we have standardized the output format of the LMM, enabling it to accurately generate negative prompts based on the provided shots. These negative prompts can then be used to rectify distorted images in subsequent steps. Given their effectiveness in purposefully addressing inconsistencies, we refer to them as "Enhanced Negative Prompts". We formalize this process as follows:

$$P_E^- = \boldsymbol{D}_\psi(\mathbf{x}, P_I), \tag{5}$$

where $\boldsymbol{D}_\psi$ is the LMM parameterized by ψ, and P_E^- is the enhanced negative prompt.

3.4 Prompt-Enhanced Re-consistency

With the enhanced negative prompt P_E^- introduced in Sec. 3.3, a straightforward method to refine 2D renderings involves employing image editing algorithms to address inconsistencies. However, existing approaches predominantly focus on adjustments to the null prompt [34] or the modification of positive prompts [14], which are generally ineffectual for altering the geometric structures in 2D images derived from 3D models. To address this limitation, we introduce a novel module for achieving re-consistency in 2D renderings, termed “Prompt-Enhanced Re-consistency,” which leverages P_E^- to effectively refine the geometric fidelity of the rendered images.

We regenerate the 2D rendered image $\mathbf{x}_0$ under the guidance of P_E^-. Specifically, to preserve the original semantic information of $\mathbf{x}_0$, we employ Denoising Diffusion Implicit Models (DDIM) [51] to invert theimage $\mathbf{x}_0$ to its noisy representation $\mathbf{x}_T$. Subsequently, we apply DDIM sampling to generate the consistent versions of the image, denoted as $\hat{\mathbf{x}}_0$, from $\mathbf{x}_T$ as follows:

$$\hat{\mathbf{x}}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\hat{\mathbf{x}}_t + (\sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}})\tilde{\epsilon}_\phi(\hat{\mathbf{x}}_t, t, P^+, P_E^-) \tag{6}$$

where $\tilde{\boldsymbol{\epsilon}}_\phi(\hat{\mathbf{x}}_t, t, P^+, P_E^-)$ is the denoising strategy incorporated the attention mechanism Attn$(\cdot)$ in Sec. 3.2, and is adjusted by Classifier-Free Guidance (CFG) [16], with the null text prompt replaced by the enhanced negative prompt P_E^-. This approach ensures that the regenerated image retains its core semantic integrity while improving its multi-view consistency. After completing T iterations as delineated by Eq. 1, we successfully achieve the re-consistent image $\hat{\mathbf{x}}_0$, effectively reconciling the image consistency with its original semantic information.

Finally, we train the 3D model θ using the MSE loss $\mathcal{L}_{\mathrm{CG}}$ between $\mathbf{x}_0$ and $\hat{\mathbf{x}}_0$ in the image space:

$$\mathcal{L}_{\mathrm{CG}} \triangleq \mathbb{E}\left[(\hat{\mathbf{x}}_0 - \mathbf{x}_0)\right], \tag{7}$$

It is worth noting that we apply Prompt-Enhanced Reconsistency only when the rendered image exhibit complete semantic structure. Our detector, $\boldsymbol{D}_\psi$, assesses the semantic completeness of the image. If the semantic structure is deemed incomplete or unclear, $\boldsymbol{D}_\psi$ returns None, precluding further processing. This ensures that enhancements are only applied to images that are adequately prepared. The dependency of our enhancement process on the state of semantic completeness directly influences the formulation of the final training loss for the 3D model, as detailed below:

$$\mathcal{L}(\theta) = \begin{cases} \mathcal{L}_{\mathrm{SDS}} + w\mathcal{L}_{\mathrm{CG}}, & \text{if } \boldsymbol{D}_\psi(\mathbf{x}, P_I) \text{ is not None,} \\ \mathcal{L}_{\mathrm{SDS}}, & \text{otherwise.} \end{cases} \tag{8}$$

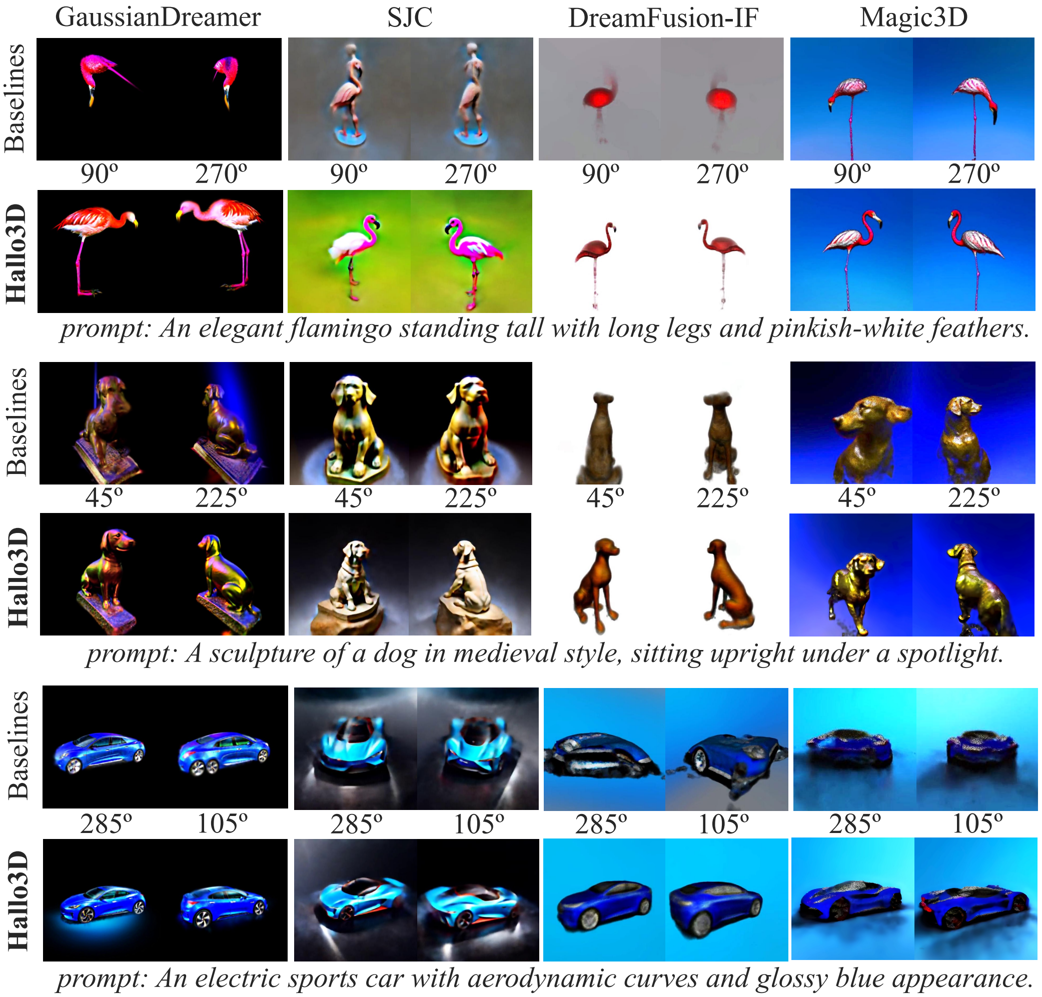
where w is set to balance the magnitude of $\mathcal{L}_{\mathrm{SDS}}$ and $\mathcal{L}_{\mathrm{CG}}$. By incorporating $\mathcal{L}_{CG}$, which is only applied when D_ψ confirms the semantic readiness of the image, we ensure that our model focuses on enhancing well-formed images. This selective application of $\mathcal{L}_{CG}$ prevents further exacerbating the quality of images already of poor quality. Simultaneously, it avoids misallocating resources to images that do not benefit from the intended enhancements, thereby improving the efficiency and effectiveness of our training process. For more implementation details, see the Appendix.A.

4 Expremients

In this section, we comprehensively evaluate Hallo3D’s performance within two categories of 3D generation frameworks: text-to-3D and image-to-3D. We present comparative results with other baseline models to highlight its capabilities. Additionally, to further substantiate the effectiveness of Hallo3D in enhancing multi-view consistency in 3D generation, we have conducted an extensive user study. Finally, we designed ablation experiments to validate the necessity of the framework’s design.

4.1 Experiment Setup

Baselines. We evaluated our method against several established baselines, demonstrating strong performance across diverse frameworks. These include text-to-3D models like GaussianDreamer [63], Score Jacobian Chain (SJC) [58], DreamFusion-IF [40], and Magic3D [24], as well as image-to-3D models such as DreamGaussian [53] and Zero-1-to-3 [28], an extension of DreamFusion. We also included methods based on NeRF [33] and 3DGS [20] for a comprehensive comparison. Identical parameter configurations and seed values were maintained for fair comparison, using default hyperparameters from the baselines’ open-source implementations. We employed the Threestudio library [13] for SJC and Magic3D, and the official codebases for the other methods. Additionally, we conducted experiments to evaluate the time consumption of Hallo3D, detailed in the Appendix.B.



prompt: An elegant flamingo standing tall with long legs and pinkish-white feathers.

prompt: A sculpture of a dog in medieval style, sitting upright under a spotlight.

prompt: An electric sports car with aerodynamic curves and glossy blue appearance.

Figure 5: **Qualitative comparison in text-driven 3D generation** of HalloD and baseline models. To provide a more straightforward comparison, we rendered both Hallo3D and the baseline models from two identical and complementary angles.

Table 1: Quantitative comparisons in text-driven 3D generation

Metrics	GaussianDreamer	**Hallo3D**	SJC	**Hallo3D**	DreamFusion-IF	**Hallo3D**	Magic3D	**Hallo3D**
CLIP-Score B/32 ↑	21.31	**24.53**	20.13	**24.34**	14.09	**22.15**	14.93	**22.05**
CLIP-Score B/16 ↑	22.67	**27.00**	21.36	**26.36**	15.98	**23.79**	16.41	**24.29**
CLIP-Score L/14 ↑	23.70	**30.12**	23.95	**28.04**	18.19	**26.72**	17.99	**27.72**

Table 2: User study in text-driven 3D generation

Metrics	GaussianDreamer	**Hallo3D**	SJC	**Hallo3D**	DreamFusion-IF	**Hallo3D**	Magic3D	**Hallo3D**
Multi-view Consistency ↑	6.00	**8.87**	4.53	**7.63**	4.63	**6.33**	5.13	**7.53**
Overall Quality ↑	5.53	**8.67**	4.77	**7.80**	4.17	**7.37**	4.60	**8.03**
Alignment with Prompt ↑	5.57	**8.87**	5.63	**7.40**	4.70	**7.03**	5.17	**7.37**

Metrics. The field of 3D generation struggles with the absence of ground truth, complicating the development of a unified evaluation metric. To address multi-view consistency, we reviewed existing evaluation methods and identified 3D inconsistencies using CLIP-Score [43]. We generated 80 unique 3D prompts using ChatGPT [37] and arranged 16 cameras in a 360-degree configuration around the z-axis. The average CLIP-Score across all views measured consistency.

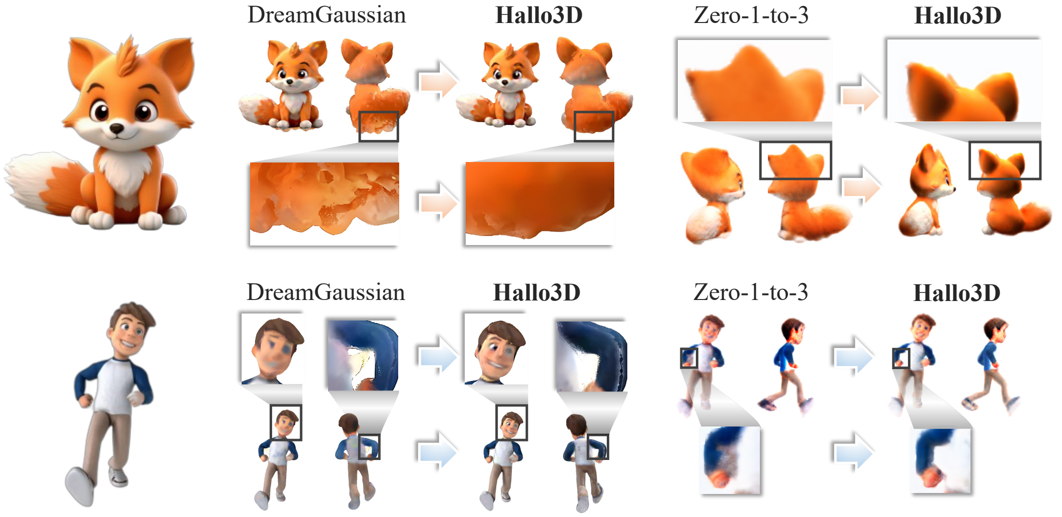


Figure 6: **Qualitative comparison in image-driven 3D generation** of Hallo3D and baseline models. To facilitate a more direct comparison, we rendered both Hallo3D and the baseline models from two complementary angles and magnified specific details.

Table 3: User study in image-driven 3D generation

Metrics	DreamGaussian	**Hallo3D**	Zero-1-to-3	**Hallo3D**
Multi-view Consistency ↑	8.40	**9.15**	7.25	**7.81**
Overall Quality ↑	9.23	**9.52**	6.10	**7.20**
Alignment with Prompt ↑	8.55	**8.00**	8.30	**8.95**

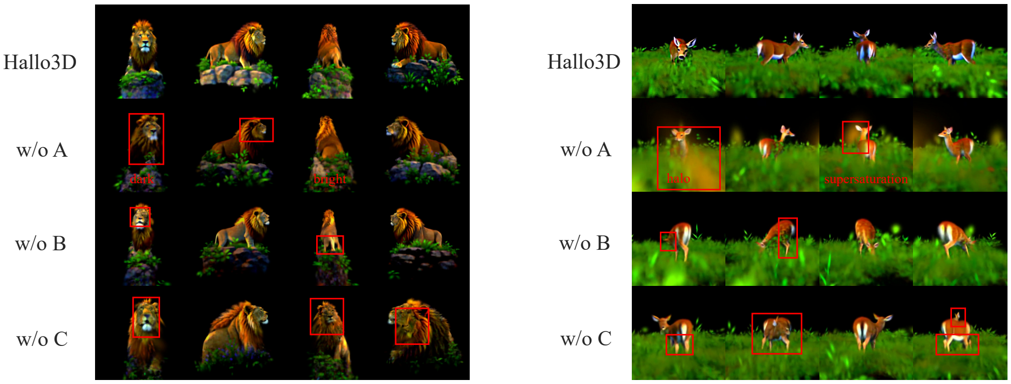
4.2 Quantitative Comparison with Baselines

In our qualitative evaluation for text-driven 3D content generation, we randomly selected three prompts from a dataset of 80 and used two high-definition images from Google Images for image-driven 3D generation. The results, shown in Fig. 5 and Fig. 6, reveal significant enhancements in multi-view consistency. Baseline models often produced flawed figures, such as headless "flamingos" or "dogs" with multiple heads and ears. In contrast, our models achieved more realistic and consistent outputs, confirming the effectiveness of our approach. The 360-degree visualization is shown in Appendix.C.

4.3 Qualitative Comparison with Baselines

Computational results. Following [63, 42, 53], we evaluated the CLIP-Score to assess the quality and consistency of 3D generated contents, as presented in Tab.1. The results indicate that our approach outperforms all baseline models, confirming the effectiveness of our method. It should be noted that the existence of a ground truth corresponding to the front view in image-driven 3D generation generally leads to higher generation quality.

Consequently, for image-3D tasks, we adhered to the experimental setup outlined in [31, 60]. Specifically, we selected 60 objects from the GSO [8] and Objaverse [6] datasets, replacing overly simple objects to ensure a more robust evaluation. These objects were rendered in frontal views at a resolution of 256x256. To comprehensively assess performance, we utilized Chamfer Distance (CD) and Volume IoU (Vol. IoU) for evaluating geometric accuracy, along with PSNR, SSIM, and LPIPS for measuring visual quality. As presented in Tab.4, the experimental results clearly indicate that our method surpasses the baseline across all metrics, achieving significant improvements in both geometry and texture quality. This further substantiates the broad applicability of our approach, demonstrating its capacity to enhance both text-to-3D and image-to-3D tasks.



prompt: A majestic lion standing on a rock.

prompt: A deer grazing on a grassy plain.

Figure 7: **Ablation study of our method.** In the figure, module A represents Multi-view Appearance Alignment in Sec. 3.2, module B stands for Multi-modal Hallucination Detection in Sec. 3.3, and module C denotes Prompt-Enhanced Re-Consistency in Sec. 3.4. We conducted ablation studies on each of these three modules respectively.

Table 4: Quantitative comparisons in image-driven 3D generation

Metrics	DreamGaussian	**Hallo3D**	Zero-1-to-3	**Hallo3D**
CD↓	0.0185	**0.0171**	0.0370	**0.0283**
Vol. IoU↑	0.5861	**0.6099**	0.4824	**0.5602**
PSNR↑	16.502	**16.518**	13.433	**14.930**
SSIM↑	0.8543	**0.8793**	0.7210	**0.7527**
LPIPS↓	0.2025	**0.1726**	0.3926	**0.3328**

User study. In our user study, we recruited 58 volunteers with expertise in artificial intelligence to participate in our experiment. To comprehensively assess the quality discrepancies among various generated 3D models, we developed an extensive scale for the volunteers to fill out. Specifically, we generated 120-frame videos for each 3D model, totaling 32 video sets. Our comparative approach involved evaluating each model independently on three criteria: "Multi-view Consistency," "Overall Quality," and "Alignment with Prompt," with ratings on a scale from 1 to 10. The findings were summarized by compiling the average scores, and can be seen in Tab.2 and Tab.3.

4.4 Ablation Study

We conducted ablation experiments on the three Hallo3D modules, as shown in Fig. 7. Starting from the complete model, we independently removed each module and assessed their effects. Notably, in the "w/o C" setting, the output of LMM, P_E^- is applied to $\mathcal{L}_{\text{SDS}}$ calculations to demonstrate the necessity of $\mathcal{L}_{\text{CG}}$. Additionally, we conducted an ablation on "w/o C & P_E^-", where P_E^- is not applied anywhere in the "w/o C" setting, to further highlight the effectiveness of the module.

In Fig.7, we focus on how module A primarily affects color and texture, while module B and module C enhance cross-view consistency. Specifically,

- In Row 2, w/o A: The lion appears significantly darker than in Row 1, and the deer exhibits a blurry halo accompanied by an unnatural color shift.
- In Row 3, w/o B: The lion's head is noticeably deformed in both the first and third columns, while the deer entirely loses its head.
- Row 4, w/o C: The "second face" appears on the lion's back in the third column, and the right deer's image shows a clear Janus Problem, with multiple legs and a distorted body visible in the second and fourth columns.

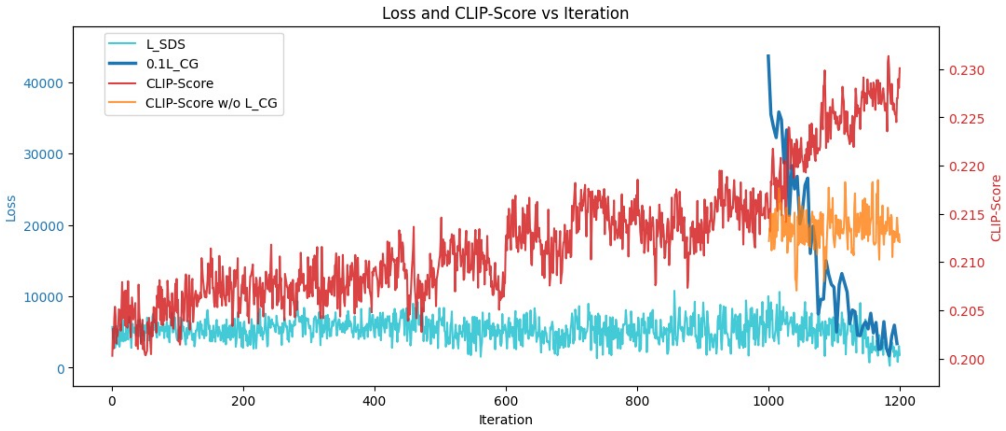


Figure 8: Loss curves for $\mathcal{L}_{\text{CG}}$ and $\mathcal{L}_{\text{SDS}}$, along with the CLIP-Score curves with and without $\mathcal{L}_{\text{CG}}$.

Table 5: Quantitative Ablation Results

Metrics	Hallo3D	w/o A	w/o B	w/o C	w/o C & P_E^-	Baseline
CLIP-Score B/32↑	24.25	23.98	23.65	22.46	22.23	21.27
CLIP-Score B/16↑	26.83	25.88	25.10	23.59	23.23	22.67
CLIP-Score L/14↑	30.00	29.36	28.71	26.92	25.58	23.71

Additionally, we conducted quantifiable ablation experiments, as shown in Tab.4.4, further demonstrating the effectiveness and necessity of each module in Hallo3D. Moreover, the better performance of "w/o C" compared to "w/o C & P_E^-" also supports the necessity of introducing $\mathcal{L}_{\text{CG}}$.

4.5 Balance between $\mathcal{L}_{\text{CG}}$ and $\mathcal{L}_{\text{SDS}}$.

In our experiments, we observed that the loss function $\mathcal{L}_{\text{CG}}$, which is computed on a per-pixel basis, typically exhibits a larger magnitude in comparison to $\mathcal{L}_{\text{SDS}}$. To achieve a balanced scale between these losses, we assigned a weight of $w = 0.1$ to $\mathcal{L}_{\text{CG}}$ in Eq.8. It is important to note that this adjustment in weight does not diminish the importance of $\mathcal{L}_{\text{CG}}$ in any way. Specifically, as shown in Fig.8, even after scaling $\mathcal{L}_{\text{CG}}$ by the factor w, it maintains a larger magnitude compared to $\mathcal{L}_{\text{SDS}}$. This demonstrates that $\mathcal{L}_{\text{CG}}$ provides ample guidance for 3D generation, ensuring effective optimization throughout the process.

5 Conclusion

In this paper, we introduce Hallo3D, a novel approach designed to enhance 3D content generation through both text-driven and image-driven methods. We demonstrate the capability of large multimodal models to infer geometric structures and detect hallucination arising from 2D diffusion models. By combining LMM with diffusion models, we achieve re-consistent 2D images applicable in the 3D domain. Extensive experimental evidence substantiates that our method significantly improves consistency and mitigates hallucinations in 3D content generation. Additionally, we thank Haoyang Tong from the University of Chinese Academy of Sciences for his contributions to this work.

Acknowledgements

This research is supported in part by several funding sources, including the National Natural Science Foundation of China (Grant No. 32341009, 62206277, U21B2045, 62425606). Additional support has been provided by the Beijing Nova Program (20230484276), as well as the Youth Innovation Promotion Association of the Chinese Academy of Sciences (Grant No. 2022132). These contributions have been instrumental in enabling the advancement of this work.

References

[1] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-Image 3D Human Digitization with Shape-Guided Diffusion. *SIGGRAPH*, 2023.

[2] Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond. *arXiv:2304.04968*, 2023.

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *CVPR*, pages 5460–5469, New Orleans, LA, USA, 2022.

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *ICCV*, pages 22560–22570, 2023.

[5] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3D using Gaussian Splatting. *arXiv:2309.16585*, 2023.

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023.

[7] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *CVPR*, 2022.

[8] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022.

[9] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. In *NeurIPS*, volume 33, pages 6637–6647, 2020.

[10] Lincong Feng, Muyu Wang, Maoyu Wang, Kuo Xu, and Xiaoli Liu. MetaDreamer: Efficient Text-to-3D Creation With Disentangling Geometry and Texture. *arXiv:2311.10123*, 2023.

[11] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. COLMAP-Free 3D Gaussian Splatting. In *CVPR*, 2023.

[12] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. GraphDreamer: Compositional 3D Scene Synthesis from Scene Graphs. In *CVPR*, 2023.

[13] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. Threestudio. 2023.

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. In *ICLR*, 2022.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020.

[16] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS*, 2021.

[17] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing Scores and Prompts of 2D Diffusion for View-consistent Text-to-3D Generation. In *NeurIPS*, 2023.

[18] Yifan Jiang, Hao Tang, Jen-Hao Rick Chang, Liangchen Song, Zhangyang Wang, and Liangliang Cao. Efficient-3DiM: Learning a Generalizable Single-image Novel-view Synthesizer in One Day. In *ICLR*, 2023.

[19] Heewoo Jun and Alex Nichol. Shap-E: Generating Conditional 3D Implicit Functions. *arXiv:2305.02463*, 2023.

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4):1–14, 2023.

[21] Chenghao Li, Chaoning Zhang, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era. *arXiv:2305.06131*, 2023.

[22] Ming Li, Pan Zhou, Jia-Wei Liu, Jussi Keppo, Min Lin, Shuicheng Yan, and Xiangyu Xu. Instant3D: Instant Text-to-3D Generation. *IJCV*, 2023.

[23] Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. Advances in 3D Generation: A Survey. *arXiv:2401.17807*, 2024.

[24] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-Resolution Text-to-3D Content Creation. In *CVPR*, 2023.

[25] Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3D: Boosting High-Fidelity Text-to-3D Generation via Coarse 3D Prior. In *CVPR*, 2023.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *AAAI*, 2023.

[27] Jin Liu, Huaibo Huang, Chao Jin, and Ran He. Portrait diffusion: Training-free face stylization with chain-of-painting. *arXiv preprint arXiv:2312.02212*, 2023.

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot One Image to 3D Object. In *ICCV*, 2023.

[29] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *ICLR*, 2024.

[30] Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J. Black, and Bernhard Schölkopf. Ghost on the Shell: An Expressive Representation of General 3D Shapes. In *ICLR*, 2023.

[31] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024.

[32] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. ATT3D: Amortized Text-to-3D Object Synthesis. In *ICCV*, 2023.

[33] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2022.

[34] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *CVPR*, 2023.

[35] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. DreamMatcher: Appearance Matching Self-Attention for Semantically-Consistent Text-to-Image Personalization. In *CVPR*, 2024.

[36] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv:2212.08751*, 2022.

[37] OpenAI. ChatGPT.

[38] OpenAI. GPT-4 Technical Report. 2023.

[39] Hao Ouyang, Kathryn Heal, Stephen Lombardi, and Tiancheng Sun. Text2Immersion: Generative Immersive Scene with 3D Gaussians. *arXiv:2312.09242*, 2023.

[40] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*, 2022.

[41] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. In *ICLR*, 2023.

[42] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *CVPR*, 2024.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[44] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. DreamBooth3D: Subject-Driven Text-to-3D Generation. In *ICCV*, 2023.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*, 2023.

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*, 2022.

[48] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. DITTO-NeRF: Diffusion-based Iterative Text To Omni-directional 3D Model. *arXiv:2304.02827*, 2023.

[49] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. In *ICLR*, 2023.

[50] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view Diffusion for 3D Generation. In *ICLR*, 2024.

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2022.

[52] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 2022.

[53] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. In *ICLR*, 2023.

[54] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior. In *ICCV*, pages 22762–22772, Paris, France, 2023.

[55] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*, pages 1921–1930, 2023.

[56] Alexander Vilesov, Pradyumna Chari, and Achuta Kadambi. CG3D: Compositional Generation for Text-to-3D via Gaussian Splatting. *arXiv:2311.17907*, 2023.

[57] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion. *arXiv:2403.12008*, 2024.

[58] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *CVPR*, 2022.

[59] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. In *NeurIPS*, 2023.

[60] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv:2405.20343*, 2024.

[61] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3D-aware Image Generation using 2D Diffusion Models. In *ICCV*, 2023.

[62] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. ConsistNet: Enforcing 3D Consistency for Multi-view Images Diffusion. In *CVPR*, 2023.

[63] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. In *CVPR*, 2023.

[64] Lvmin Zhang and Maneesh Agrawala. Transparent Image Layer Diffusion using Latent Transparency. *arXiv:2402.17113*, 2024.

[65] Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Zhipeng Hu, Changjie Fan, and Xin Yu. EfficientDreamer: High-Fidelity and Robust 3D Creation via Orthogonal-view Diffusion Prior. In *CVPR*, 2023.

[66] Linqi Zhou, Andy Shih, Chenlin Meng, and Stefano Ermon. DreamPropeller: Supercharge Text-to-3D Generation with Parallel Sampling. In *CVPR*, 2023.

A Implementation details

Selection of focal view. We use Fovy, the camera's vertical view, as our selection standard. The first view with Fovy exceeding 120% of each baseline's default becomes our focal view, enabling a broader shot capturing more object details.

Details of LMM setup. We used different P_I than in Fig.4. The primary purpose of Fig.4 is to use a case study to demonstrate how LMMs can infer structural consistency and respond in specific formats. To highlight this capability, we employed two dialogues. In practice, we used a single interaction to query the LMM to achieve faster runtime. The specific setting is as follows.

> "You are a master of 3D generation, and please refer to the 'Prompt' and 'Negative Prompt' below to identify the inconsistency in the image I provided you with, with body shape, perspective, texture, and so on.
>
> Reference:
>
> 'Prompt': '3d render of xx, front view, standing, high quality, 4K',
>
> 'Negative Prompt': 'multi-head, unnatural lighting, smooth appearance, distorted color, long neck, two-nosed, extra limbs' ".

For LMM, we chose the locally deployed LLaVA [26], using the version llava-v1.6-34b.

General prompt. Our method acts as a universal enhancement for 3D generation, considering the common use of general negative prompts in baseline methods [63, 58, 40, 53, 24, 28]. The specific setting is as follows.

> "unnatural colors, poor lighting, low quality, artifacts, smooth texture".

B Time Consumption

We recorded the runtime using two baselines: GaussianDreamer [63] based on 3DGS with fewer iterations and faster speed, and DreamFusion [40] based on NeRF with more iterations and slower speed, on NVIDIA V100.

To optimize the process, we begin calculating $\mathcal{L}_{\mathrm{CG}}$ later in the training and only every 4 iterations in our experiments. This approach is consistent with the statement in Sec.3.4 that *"this module only works when the rendered images exhibit complete semantic structures."* The rationale is twofold: first, during the early stages of training, the 3D assets are relatively disorganized and lack clear semantic structures, making it challenging for LMMs to reason accurately. Therefore, we delay the introduction of $\mathcal{L}_{\mathrm{CG}}$. Second, we empirically found that calculating $\mathcal{L}_{\mathrm{CG}}$ every 4 iterations does not affect performance, allowing us to reduce training time. The results are presented in the Tab.6.

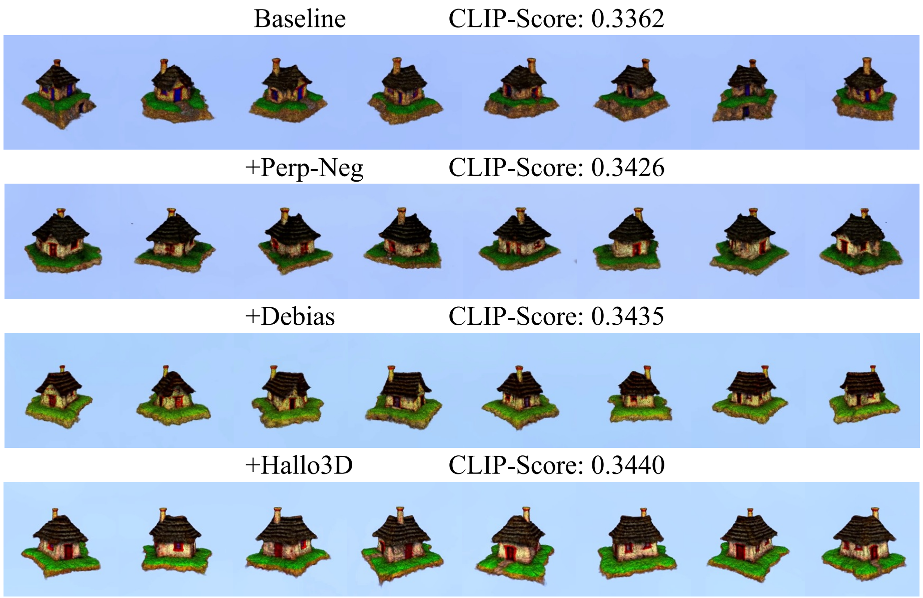
Notably, since our method includes the "Multi-View Appearance Alignment" module, which requires attention calculations across four differently angled rendered images, we set the batch size to 4 for all baselines. To ensure a fair comparison, we reduced the number of iterations to 1/4 of the original. For example, DreamFusion originally trained for 10,000 iterations, and we adjusted it to 2,500 for optimization. GaussianDreamer(iteration=1200) already uses batch=4, so we matched its iteration count at 1,200.

The experimental results indicate that while our method introduces some additional time overhead, this is fully justified by the significant improvements in performance and quality, especially considering the challenging nature of addressing the Janus Problem.

C Additional Experiments

C.1 The Effectiveness and Necessity of $\mathcal{L}_{\mathrm{CG}}$

To further underscore the necessity of incorporating $\mathcal{L}_{\mathrm{CG}}$, we plotted the curve of $\mathcal{L}_{\mathrm{CG}}$ over the course of iterations. In conjunction with this, we also plotted the CLIP-Score for both the complete model and an ablated version that omits $\mathcal{L}_{\mathrm{CG}}$. As illustrated in Fig.8, it is evident that $\mathcal{L}_{\mathrm{CG}}$ steadily decreases with increasing iterations, contributing to a marked improvement in the CLIP-Score. In



Baseline CLIP-Score: 0.3362

+Perp-Neg CLIP-Score: 0.3426

+Debias CLIP-Score: 0.3435

+Hallo3D CLIP-Score: 0.3440

prompt: A 3D model of an adorable cottage with a thatched roof.

Figure 9: Comparison experiments with Perp-Neg and Debias.

Table 6: The time consumption introduced by Hallo3D.

Baseline	**Iteration**	$\mathcal{L}_{\text{CG}}$ **Start Rounds**	**Original Time**	**Extra Time**	**Total Time**
GaussianDreamer	1200	1000	∼28 min	∼10 min	∼38 min
DreamFusion	2500	2200	∼51 min	∼15 min	∼66 min

contrast, the CLIP-Score for the model without $\mathcal{L}_{\text{CG}}$ exhibits only a marginal improvement. These findings clearly highlight both the necessity and effectiveness of incorporating $\mathcal{L}_{\text{CG}}$ into the model.

C.2 Comparison Experiments with Other 3D Consistency Enhancement Methods.

To further demonstrate the advantages of our method, we compared it with other approaches[2, 17] aimed at improving 3D consistency. As shown in Fig.9, Hallo3D more effectively addresses the Janus problem and achieves a greater improvement in CLIP-Score.

C.3 360-degree Visualization Results

Due to space constraints, Fig.5 in the main text only presents 3D generation results from selected viewpoints. The full 360-degree visualizations can be found in Fig.10 and Fig.11.

D Limitation

Our method has shown improvements in 3D generation consistency across various baselines, including both text-based and image-based approaches. However, as a method focused on enhancing view consistency, the quality of our experimental results is inherently tied to the performance of the baseline models. Moreover, the potential misuse of advanced 3D generation technologies poses risks to social trust and information integrity. Looking ahead, we will prioritize the Janus Problem as a key research direction and are committed to contributing further to the field of 3D generation alongside our fellow researchers.

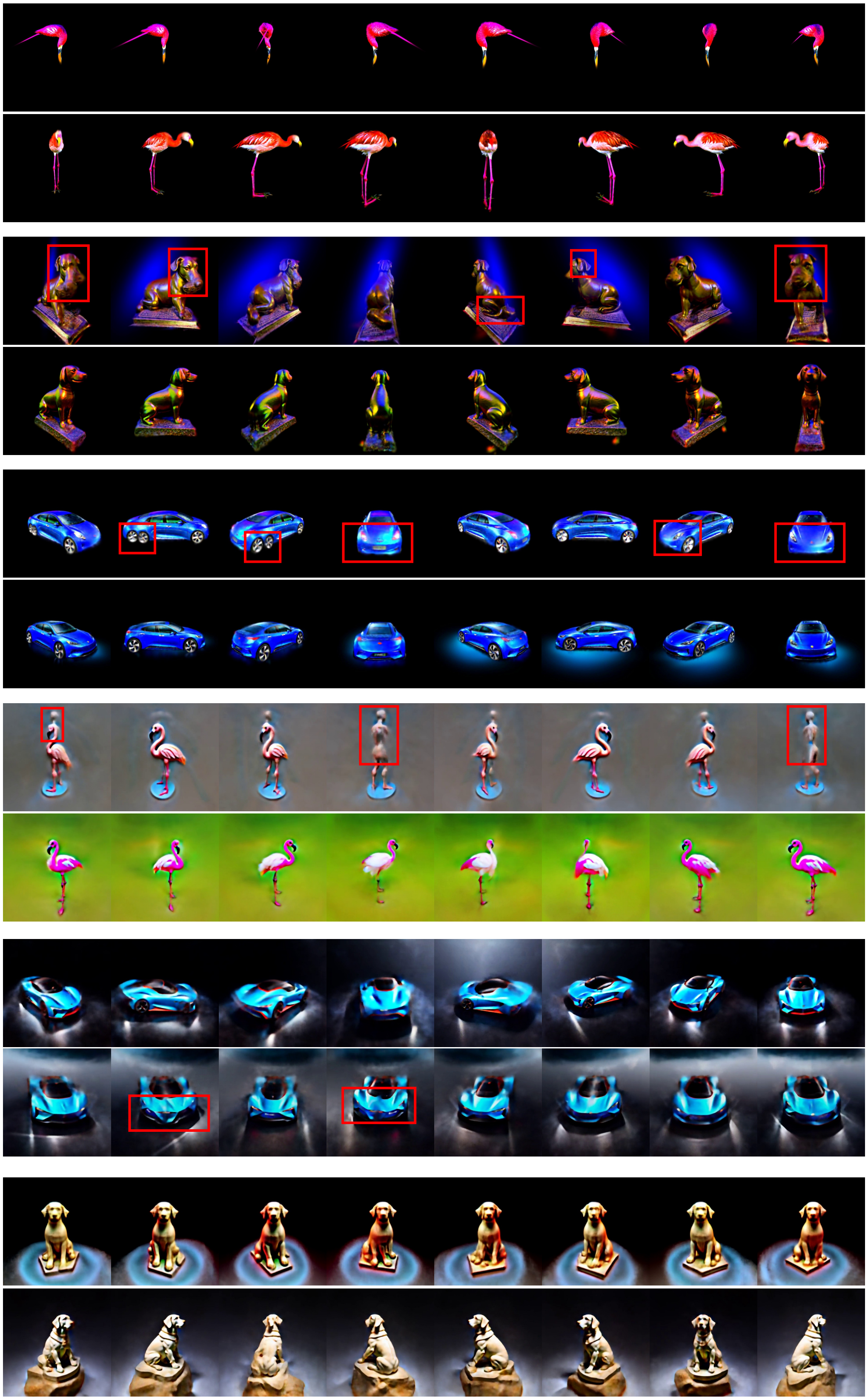


Figure 10: 360-degree visualization results in Fig.5 (1).

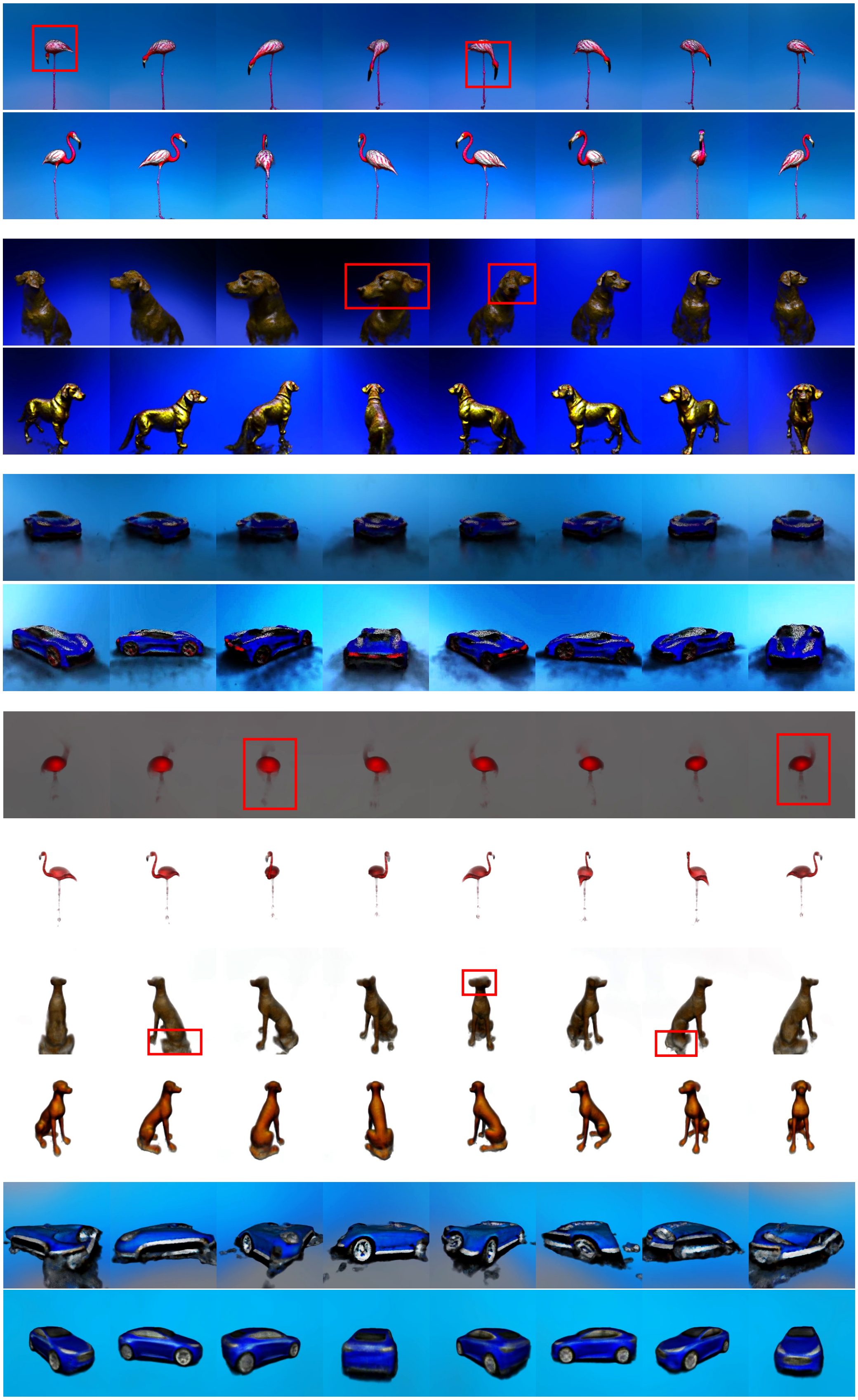


Figure 11: 360-degree visualization results in Fig.5 (2).

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe it in Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe it in Section D.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Without theory assumptions and proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe it in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are currently organizing the code and plan to release it as open-source in the future.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe it in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our main results are based on user study.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe it in Section A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We describe it in Section D.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: None.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets such as code, data, and models used in our research are properly credited to their respective creators or owners. We have explicitly mentioned the licenses and terms of use for each asset in our documentation. Additionally, we have ensured that all terms of use are strictly adhered to, including obtaining necessary permissions for assets with restrictive licenses. This practice supports ethical research standards and ensures legal compliance in the use and distribution of third-party resources.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: We are currently organizing the code and plan to release it as open-source in the future.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We have included all data collected from the survey in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: None.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.