CONSISTENT123: ONE IMAGE TO HIGHLY CONSISTENT 3D ASSET USING CASE-AWARE DIFFUSION PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing 3D objects from a single image guided by pretrained diffusion models has demonstrated promising outcomes. However, due to utilizing the case-agnostic rigid strategy, their generalization ability to arbitrary cases and the 3D consistency of reconstruction are still poor. In this work, we propose Consistent123, a case-aware two-stage method for highly consistent 3D asset reconstruction from one image with both 2D and 3D diffusion priors. In the first stage, Consistent123 utilizes only 3D structural priors for sufficient geometry exploitation, with a CLIP-based case-aware adaptive detection mechanism embedded within this process. In the second stage, 2D texture priors are introduced and progressively take on a dominant guiding role, delicately sculpting the details of the 3D model. Consistent123 aligns more closely with the evolving trends in guidance requirements, adaptively providing adequate 3D geometric initialization and suitable 2D texture refinement for different objects. Consistent123 can obtain highly 3D-consistent reconstruction and exhibits strong generalization ability across various objects. Qualitative and quantitative experiments show that our method significantly outperforms state-of-the-art image-to-3D methods. See `https://Consistent123.github.io` for a more comprehensive exploration of our generated 3D assets.

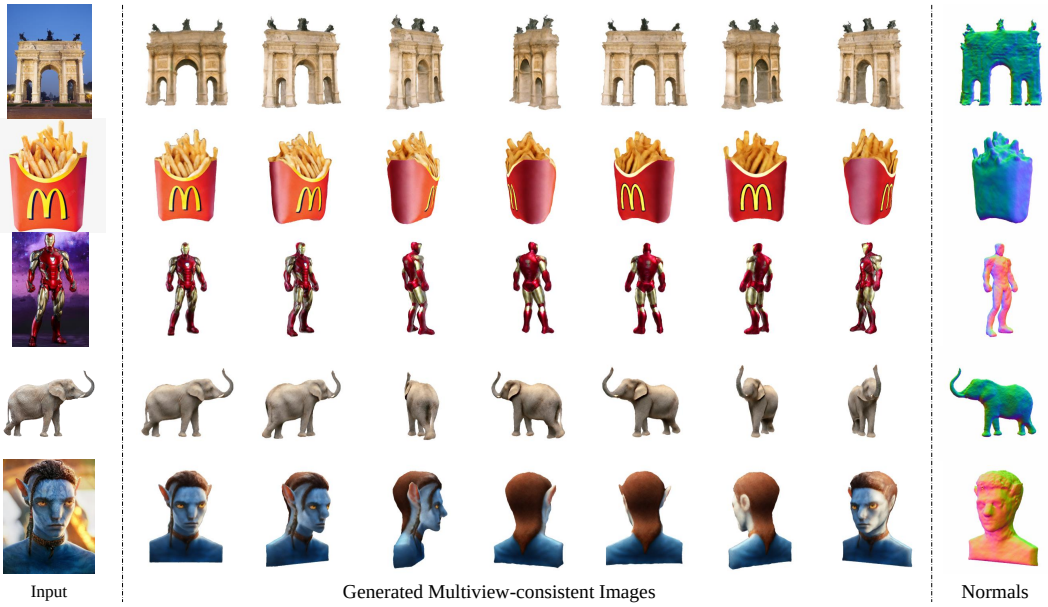


Figure 1: **The reconstructed highly consistent 3D assets from a single image of Consistent123.** Rendered 3D models are presented by seven views (middle part) and normals (right part).

1 INTRODUCTION

The experienced 3D artists can craft intricate 3D models from images, however, this demands hundreds of hours of manual effort. In this study, we aim to efficiently generate highly consistent 3D model from a single image. This endeavor promises to furnish a potent adjunct for 3D creation and offers a swift means of procuring 3D objects for virtual three-dimensional environments construction.

Despite decades of extensive research efforts (Mescheder et al., 2019; Park et al., 2019; Wang et al., 2018; Hanocka et al., 2020; Mildenhall et al., 2020), the task of reconstructing 3D structure and texture from a single viewpoint remains inherently challenging due to its ill-posed nature. To address this challenge, one category of approaches relies on costly 3D annotations obtained through CAD software or tailored domain-specific prior knowledge (Wang et al., 2023; Zhang et al., 2023), e.g. human and clothing templates, which contribute to consistent results while also limiting applicability to arbitrary objects. Another cue harnesses the generalization ability of 2D generation models like CLIP (Radford et al., 2021a) and Stable Diffusion (Rombach et al., 2022). However, Melas-Kyriazi et al. (2023) and Tang et al. (2023) suffer from severe multi-face issue, that is, the face appears at many views of the 3D model. With 3D structure prior, Liu et al. (2023) and Qian et al. (2023) can stably recover the 3D structure of an object, but struggle to obtain highly consistent reconstruction. All these methods do not take into account the unique characteristics of object, and utilize fixed strategy for different cases. These case-agnostic approaches face difficulty in adapting optimization strategies to arbitrary objects.

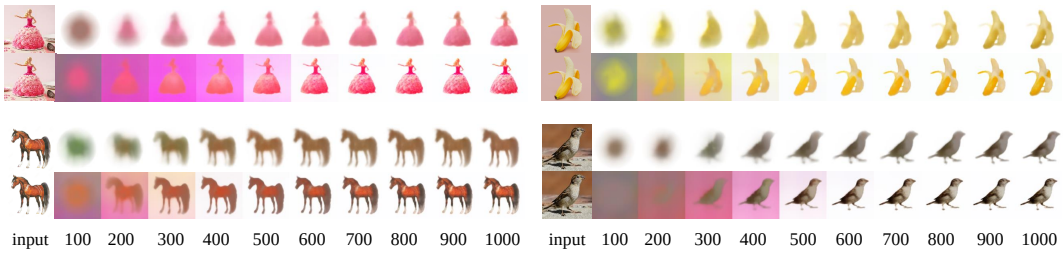


Figure 2: **The observation of optimization.** For each case, the top row shows the optimization process using 2D priors, and the bottom row using 3D priors.

However, our objective is to establish a versatile approach applicable to a broad spectrum of objects, endowed with the capability to dynamically adapt guidance strategy according to the extent of reconstruction progress. To achieve this aim, we draw attention to two pivotal **observations**: **(1)** Across various objects, a case-aware optimization phase, driven solely by 3D structural prior in the early stage, ensures the fidelity and consistency of the eventual reconstruction. **(2)** During the reconstruction process, the initial focus lies on capturing the object's overall structure, followed by the meticulous refinement of geometric shape and texture details, as illustrated in Fig 2.

Considering these, we propose ***Consistent123***, a novel approach for one image to highly consistent 3D asset using case-aware 2D and 3D diffusion priors. Specifically, Consistent123 takes two stages. *Stage 1*: Consistent123 initializes the 3D content solely with 3D prior, thereby mitigating any disruption from 2D prior in structure exploitation. This process involves a case-aware boundary judgement, where we periodically sample the 3D content from fixed perspectives and measure their similarity with textual information. Once the changing rate of the similarity falls below a threshold, Consistent123 switches to stage 2. *Stage 2*: Consistent123 optimizes the 3D content with dynamic prior, namely the combination of 2D and 3D prior. Our rationale is to reduce the emphasis on 3D prior over time, while accentuating the significance of 2D prior, which serve as the principal guidance for exploring texture intricacies. Consistent123 adaptively tailors an continuous optimization procedure for different input, facilitating the creation of exceptionally coherent 3D assets.

We evaluate Consistent123 on the RealFusion15 (Melas-Kyriazi et al., 2023) dataset and our collected C10 dataset. Through quantitative and qualitative analysis, we demonstrate the superiority of Consistent123 when compared to state-of-the-art methods. In summary, our contributions can be summarized as follows:

- We propose a case-aware image-to-3D method, **Consistent123**, which aligns more effectively with the demands of prior knowledge. It places a heightened emphasis on 3D structural guidance in the initial stage and progressively integrates 2D texture details in the subsequent stage.
- Consistent123 incorporates a adaptive detection mechanism, eliminating the necessity for manual adjustments to the 3D-to-2D prior ratio. This mechanism autonomously identifies the conclusion of 3D optimization and seamlessly transitions to a 3D-to-2D reduction strategy, improving its applicability across objects with diverse geometric and textural characteristics.
- Consistent123 demonstrates excellent 3D consistency in contrast to purely 3D, purely 2D, and 3D-2D fusion methodologies. Furthermore, our approach yields superior geometric and textural quality, concurrently addressing the challenge of multi-face problem.

2 Related Work

2.1 Text-to-3D Generation

Generating 3D models is a challenging task, often hindered by the scarcity of 3D data. As an alternative, researchers have turned to 2D visual models, which are more readily available. One such approach is to use the CLIP model (Radford et al., 2021a), which has a unique cross-modal matching mechanism that can align input text with rendered perspective images. Mohammad Khalid et al. (2022) directly employed CLIP to optimize the geometry and textures of meshes. Jain et al. (2022) and Wang et al. (2022a) utilized the neural implicit representation, NeRF (Mildenhall et al., 2020), as the optimization target for CLIP.

Due to the promising performance of the Diffusion model in 2D image generation (Rombach et al., 2022; Ramesh et al., 2022; Wang et al., 2022b), some studies have extended its application to 3D generation. Poole et al. (2023) directly used a 2D diffusion model to optimize the alignment between various rendered perspectives and text with SDS loss, thereby generating 3D objects that match the input text. Lin et al. (2023) used the two-stage optimization with diffusion model to get a higher resolution result. Seo et al. (2023) generated a 2D image as a reference and introduced a 3D prior based on the generated image. It also incorporated optimization with a prompt embedding to maintain consistency across different perspectives. Richardson et al. (2023) generated textures using a depth-to-image diffusion model and blended textures from various perspectives using a Trimap. Wang et al. (2023) and Xu et al. (2023) bridged the gap between vision and language with CLIP, and achieved a unified 3D diffusion model for text-conditioned and image-conditioned 3D generation. Cao et al. (2023) transformed the observation space to a standard space with a human prior and used a diffusion model to optimize NeRF for each rendered perspective.

2.2 Single Image 3D Reconstruction

Single-image 3D reconstruction has been a challenging problem in the fields of graphics and computer vision, due to the scarcity of sufficient information. To address this issue, researchers have explored various approaches, including the use of 3D-aware GANs and Diffusion models. Some work (Chan et al., 2022; Yin et al., 2022; Xiang et al., 2022; Xie et al., 2023) leveraged 3D-aware GANs to perform 3D face generation with GAN inversion techniques (Roich et al., 2022; Wang et al., 2022c). Other works used Diffusion models to generate new perspectives in reconstruction. Wang et al. (2023) proposed a 3D diffusion model for high-quality 3D content creation, which is trained on synthetic 3D data. Liu et al. (2023) fine-tuned Stable Diffusion with injected camera parameters on a large 3D dataset Deitke et al. (2023) to learn novel view synthesis.

Another line of work adopted 2D diffusion prior to directly optimize a 3D object without the need for large-scale 3D training data. These approaches represent promising avenues for addressing the challenge of single-image 3D reconstruction. As a seminal work, Tang et al. (2023) used an image caption model (Li et al., 2022) to generate text descriptions of the input image. The researchers then optimized the generation of novel views with SDS loss, as well as introducing a denoised CLIP loss to maintain consistency among different views. Meanwhile, Melas-Kyriazi et al. (2023) utilized textual inversion to optimize prompt embedding from input images and then employed SDS loss to optimize the generation of new perspectives. Qian et al. (2023) leveraged a rough 3D prior generated by zero 1-to-3 (Liu et al., 2023) and combined it with textual inversion to optimize prompt embedding using SDS loss with fixed weighting.

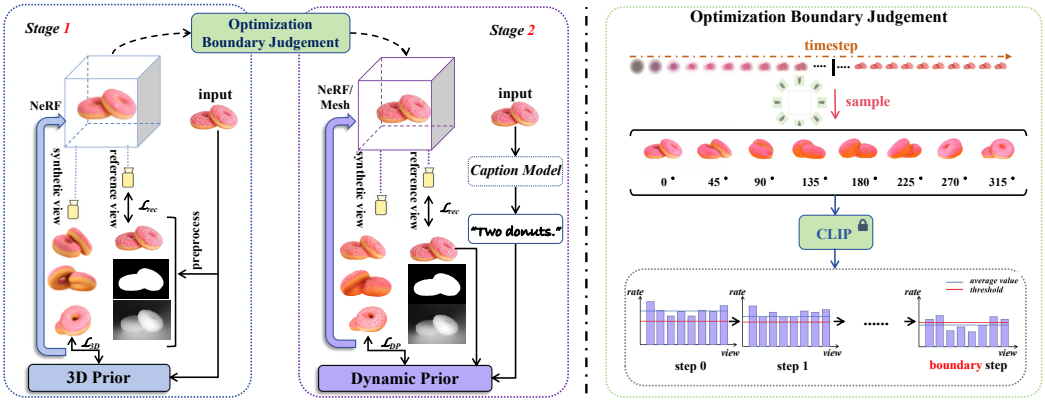


Figure 3: **The framework of Consistent123.** Consistent123 consists of two stages. In the first stage, we take advantage of 3D prior to optimize the geometry of 3D object. With the help of an optimization boundary judgment mechanism based on CLIP, we ensure the geometry initial optimization process is well conducted. Then, in the second stage, the output from the last stage continues to be optimized by the fusion of 2D prior and 3D prior in a specific ratio based on timestep, which is also named Dynamic Prior. To access a high-consistence and high-quality asset, we employ enhanced representation like Mesh instead of NeRF in the final period of optimization. The eventual result of the framework has correct geometry and exquisite texture from visual observation.

3 METHODOLOGY

As shown in Fig 3, the optimization process of Consistent123 can be categorized from a perspective standpoint into two phases: the reference view and the novel view. In the reference viewpoint, we primarily employ the input image as the basis for reconstruction, a topic comprehensively addressed in Section 3.1. The optimization of the novel view unfolds across two distinct stages. These two stages are thoroughly explored in Sections 3.2 and Section 3.3, respectively. The resultant model output consistently exhibits a high degree of 3D consistency and exceptional texture quality.

3.1 REFERENCE VIEW RECONSTRUCTION

Imported a 2D RGB image, Consistent123 adopts a preprocess operation to get derivative ground truth which can be used in the loss calculation in the reference view. We utilize pretrained model (Eftekhar et al., 2021; Kar et al., 2022) to acquire the demerger $\mathbf{I}^{gt}$, the binary mask $\mathbf{M}^{gt}$ and the depth of object $\mathbf{D}^{gt}$. $\mathcal{L}_{rgb}$ ensures the similarity between the input image and the rendered reference view image. Mean Squared Error (MSE) loss is leveraged to calculate the $\mathcal{L}_{rgb}$ in the form as follows:

$$\mathcal{L}_{rgb} = \|\mathbf{I}^{gt} - \mathcal{G}_{\theta}\left(v^{r}\right)\|_{2}^{2} \tag{1}$$

where $\mathcal{G}_{\theta}$ stands for the representation model in the optimization process, v^{r} represents the viewpoint of reference view in the rendering process. The design of $\mathcal{L}_{mask}$ likewise employs MSE to operate calculation whose concrete expression as follows:

$$\mathcal{L}_{mask} = \|\mathbf{M}^{gt} - \mathbf{M}\left(\mathcal{G}_{\theta}\left(v^{r}\right)\right)\|_{2}^{2} \tag{2}$$

where $\mathbf{M}\left(\cdot\right)$ means the operation of extracting the mask of the rendered image. Seeing that the method of using depth prior in the former of this area, we decide to adopt the normalized negative Pearson correlation $\mathcal{L}_{depth}$ in-depth loss computation. Given three vital parts of reference view reconstruction loss, we merge them into a modified form of expression:

$$\mathcal{L}_{rec} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{depth}\mathcal{L}_{depth} \tag{3}$$

where λ_{rgb}, λ_{mask} and λ_{depth} are controllable parameters which are used to regulate the ratio of each supervision. With the help of merged loss $\mathcal{L}_{rec}$, we can restore a high detail and correct geometry target on the reference viewpoint.

3.2 Optimization Boundary Judgement

The optimization process illustrated in Fig 2 demonstrates the efficiency of 3D structural priors in capturing the shape of object, and the 3D priors play a crucial role mainly in the initial stage of reconstruction. To ensure the comprehensive recovery of the object's shape as depicted in the image, we establish a structural initialization stage, namely stage 1, where only 3D structural priors guide the optimization. The guidance of the 3D prior can be expressed as the gradient which is used to update the parameter θ:

$$\nabla_\theta \mathcal{L}_{3D}(\phi, \mathcal{G}_\theta) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi\left(\mathbf{z}_t; \mathbf{I}^r, t, R, T\right) - \epsilon\right)\frac{\partial \mathbf{I}}{\partial \theta}\right] \tag{4}$$

where t denotes the training timestep, $\mathbf{z}$ represents the latent variable generated through the encoding of the image $\mathbf{I}$, R and T mean the rotation and translation parameters of the camera. The function $w(t)$ corresponds to a weighting function, while ϵ_ϕ and ϵ respectively denote the noise prediction value generated by the U-Net component of the 2D diffusion model and the ground truth noise. During stage 1, 2D priors are deliberately excluded, effectively mitigating the multi-face issue. The output of this stage is 3D content with high-quality structure, yet it significantly lags in terms of texture fidelity compared to the image representation. That's mainly because of the deficiency of texture information, which is primarily driven by 2D priors.

Consequently, we embed a case-aware CLIP-based detection mechanism within stage 1 to determine whether the shape of the current 3D content has been accurately reconstructed. If so, a transition is made to stage 2, with 2D priors introduced gradually. During the first-stage training, we conduct boundary judgement at specific iterations. Specifically, we periodically perform detection at intervals of h iterations, set to 20 in our experiments. For each detection step k, we render the current 3D content from different viewpoints, resulting in M images, and then calculate the average similarity score between these images and textual descriptions using the CLIP model:

$$\mathcal{S}^k_{CLIP}\left(y, \mathcal{G}^k_\theta\right) = \frac{1}{M}\sum_{v\in V}\varepsilon_{CLIP}\left(\mathcal{G}^k_\theta\left(v\right)\right)\cdot\varphi_{CLIP}\left(y\right) \tag{5}$$

where y is the description of the reference image, and v is a rendering perspective belonging to sample views set V. ε_{CLIP} is a CLIP image encoder and φ_{CLIP} is a CLIP text encoder. To determine whether the shape of the current 3D content has been adequately recovered, we compute the moving average of changing rate of $\mathcal{S}_{CLIP}$:

$$R^k = \frac{1}{L}\sum_{i=k-L+1}^{k}\left(\mathcal{S}^i_{CLIP} - \mathcal{S}^{i-1}_{CLIP}\right)/\mathcal{S}^{i-1}_{CLIP} \tag{6}$$

where L is the size of the sliding window. When this rate falls below a threshold δ, the current 3D content is considered to possess a structure similar to that represented in the image.

3.3 Dynamic Prior

Recognizing that 3D prior optimization is characterized by consistent structure guidance but weak texture exploration, while 2D prior optimization leads to high texture fidelity but may occasionally diverge from the input image, we posit these two priors exhibit complementarity, each benefiting the quality of the final 3D model. Consequently, in Stage 2, we introduce a 2D diffusion model as the guiding 2D prior to enrich the texture details of the 3D object. Throughout the optimization process, the 2D diffusion model primarily employs Score Distillation Sampling (SDS) (Poole et al., 2023) loss to bridge the gap between predicted noise and ground truth noise. This concept is elucidated as the follows:

$$\nabla_\theta \mathcal{L}_{2D}(\phi, \mathcal{G}_\theta) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi\left(\mathbf{z}_t; y, t\right) - \epsilon\right)\frac{\partial \mathbf{z}}{\partial \mathbf{I}}\frac{\partial \mathbf{I}}{\partial \theta}\right] \tag{7}$$

where y, originating from either user observations or the output of a caption model, represents the text prompt describing the 3D object. However, we have observed that, in the stage 2, when the optimization relies solely on the 2D prior, the resulting 3D asset often exhibits an unfaithful appearance. This is attributed to the low-resolution output of stage 1 possessing poor low-level information such as color, shading, and texture, which makes room for 2D prior to provide high-resolution but unfaithful guidance. Moreover, the alignment relationship between the input text

prompt and each individual novel view which is waiting to be optimized by the 2D prior varies. This variability leads the 2D prior to introduce certain unfaithful details, which we refer to as the 'Over Imagination' issue. Consequently, the eventual output typically maintains a reasonable structure but displays an unfaithful novel view, resulting in an inconsistent appearance.

To resolve the above problem, we incorporate 3D prior and 2D prior in an incremental trade-off method instead of only using 2D diffusion model in stage 2, which we call it **Dynamic Prior**. More specifically, we design a timestep-based dynamic integration strategy of two kinds of prior to gradually introduce exquisite guidance information while maintaining its faithfulness to input image. The loss formula of dynamic prior using both $\mathcal{L}_{3D}$ and $\mathcal{L}_{2D}$ is as follows:

$$\mathcal{L}_{DP} = e^{-\frac{t}{T}}\mathcal{L}_{3D} + \left(1 - e^{-\frac{t}{T}}\right)\mathcal{L}_{2D} \tag{8}$$

where T represents total timesteps of optimization. As shown in Equation (8), we determine the weighting coefficients of two losses using an exponential form which is dependent on the parameter t. As t increases, $\mathcal{L}_{3D}$ which is primarily contributing structural information undergoes a gradual reduction in weight, while $\mathcal{L}_{2D}$ which is mainly responsible for optimizing texture information exhibits a progressive increase of influence. We have also considered expressing $\mathcal{L}_{DP}$ in the form of other basis functions, but extensive experimental results have shown that the expression in Equation (8) yields many excellent and impressive results, and more details of the comparison can be found in Section 4.4. Compared to single prior or fixed ratio prior, the outputs of Consistent123 are more consistent and exquisite from the perspective of texture and geometry.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

For the diffusion prior, we adopt the open-source Stable Diffusion (Rombach et al., 2022) of version 2.1 as 2D prior, and employ the Zero-1-to-3 (Liu et al., 2023) as the 3D prior. We use Instant-NGP (Müller et al., 2022) to implement the NeRF representation and for mesh rendering , we utilize DMTet (Shen et al., 2021), a hybrid SDF-Mesh representation. The rendering resolutions are configured as 128 × 128 for NeRF and 1024 × 1024 for mesh. Following the camera sampling approach adopted in Dreamfusion (Poole et al., 2023), we sample the reference view with a 25% probability and the novel views with a 75% probability. For the case-aware detection mechanism, we sample from 8 viewpoints each time, that is $V = \{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$. The sliding window size L is set to 5 and the threshold δ of 0.00025. We use Adam optimizer with a learning rate of 0.001 throughout the reconstruction. For an image, the entire training process with 10,000 iterations takes approximately 30 minutes on a single NVIDIA A100 GPU.

4.2 COMPARISON WITH STATE-OF-THE-ART

Datasets. We consider a classic benchmark, RealFusion15, released by RealFusion (Melas-Kyriazi et al., 2023). RealFusion15 consists of 15 images featuring a variety of subjects. In addition, we introduced a C10 dataset consisting of 100 images collected from 10 categories which covers a wider range of items. These 10 categories broadly encompass common objects found in daily life, including fruits, balls, furniture, scenes, flora and fauna, food, transportation, clothing and footwear, cartoon characters, and artwork. Thus, the results on C10 can serve as an effective evaluation of the method's generalization ability.

Baselines and metrics. We evaluate Consistent123 against state-of-the-art baselines, including RealFusion (Melas-Kyriazi et al., 2023), Make-it-3D (Tang et al., 2023), Zero-1-to-3 (Liu et al., 2023), and Magic123 (Qian et al., 2023), on both the RealFusion15 and C10 datasets. Like Magic123, we use an improved implementation (Tang, 2022) of Zero-1-to-3, and the original released code for other works. For quantitative evaluation, we adopt three metrics, namely CLIP-similarity (Radford et al., 2021b), PSNR and LPIPS (Zhang et al., 2018). CLIP-similarity quantifies the average CLIP distance between the rendered image and the reference image, serving as a measure of 3D consistency by assessing appearance similarity across novel views and the reference view. PSNR and LPIPS assess the reconstruction quality and perceptual similarity at the reference view.

Quantitative comparison. As demonstrated in Table 1, on the RealFusion15 dataset, Consistent123 attains the most favorable results in the CLIP-Similarity metric which gain an increment of **11.2%**

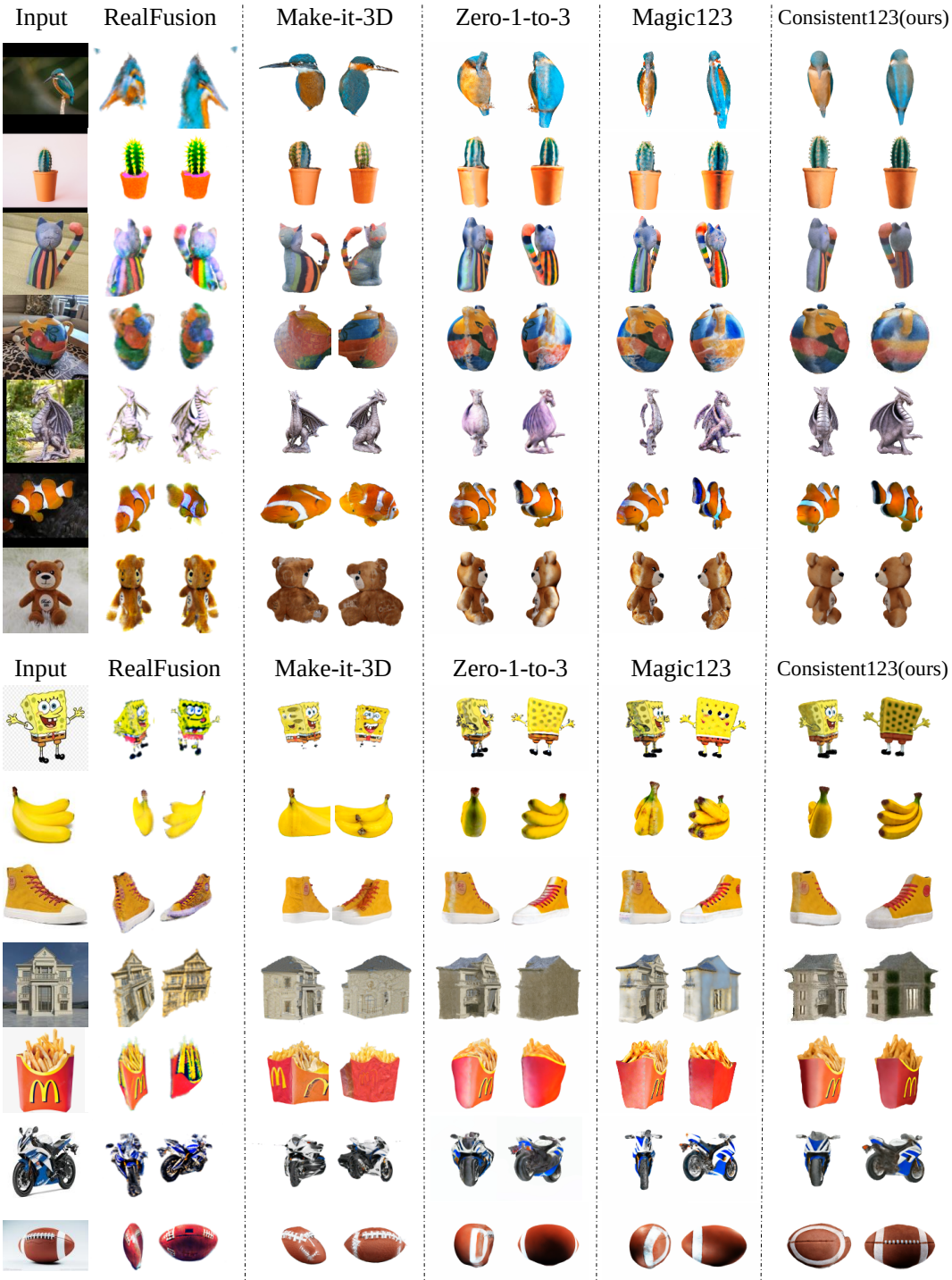


Figure 4: **Qualitative comparison vs SOTA methods.** The results on the RealFusion15 dataset is shown on top, and results on the C10 dataset on the bottom. We randomly sample 2 novel views to showcase, and reference view and other views are included in Appendix A.1. Please visit `https://Consistent123.github.io` for a more intuitive comparison by watching videos.

compared to the original SOTA, signifying that our method yields the most consistent 3D models. Regarding reference view reconstruction, Consistent123 performs comparably to Magic123 and Zero-1-to-3, and significantly outperforms RealFusion and Make-it-3D. On the C10 dataset, encompassing images from 10 distinct categories, Consistent123 outpaces its counterparts by a substantial margin across all evaluation metrics. Moreover, there is a notable enhancement in CLIP-Similarity, accompanied by an improvement of **2.972** in PSNR and **0.066** in LPIPS metrics when compared to the previously top-performing model, which underscore robust generalization capability of Consistent123 across diverse object categories.

Table 1: Qualitative results on the RealFusion15 and C10 datasets. Make-it-3D uses CLIP similarity to supervise the training, so its value† is not considered for Make-it-3D in the comparison.

Dataset	Metrics/Methods	RealFusion	Make-it-3D	Zero-1-to-3	Magic123	Consistent123(ours)
RealFusion15	CLIP-Similarity↑	0.735	0.839†	0.759	0.747	**0.844**
	PSNR↑	20.216	20.010	25.386	25.637	**25.682**
	LPIPS↓	0.197	0.119	0.068	0.062	**0.056**
C10	CLIP-Similarity↑	0.680	0.824†	0.700	0.751	**0.770**
	PSNR↑	22.355	19.412	18.292	15.538	**25.327**
	LPIPS↓	0.140	0.120	0.229	0.197	**0.054**

Qualitative comparison. We present a comprehensive set of qualitative results featuring 14 images drawn from the RealFusion15 and C10 datasets in Fig 4. In contrast to our method, RealFusion often yields flat 3D results with colors and shapes that exhibit little resemblance to the input image. Make-it-3D displays competitive texture quality but grapples with a prominent issue of multi-face. For instance, when reconstructing objects like teddy bears and Spongeboy, it introduces facial features at different novel views, which should only appear in the reference view. Zero-1-to-3 and Magic123 produce visually plausible structures, but the consistency of texture among all views, especially in side views, is poor. For example, in the cases of fish and rugby, their textures fail to achieve a smooth transition when observed from the side view. In contrast, our methodology excels in generating 3D models that not only exhibit semantic consistency with the input image but also maintain a high degree of consistency in terms of both texture and geometry across all views.

4.3 Ablation Study of Two Stage Optimization

In this section, we emphasize the significance of boundary judgment. We divide the reconstruction process into three parts, namely: 3D structural initialization, boundary judgment, and dynamic prior-based optimization. In cases where boundary judgment is absent, the optimization process can be categorized into two approaches: full 3D structural initialization (boundary at the training starting point) or full dynamic prior-based optimization (boundary at the training endpoint), denoted as Consistent123$_{3D}$ and Consistent123$_{dynamic}$, respectively. As illustrated in Fig 5, without the guidance of 2D texture priors, Consistent123$_{3D}$ produces visually unrealistic colors in the new view of the car, and in the absence of 3D structural initialization, Consistent123$_{dynamic}$ exhibits inconsistency and multi-face issue in Mona Lisa's face. In contrast, results with boundary judgment showcase superiority in both texture and structure.

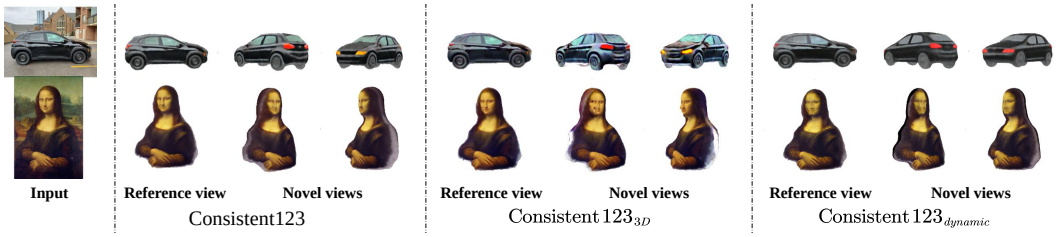


Figure 5: **The ablation of two stage optimization.**

4.4 Ablation Study of Dynamic Prior

Dynamic prior refers to the method of dynamically adjusting the ratio of 2D and 3D priors based on different time steps during the optimization process. Depending on the transformation method, we compare the optimization effects of three different approaches: exponential (Equation (8)), linear

and logarithmic (Equation (9)). We assessed them across ten categories, each comprising 5 images from the RealFusion15 and C10 datasets. As shown in the Table 2, the exponential variation process, which is the our adopted method, can achieve a higher CLIP-Similarity on most of the categories, which to some extent reflects the reconstruction consistency. The actual reconstruction results also support this, as the exponential variation method can effectively mitigate the multi-head problem, leading to higher reconstruction quality and better consistency.

$$\mathcal{L}_{linear} = \frac{t}{T}\mathcal{L}_{3D} + \left(1 - \frac{t}{T}\right)\mathcal{L}_{2D} \quad , \quad \mathcal{L}_{log} = \log_2 \frac{t}{T}\,\mathcal{L}_{3D} + \left(1 - \log_2 \frac{t}{T}\right)\mathcal{L}_{2D} \tag{9}$$

The key difference between exponential transformation and the other two lies in the fact that exponential transformation can inject 2D priors more quickly. In the previous optimization stage, 3D priors were used to ensure the correctness of the basic geometric structure of the reconstruction. The purpose of dynamic priors is to optimize the quality and consistency of the reconstruction while maintaining the correctness of the 3D structure. The former has already undergone optimization in the first stage, requiring only a small amount of injection during the dynamic prior stage to maintain the effectiveness of the 3D prior.

Table 2: Ablation Study of Dynamic Prior on the RealFusion15 and C10 datasets.

Methods	Metrics/class	ball	biont	furniture	cartoon	fruit	statue	food	vehicle	costume	scene	average
log	CLIP-Similarity↑	0.79	0.85	**0.58**	0.77	0.87	0.71	0.87	0.74	**0.67**	0.68	0.76
	PSNR↑	26.45	25.46	23.19	23.97	24.62	22.94	27.33	24.24	**26.14**	21.71	24.59
	LPIPS↓	**0.04**	0.06	**0.12**	**0.06**	0.06	0.11	**0.03**	0.07	0.06	0.10	0.07
linear	CLIP-Similarity↑	0.82	0.85	0.55	0.74	**0.88**	0.73	**0.88**	0.72	0.65	0.70	0.76
	PSNR↑	26.32	25.51	22.96	23.43	25.31	**25.71**	**27.41**	24.57	25.36	21.63	24.96
	LPIPS↓	**0.04**	0.05	0.13	0.09	**0.04**	**0.06**	**0.03**	0.07	0.06	0.10	0.07
exp	CLIP-Similarity↑	**0.87**	**0.88**	0.54	**0.78**	0.87	**0.77**	**0.88**	**0.76**	**0.67**	**0.72**	**0.79**
	PSNR↑	**27.50**	**26.09**	**23.28**	**24.29**	**25.39**	25.63	27.02	**25.16**	25.65	**21.78**	**25.30**
	LPIPS↓	**0.04**	**0.04**	**0.12**	**0.06**	0.05	0.07	0.04	**0.05**	**0.05**	**0.09**	**0.06**

4.5 User Study

Due to the absence of ground-truth 3D models, we conducted a perceptual study to compare Consistent123 against SOTA baselines. Participants were tasked with selecting the best result that represents the texture and structure of the object depicted in the image. To quantify the likelihood of participants favoring SOTA methods over Consistent123, we present the corresponding results in Fig 6. Our method demonstrates superior performance compared to the alternatives, exhibiting a **65.7%** advantage in the user study. More details are available in the Appendix A.3.

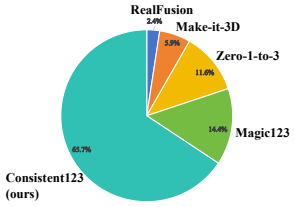


Figure 6: **User Study.** The collected results of preference.

5 Conclusion and Discussion

Conclusion. In this study, we introduce Consistent123, a two-stage framework designed for achieving highly detailed and consistent 3D reconstructions from single images. By recognizing the complementary nature of 3D and 2D priors during the optimization process, we have devised a training trade-off strategy that prioritizes initial geometry optimization with 3D priors, followed by the gradual incorporation of exquisite guidance from 2D priors over the course of optimization. Between the two optimization stages, we employ a large-scale pretrained image-text pair model as a discriminator for multi-view samples to ensure that the 3D object gains sufficient geometry guidance before undergoing dynamic prior optimization in stage 2. The formulation of our dynamic prior is determined through the exploration of various foundational function forms, with a subsequent comparison of their categorized experimental results. Our approach demonstrates enhanced 3D consistency, encompassing both structural and textural aspects, as demonstrated on existing benchmark datasets and those we have curated.

Limitation. Our study reveals two key limitations. Firstly, during stage 1, heavy reliance on 3D priors influences the 3D object, with reconstruction quality notably affected by the input image's viewpoint. Secondly, output quality depends on the description of asset in stage 2. Finer-grained descriptions enhance output consistency, while overly brief or ambiguous descriptions lead to the 'Over Imagination' issue in Stable Diffusion (Rombach et al., 2022), introducing inaccurate details.

REFERENCES

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 10786–10796, 2021.

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

Rana Hanocka, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Point2mesh: A self-prior for deformable meshes. *arXiv preprint arXiv:2005.11084*, 2020.

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 857–866, 2022.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18963–18974, 2022.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2022.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, 2023.

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. *arXiv preprint arXiv:2303.11328*, 2023.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4460–4470, 2019.

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. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Proceedings of the European conference on computer vision (ECCV)*, pp. 405–421, Cham, 2020. Springer International Publishing.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, 2022.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):102:1–102:15, July 2022.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 165–174, 2019.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

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. *arXiv preprint arXiv:2306.17843*, 2023.

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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763. PMLR, 18–24 Jul 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021b.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

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. *arXiv preprint arXiv:2303.07937*, 2023.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion.

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. *arXiv preprint arXiv:2303.14184*, 2023.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3825–3834, 2022a.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–67, 2018.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022b.

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022c.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4563–4573, 2023.

Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022.

Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 321–331, June 2023.

Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. *arXiv preprint arXiv:2307.11545*, 2023.

Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Öztireli Cengiz, and Yujiu Yang. 3d gan inversion with facial symmetry prior. *arxiv:2211.16927*, 2022.

Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 501–511, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

A APPENDIX

A.1 ADDITIONAL RESULTS

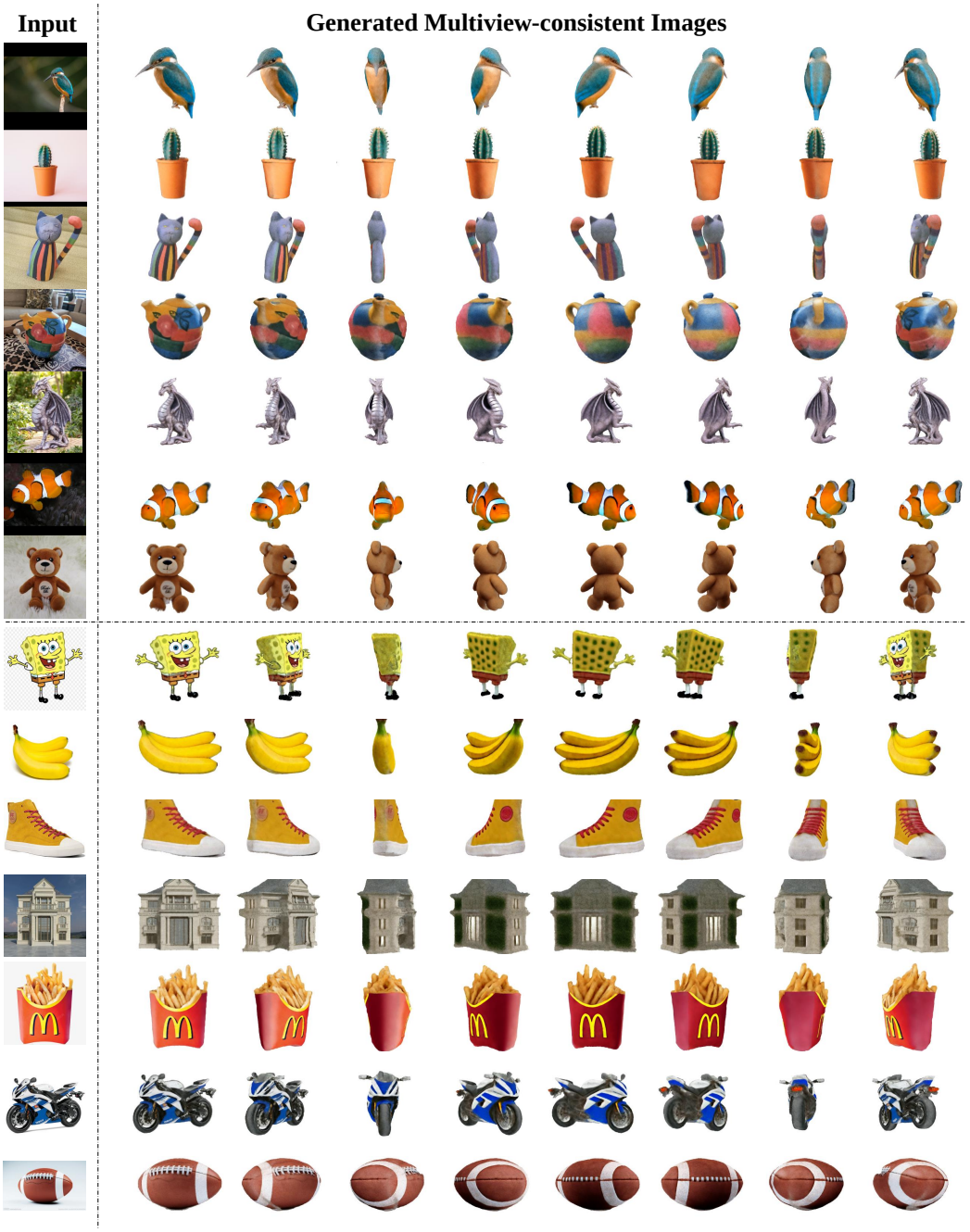


Figure 7: **The reconstruction of Consistent123.** We showcase the 3D model mentioned in Section 4.2 from 8 perspectives.

In order to better demonstrate 3D consistency, we visualize the 3D model reconstructed by our method from 8 perspectives in Fig 7. Besides, we present the 3D assets obtained by our method in Fig 8. The corresponding videos can be found at `https://Consistent123.github.io`.

Figure 8: **The 3D assets obtained by Consistent123.**

A.2 Ablation Study of Concept Injection

Several recent works (Melas-Kyriazi et al., 2023; Qian et al., 2023; Tang et al., 2023) within the same research domain have also highlighted the potential issue of 2D prior reliance, wherein the reconstructed object may not faithfully align with the original image. A common approach to addressing this challenge involves the adoption of personalized concept injection methods, such as textual inversion (Gal et al., 2022). Specifically, textual inversion employs a specific token in place of a purely textual prompt to represent the object depicted in the reference image. As depicted in Fig 9, we have also incorporated this personalized customization technique as part of our control experimental group. The experimental outcomes, however, reveal that textual inversion does not

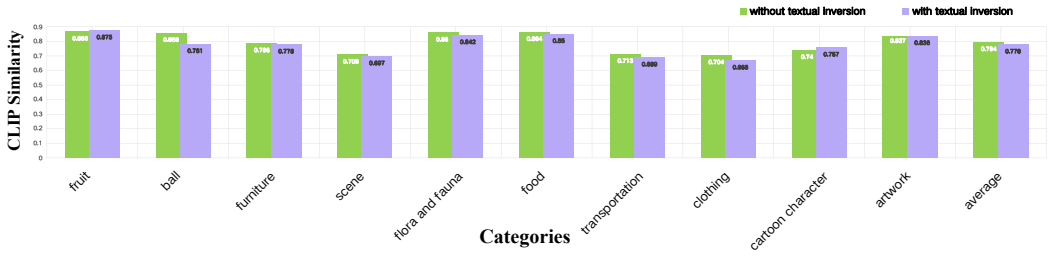


Figure 9: **Ablation study of concept injection (quantitative).** We quantitatively compare the experimental results of 2D prior using textual inversion (Gal et al., 2022) and the one without on 10 categories of input images in Realfusion15 and C10 datasets.

yield significant performance improvements across most data categories. Consequently, considering the relatively time-consuming nature of prompt tuning operations (approximately 2 hours per case on a single NVIDIA A100 GPU), we have opted to exclude it as the default choice in our pipeline, instead offering it as an optional selection.

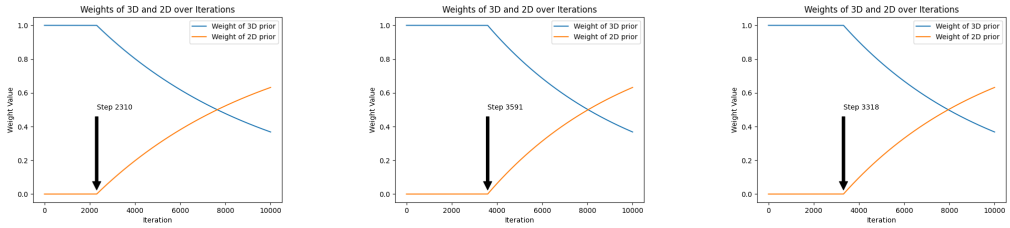
A.3 User Study

To delve deeper into the qualitative assessment of model outputs as perceived by the sense of competence, we conducted a user study comprising 784 feedbacks from 56 users to gather statistical data, as depicted in the Fig 6. From the RealFusion15 and C10 datasets, we carefully selected 14 representative cases to gauge user preferences. For each of these 14 cases, we compared the outputs of our proposed method against those of four other existing methods. Participants were asked to rank the five available outputs for each case based on their preferences. The final determination of user preference was derived from a statistical analysis of the collected data. As illustrated in the pie chart, our method demonstrates a clear and substantial lead over previous methods within the same research domain.

A.4 Additional Comparison Results

Figure 10: **Qualitative comparison vs SOTA methods on NeRF4 dataset.**

A.5 Visualization of the weight change



cactus dragon fish_real_nemo

Figure 11: **Visualization of the weight change.** We randomly pick three cases from the RealFusion15 dataset, including cactus, dragon and fish.