Text2Stereo: Repurposing Stable Diffusion for Stereo Generation with Consistency Rewards

Aakash Garg[1] Libing Zeng[1] Andrii Tsarov[2] Nima Khademi Kalantari[1]
[1]Texas A&M University, [2]Leia Inc.
{aakash.garg80,libingzeng,nimak}@tamu.edu, andrii.tsarov@leiainc.com

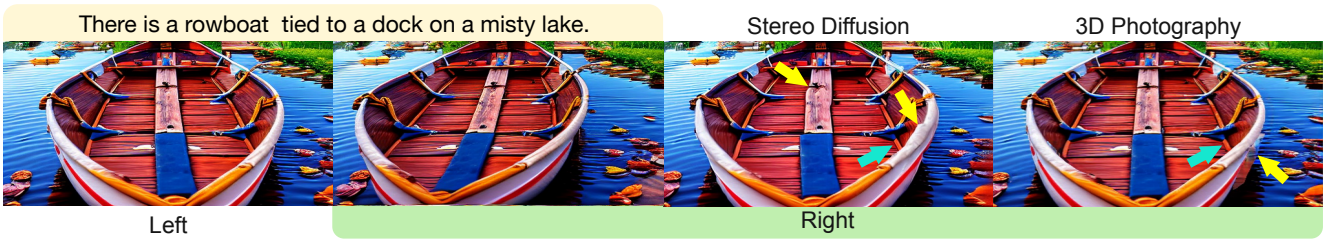

Figure 1. Given an input text prompt, our method synthesizes a stereo pair of left and right images. We use the generated left image as input to StereoDiffusion [46] and 3D Photography [39] to generate the right image. Both StereoDiffusion and 3D Photography use depth-based warping to transfer content from the input to the novel view. As such, they often struggle to create appropriate parallax effects for objects with continuously varying depth, as indicated by the cyan arrows. Moreover, since StereoDiffusion performs depth warping in the latent space, the warping is often not pixel-perfect, resulting in objectionable artifacts, as indicated by the yellow arrows. Finally, 3D Photography frequently struggles to reconstruct occluded regions (see the yellow arrow). Our approach, however, produces consistent, high-quality stereo images with wide baselines

Abstract

In this paper, we propose a novel diffusion-based approach to generate stereo images given a text prompt. Since stereo image datasets with large baselines are scarce, training a diffusion model from scratch is not feasible. Therefore, we propose leveraging the strong priors learned by Stable Diffusion and fine-tuning it on stereo image datasets to adapt it to the task of stereo generation. To improve stereo consistency and text-to-image alignment, we further tune the model using prompt alignment and our proposed stereo consistency reward functions. Comprehensive experiments demonstrate the superiority of our approach in generating high-quality stereo images across diverse scenarios, outperforming existing methods.

1. Introduction

With the rise in popularity of VR headsets (e.g., Meta Quest) and light field displays (e.g., Lume Pad), generating suitable content for these devices is becoming increasingly important. Although powerful diffusion models, such as Stable Diffusion [12, 34], allow the average user to produce creative images from text prompts, generating stereo images remains a major challenge.

One potential approach for generating stereo images is to first produce a single image using an existing diffusion model and then apply a single-image view synthesis method [17, 33, 39, 41, 44, 49] to reconstruct the other view. Most of these techniques [17, 33, 39, 49], however, generate novel views by warping the input image using monocular depth and inpainting the occluded regions. While these methods produce reasonable results with a small baseline, their results for larger baselines—this paper's focus—often contain objectionable artifacts. Specifically, depth-based warping often produces incorrect parallax effect for objects with continuous varying depth (see Fig. 1). Additionally, these methods usually reconstruct the occluded regions in a plausible but contextually inaccurate manner.

Recently, Wang et al. [46] tackle the problem of stereo image generation using a pre-trained Stable Diffusion model. Specifically, they follow the pipeline of previously mentioned methods, reconstructing stereo images through depth-based warping in the latent space of the diffusion model. As a result, they inherit the limitations of single-image view synthesis techniques. Furthermore, due to operating in the latent space, their warping is not pixel-perfect.

In this paper, we propose a novel diffusion-based approach for generating stereo image pairs from text prompts. Since stereo image datasets with large baselines are scarce,

we supplement the existing data [43] by creating additional data using the multi-view MVImgNet [54] dataset. Specifically, for each scene, we reconstruct it in 3D by optimizing a 3D Gaussian Splatting representation [15] on the input images. We then rendering several stereo pairs from the optimized representation for each scene and include them as training data.

Training a diffusion model from scratch is challenging due to the limited number of scenes in our dataset; the model can easily overfit and fail to generalize well. To address this issue, we leverage the strong priors of the pretrained Stable Diffusion [12, 34] and fine-tune it on our data, adapting it to the stereo generation task while retaining its generalization capabilities. However, the diffusion model outputs a single RGB image, while we are dealing with stereo image pairs. Therefore, we propose to vertically stack the left and right images to form a single RGB image, matching the output format of the diffusion model.

This fine-tuning process adapts the diffusion model to produce stereo images, but the tuned model suffers from two issues: **1)** the generated stereo pairs are often not geometrically consistent, as consistency is not enforced during the initial fine-tuning; and **2)** while the tuned model can generate stereo images for test prompts, the generated content is often not fully aligned with the text. To address these issues, we propose using the approach by Prabhudesai et al. [30] (AlignProp), which enables fine-tuning of diffusion models according to arbitrary but differentiable reward functions. Specifically, we introduce a stereo consistency reward function to improve geometric consistency, and use human preference score v2 (HPSv2) [50] to enhance text-to-image alignment.

Experimental results demonstrate that our approach produces consistent, high-quality stereo images that outperform existing methods. In summary, our paper makes the following key contributions:

- We propose fine-tuning Stable Diffusion on stereo images, adapting the model to generate stereo pairs while retaining its generalization capabilities.
- To improve geometric consistency and text-to-image alignment, we further tune the model using prompt alignment and our proposed stereo consistency rewards.
- We demonstrate that our approach produces stereo images with improved consistency and quality compared to existing methods.

2. Related Work

In this section, we review closely related work, focusing on 3D generation and single-image novel view synthesis. Additionally, we discuss approaches that fine-tune diffusion models based on specific reward functions, as we leverage this technique to enhance our stereo diffusion model.

2.1. 3D Generation

With recent advances in generative methods and 3D scene representations, such as neural radiance fields (NeRF) [26] and 3D Gaussian Splatting [15], there has been growing interest in 3D generation. One group of methods [2, 8, 11, 27, 35] integrates NeRF into generative adversarial networks (GANs) [9] to synthesize 3D content. While these methods produce high-quality results, they are limited to generating single objects.

Another category of methods leverages powerful 2D diffusion models [12, 34] as priors to reconstruct 3D scenes or objects. Specifically, DreamFusion [29] and its follow-up works [25, 38, 42, 48] use score distillation sampling (SDS) to optimize 3D representations like NeRF and 3D Gaussian Splatting. However, these methods often yield oversmoothed results due to SDS loss limitations, and their optimization process is computationally intensive.

Closer to our approach, some methods [21, 51] fine-tune Stable Diffusion [34] on large synthetic 3D object datasets [5] to produce multi-view images, which are then passed to a transformer network to generate the final 3D representation. Xie et al. [51] further refine the diffusion model using reinforcement learning to enhance consistency across generated multiview results. However, these methods are primarily focused on generating individual objects. In contrast, we target stereo generation for general scenes.

2.2. Single-Image Novel View Synthesis

Given a single image, a large number of methods [22, 31, 39, 44, 45] synthesize novel views by estimating intermediate 3D representations, such as layered depth images (LDI) [37] and multi-plane images (MPI) [56]. However, these methods are generally limited to narrow viewpoint changes and struggle to generate images with significant deviations from the input.

A group of recent techniques [3, 28, 53, 55] leverage powerful diffusion models [12, 34] for this task. These methods progressively project images into the 3D scene using estimated monocular depth and inpaint occluded regions with diffusion inpainting. However, depth-based warping often produces incorrect parallax, particularly for objects with continuous varying depth. Additionally, while inpainting models can fill in occluded regions with plausible content, they frequently lack contextual accuracy. Moreover, Wang et al. [46] (StereoDiffusion) focus specifically on stereo generation using depth-based warping in the latent space of Stable Diffusion [34]. However, in addition to the aforementioned issues, warping in the latent space also results in less precise pixel alignment.

2.3. Tuning Diffusion with Rewards

Reward fine-tuning has emerged as a promising approach to refining diffusion models, enabling the production of out-

puts that align with specific objectives. Inspired largely by reinforcement learning (RL), this approach has become central to applications requiring nuanced control over generation quality. For example, Lee et al. [18] apply reward-weighted regression on a curated dataset to address misalignments in factors such as object count, color consistency, and background quality. Methods like DDPO [1] and DPOK [7] use policy gradients in multi-step diffusion models [6], enhancing reward outcomes for aesthetic quality, image-text alignment, and compressibility.

In contrast to these RL-based methods, some approaches [4, 30] perform optimization by directly backpropagating gradients from a differentiable reward function, using gradient checkpointing [10] to do so efficiently. In our work, we employ such techniques, particularly AlignProp [30], to enhance the stereo consistency and prompt alignment of the generated results.

3. Background

In this section, we provide an overview of the concepts related to our approach: diffusion models [12, 34] and AlignProp [30].

3.1. Diffusion Models

Diffusion models [12] are probabilistic generative models that have recently achieved state-of-the-art performance in high-quality image synthesis. The process consists of a forward and reverse diffusion phase. In the forward process, noise is gradually added to an input image x_0 over T timesteps, producing a sequence of increasingly noisy images $x_0, \ldots, x_T$, eventually leading to a noise distribution x_T. Specifically, given a random noise image with normal distribution $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$, the image at time t is obtained by adding noise to the clean image according to $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t$ is derived based on the variance at each timestep.

The reverse process aims to denoise x_T back to x_0 using a learned denoising function ϵ_θ that takes the image at the current step x_t, and often a text prompt c, to estimate the noise $\hat{\epsilon}$, i.e., $\hat{\epsilon} = \epsilon_\theta(x_t, c, t)$. Given a set of images and corresponding text prompts $\mathcal{T} = \{(x_0^i, c^i)\}_{i=1}^N$, the model is trained by optimizing the following objective:

$$\mathcal{L}_d = \frac{1}{N} \sum_{(x_0^i, c^i) \in \mathcal{T}} \|\epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0^i + \sqrt{1-\bar{\alpha}_t}\epsilon, c^i, t) - \epsilon\|^2. \tag{1}$$

During inference, x_T is initialized with Gaussian noise, and the trained network ϵ_θ is used to progressively denoise it, ultimately producing a clean image x_0. Since both training and inference of diffusion models are computationally expensive, latent diffusion models (LDM) [34] propose performing the diffusion process in the latent space of a variational autoencoder (VAE) [16], significantly reducing computational load. In our work, we utilize an LDM, specifically Stable Diffusion, and adapt it to stereo generation task.

3.2. AlignProp

The goal of this approach is to fine-tune a pre-trained diffusion model to produce results aligned with a specific reward. Unlike the objective in Eq. 1, which operates on a single denoising step, AlignProp [30] maximizes the reward based on the model's output after multiple denoising steps. Specifically, given a dataset of training text prompts $\mathcal{C} = \{c^i\}_{i=1}^M$, AlignProp optimizes the diffusion model's parameters to maximize the following objective function:

$$\mathcal{L}_a = -\frac{1}{M} \sum_{c^i \in \mathcal{C}} R(\pi_\theta(x_T, c^i)), \tag{2}$$

where c^i represents a training prompt and R is the reward function, which might, for example, measure aesthetic quality or compressibility of the generated images x_0. Moreover, π_θ encapsulating the iterative denoising process into a single function, i.e., $x_0 = \pi_\theta(x_T, c)$.

Optimizing this objective by fully backpropagating through all denoising steps, however, leads to mode collapse. To address this issue, AlignProp proposes truncating gradient backpropagation at a random denoising step. This adjustment enables the optimization process to adapt the network according to the reward while avoiding mode collapse. In our work, we use AlignProp to enhance stereo consistency and prompt alignment in the generated results.

4. Methodology

The goal of our work is to train a diffusion model that generates consistent stereo image pairs with a large baseline from text prompts. Our approach, dubbed Text2Stereo, adapts the pre-trained Stable Diffusion [34] model to the stereo generation task. In the following sections, we first discuss our process for obtaining a dataset of stereo images and their corresponding text prompts (Sec. 4.1). We then describe our fine-tuning process, which consists of two stages: stereo adaptation (Sec. 4.2) and fine-tuning for consistency enhancement (Sec. 4.3). Overview of our approach is illustrated in Fig. 2.

4.1. Dataset Preparation

Since the goal of our work is to generate stereo images with a large baseline, we need a dataset of such stereo image pairs for training. Unfortunately, most existing stereo datasets are either captured with stereo cameras that have a small baseline [14] or are synthetic [24], making them unsuitable for our task. A notable exception is the dataset by NeRFStereo [43], which contains 270 scenes, each with 100 large baseline stereo images (total 27,000).

Figure 2. We show the overview of our approach comprising two stages (left and right). In the first stage, we fine-tune the pretrained Stable Diffusion model using LoRA [13] on our stereo image dataset, with left-right images stacked vertically. In the second stage, we further optimize our model using AlignProp [30] to enhance the stereo consistency and prompt alignment.

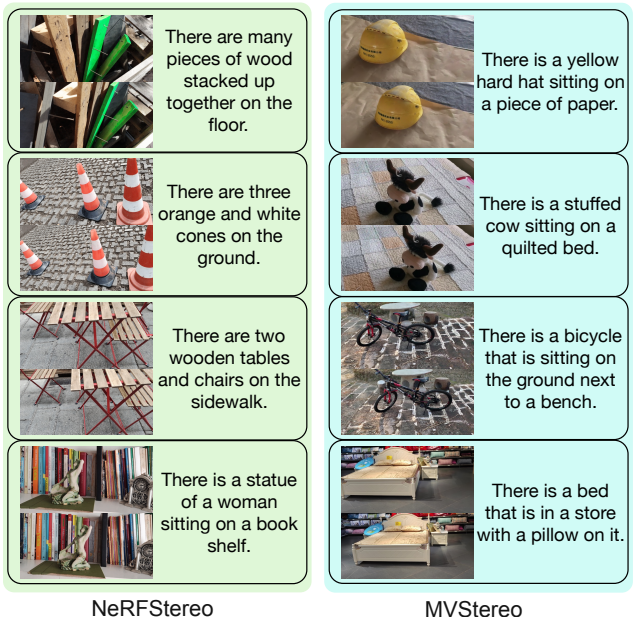

Figure 3. We present example images from our dataset. In addition to the existing stereo dataset from NeRFStereo [43], we introduce a newly generated dataset, MVStereo. This dataset is created by first reconstructing a 3D Gaussian splatting [15, 57] representation from multiview images obtained from MVImgNet [54], followed by sampling stereo images from the reconstructed model.

In our work, we supplement the NeRFStereo dataset by creating our own stereo image dataset, coined MVStereo, using the multi-view MVImgNet dataset [54]. Specifically, we select a subset of 234 scenes from MVImgNet [54] and reconstruct them in 3D by optimizing a 3DGS representation [15], utilizing the approximately 30 images available for each scene. Although the 3DGS optimization [15] yielded reasonable results, we found that the method proposed by Zhu et al. [57] produced better outcomes with fewer floaters and less blurriness, which is why we adopt their approach to reconstruct the scenes.

Once the 3DGS representation is obtained, we render 7373 stereo pairs by setting up stereo cameras at various views. We carefully position the cameras around the input scene to ensure the rendered images do not contain floaters or blurry content. Together, we use 30982 stereo images across 458 scenes; 234 from MVStereo and a subset of 224 scenes from NeRFStereo, both encompassing diverse indoor and outdoor environments. Since our goal is text-based stereo generation, we also require corresponding text prompts for each stereo image in our dataset, which we obtain using the BLIP captioning model [19]. Some examples of stereo images and their corresponding captions from both NeRFStereo and our MVStereo are shown in Fig. 3.

4.2. Stereo Adaptation

Given a dataset of stereo images and their corresponding text prompts, $\{x_l^i, x_r^i, c^i\}_{i=1}^N$, we aim to train a stereo generator. However, due to the relatively small size of our dataset, training a diffusion model from scratch is impractical. Even with a larger dataset, ensuring generalization to diverse text prompts would be challenging. To address this, we leverage the strong prior knowledge in a pre-trained diffusion model, specifically Stable Diffusion [34], and adapt it for the stereo generation task.

The primary challenge here is that Stable Diffusion is designed to produce a single RGB image, whereas our objective is to generate stereo image pairs. Inspired by Instant3D [21], we propose stacking our stereo image pairs, x_l^i and $x_r^i \in \mathbb{R}^{256\times512\times3}$, vertically to form a single RGB image, $x^i \in \mathbb{R}^{512\times512\times3}$ (see Fig. 2). This stacked representation, along with the corresponding text prompts, forms our training dataset, $\mathcal{T} = \{x^i, c^i\}_{i=1}^N$, which we use to fine-tune Stable Diffusion according to the objective in Eq. 1. Note that we do not provide camera pose information as the input to the network. However, since the training data is mainly composed of stereo images with a large baseline, our trained model has the ability to produce such stereo pairs.

Fine-tuning Stable Diffusion by directly optimizing its parameters, θ, can lead to overfitting due to the small size of our dataset. To address this, we employ Low-Rank Adaptation (LoRA) [13], which mitigates overfitting by freezing the diffusion parameters θ and modulating them through a trainable layer with significantly fewer parameters. Specifically, each linear layer, initially represented as $h = Wx$

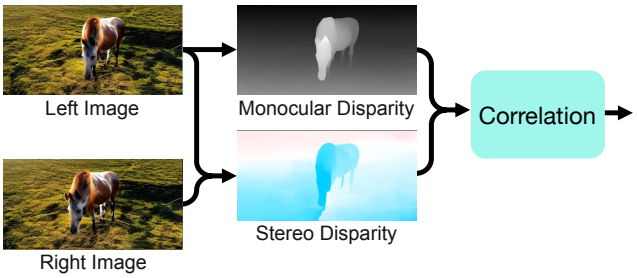

Figure 4. Given a stereo image pair, we estimate the stereo disparity using both images and monocular disparity using only the left image. The correlation between the two maps will then serve as our stereo consistency reward function.

with $W \in \mathbb{R}^{d\times d}$, is modified to $h = Wx + BAx$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$, with $r \ll d$. By freezing W and updating only A and B in each layer of Stable Diffusion, we adapt the model for stereo generation while preserving its ability to generalize.

4.3. Fine-Tuning for Consistency Enhancement

Our fine-tuned model, as shown in Fig. 9 (Base), produces vertically stacked stereo images for unseen text prompts. However, the generated results exhibit two issues. First, while the content in the left and right images shifts with the camera's perspective, there are often deformations and inconsistencies between objects in the two views. This issue primarily arises because our initial fine-tuning lacks a mechanism to enforce consistency between the left and right images. Second, we observe that after fine-tuning, the diffusion model generates images that, in some cases, are not fully consistent with the text prompt.

To address these issues, we propose further fine-tuning the model to enhance its stereo and prompt consistency using AlignProp [30]. The main challenge here is designing an appropriate reward function R that measures the stereo and prompt consistency of the generated images. To this end, we propose a reward function consisting of three terms as follows:

$$R = \alpha R_s + \beta R_p + \gamma R_c \tag{3}$$

where R_s, R_p, and R_c refer to the stereo consistency, prompt consistency, and convergence rewards, described below. Moreover, $\alpha = 0.25$, $\beta = 0.75$, and $\gamma = 0.25$ define each term's weight.

Stereo Consistency: Since there is currently no well-established mechanism for checking the stereo consistency between generated stereo pairs, we need to design our own metric. Our key idea is that for stereo images to be consistent, the stereo and monocular disparities should align. This ensures that the model avoids trivial solutions, such as duplicating content, where the stereo disparity is zero, but the monocular disparity still reflects the correct depth. To achieve this, we estimate the stereo disparity as $d^s = \Phi(x_l, x_r)$ and the monocular disparity as $d^m = \Psi(x_l)$ and measure their similarity. Since the monocular disparity is relative, comparing these two disparities directly using pixel-wise metrics, such as L_2, is not effective. Therefore, we propose measuring their similarity using Pearson correlation, as follows:

$$R_s = \frac{\sum_p (d^m(p) - \bar{d^m})(d^s(p) - \bar{d^s})}{\sqrt{\sum_p (d^m(p) - \bar{d^m})^2 \sum_p (d^s(p) - \bar{d^s})^2}}, \tag{4}$$

where $\bar{d^m}$ and $\bar{d^s}$ are the average monocular and stereo disparities over all pixel coordinates p. In our implementation, we use DepthAnythingV2 [52] to estimate monocular disparity. For stereo disparity, we initially experimented with CREStereo [20], but found it to be sensitive to imperfections in the stereo pairs. Therefore, we use SEA-RAFT [47] to estimate the optical flow between the two images, and then use the x-coordinate as the disparity. We illustrate our stereo consistency reward in Fig. 4.

Prompt Consistency: We use the human preference score v2 [50], which trains a CLIP model [32] on a large annotated dataset. This score reliably measures the consistency between the text prompt and the generated images, and we adopt it as our prompt consistency metric, R_p.

Convergence: Stereo images captured with cameras with parallel optical axis have convergence at infinity, i.e., the objects at infinite depth will have zero disparity. Through the finetuning, however, the diffusion model may generate stereo images with convergence at the middle of the scene, i.e., objects further away from the convergence will have negative disparity. To avoid this issue, we introduce the following reward to penalize the negative disparities:

$$R_c = -\frac{\|\max(-d^s(p), 0)\|_1}{\max(\max(-d^s), 1)}. \tag{5}$$

Here, the denominator is a normalization factor that prevents large negative disparities (greater than 1) from causing a spike in the loss.

5. Results

In this section, we first describe the implementation details. We then show comparisons against state-of-the-art methods and demonstrate the impact of various components of our approach.

5.1. Implementation Details

We implement our method in PyTorch and use the pre-trained Stable Diffusion v1.5 as our base model. Additionally, we utilize LoRA with rank 4, and inject it into every U-Net cross-attention layer. In the initial training phase, we

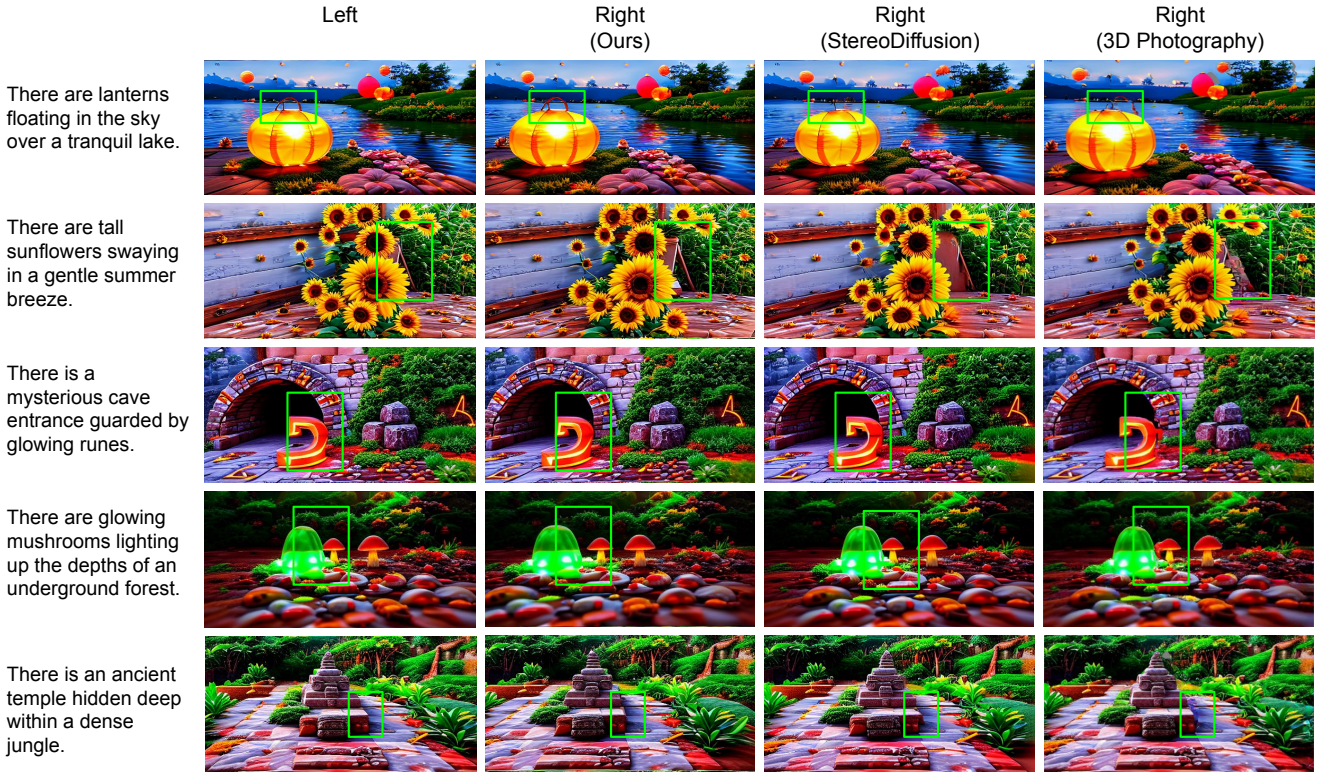

Figure 5. Qualitative comparison of our approach against StereoDiffusion [46] and 3D Photography [39] on five test prompts. Given a text prompt, we generate a stereo pair and use the left image as input for the other methods to reconstruct the right image. StereoDiffusion, which performs warping in the latent space, often distorts objects (top two rows) or fails to position them correctly (bottom three rows). For example, note that StereoDiffusion does not produce the gap between the mushrooms in the fourth example. 3D Photography struggles with depth inaccuracies (e.g., thin structures in the top row) and fails to reconstruct occluded areas (bottom four rows). In contrast, our approach produces consistent, high-quality results with wide baselines.

optimize the model with a cosine learning rate scheduler starting at 1e-4, employing a batch size of 4 with gradient accumulation over 4 steps for a total of 4000 iterations, which took roughly 6 hours on a single A100 GPU. Subsequently, for optimization using consistency rewards, the model undergoes further fine-tuning for 300 iterations with a batch size of 100 prompts per step utilizing 4 A100 GPUs for a day.

5.2. Comparisons

We demonstrate the effectiveness of our approach by providing comparisons against StereoDiffusion [46] and 3D Photography [39]. Specifically, in each case, we generate the stereo images given a text prompt with our approach and use the left image as the input to the other techniques for reconstructing the right image. As such, we only compare the reconstructed right images. 3D Photography reconstruct the novel image through depth-based warping and inpainting, while StereoDiffusion, performs the warping in the latent space of a diffusion model.

Figure 5 shows comparisons against the other approaches on five test prompts. Since the warping in the latent domain is not precise, StereoDiffusion either distort the objects (top two scenes) or is unable to move various objects to the appropriate location (last three scenes). Similarly, 3D Photography struggles in cases where the depth is inaccurate (e.g., the thin structure in the top scene) and is unable to properly reconstruct the occluded areas (bottom four scenes). In contrast, our approach produces consistent high-quality results with large baselines.

We further evaluate the consistency of generated results in Figure 6. Specifically, we use the pair of images generated by each method as the input to Splatt3R [40] to obtain the corresponding 3D Gaussian splatting (3DGS) [15] representation. We then render the 3DGS representation from a novel view and compare the renderings. The key idea is that if the stereo images are consistent, Splatt3R will produce a high-quality 3DGS representation and thus the rendered images will be of high quality. As shown in Figure [40], the novel view images for both StereoDiffusion and 3D Photography contain distracting artifacts. In contrast, the rendering by our approach has clear object boundaries, demonstrating the consistency of our generated left and right images.

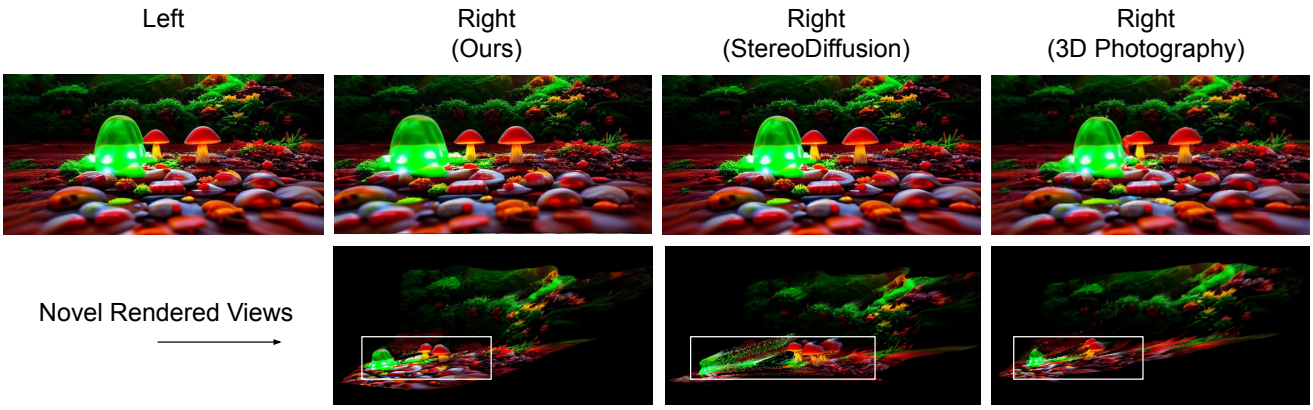

Figure 6. Evaluation of stereo consistency using Splatt3R [40]. Given the text prompt: "There are glowing mushrooms lighting up the depths of an underground forest," we generate a stereo pair with each method and use it as input to Splatt3R to obtain a 3D Gaussian splatting (3DGS)[15] representation. The 3DGS model is then rendered from a novel viewpoint. StereoDiffusion[46] and 3D Photography [39] produce stereo images with inconsistencies, leading to artifacts in the rendered view. In contrast, our approach generates more consistent stereo pairs, resulting in a high-quality 3DGS representation with clear object boundaries.

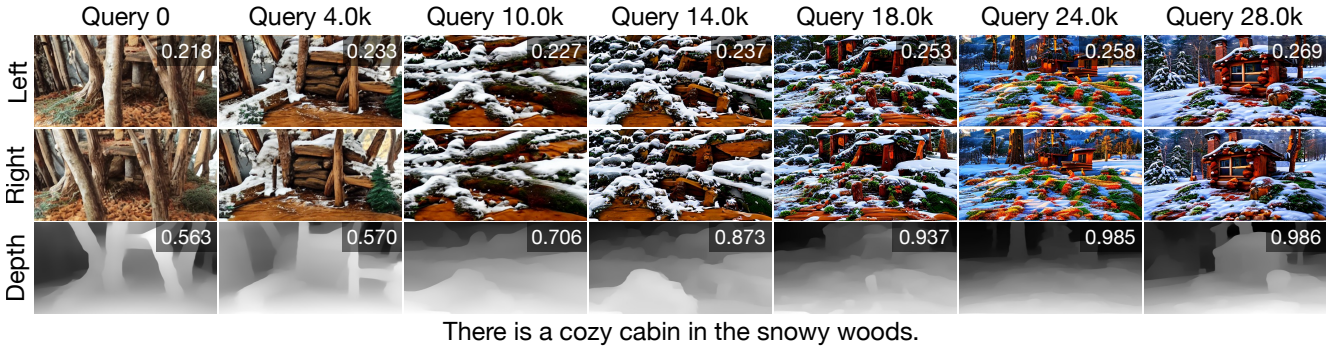

Figure 7. We demonstrate the results through various stages of the training process. The numbers in the first and third rows correspond to the prompt alignment score and stereo consistency score, respectively. As seen, both stereo consistency and prompt alignment exhibit improvements throughout the reward optimization process.

5.3. Analysis

We begin by showing the progressive improvement of stereo consistency and prompt alignment with the optimization of our consistency rewards (Eq. 3) in Fig. 7. The values presented in the first and third rows represent the prompt and stereo consistency scores, respectively. Each column is obtained by generating the results of the network after certain number of reward queries. Here the reward query refers to the number of times the reward function is evaluated during the fine-tuning process, e.g., one iteration with a batch size of 100 leads to 100 reward queries. As seen both the prompt and stereo consistency of the results improve during the optimization (despite some fluctuations). Particularly, the stereo depth maps (third row) improve significantly through the fine-tuning process.

We further evaluate how the prompt and stereo consistency rewards evolve during the optimization for both training and test prompts in Figure 8. Specifically, we perform the evaluation on 100 training prompts, drawn from the training dataset, and 100 test prompts, generated by ChatGPT 4o. We ensure that the test prompts are substantially different from the training prompts. As shown in Fig. 8, during the reward optimization process, both the mean values of the prompt and stereo consistency rewards exhibit a progressive increase, while their standard deviations decrease. These trends indicate an enhancement in the model's capability to generate images with improved prompt alignment and stereo consistency. Additionally, in all plots, the training and testing results follow similar trajectories, suggesting that optimizing for consistency objectives does not compromise the model's generalization ability.

Finally, we conduct ablation studies to evaluate the impact of various components of our system both numerically (Table 1) and visually (Figure 9). As seen, the base model (without consistency tuning) produces results with

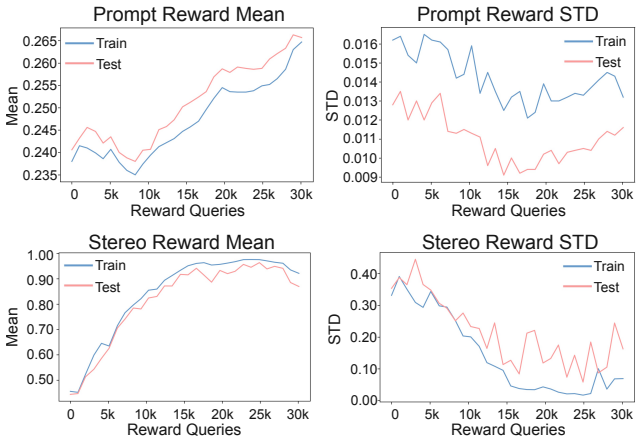

Figure 8. We show the rewards averaged over 100 training and test prompts during the optimization process. The mean values of both prompt and stereo consistency rewards increase, while their standard deviations decrease, during the reward optimization process. These trends suggest an improvement in the model's ability to generate images with enhanced prompt alignment and stereo consistency. Furthermore, across all plots, both training and testing results demonstrate similar trends, indicating that optimizing for consistency objectives does not adversely affect the model's generalization capabilities.

Table 1. We quantitatively evaluate the impact of various components of our system in terms of the stereo and prompt consistency scores.

Condition	Stereo Score	Prompt Score
Base	0.414 ± 0.327	0.237 ± 0.012
Base + Stereo	$\mathbf{0.985 \pm 0.012}$	0.205 ± 0.010
Base + Stereo + Prompt, 10 prompts	0.877 ± 0.192	0.258 ± 0.010
Base + Stereo + Prompt, 100 prompts	0.937 ± 0.098	0.254 ± 0.010
Base + Stereo + Prompt, 750 prompts (Ours)	0.949 ± 0.078	$\mathbf{0.264 \pm 0.011}$
Base + Stereo + Prompt, 2000 prompts	0.940 ± 0.091	0.263 ± 0.011

poor stereo consistency and prompt alignment. Optimizing with only the stereo consistency reward (Base + Stereo) improves stereo consistency but significantly reduces prompt alignment. Optimizing with both stereo and prompt consistency rewards improves both scores. Additionally, increasing the training prompts to 750 consistently enhances the results, after which the improvement stabilizes. Note that compared to the variant without prompt consistency (Base + Stereo), our method produces results with slightly lower stereo scores. However, as shown in Fig. 9, our approach produces stereo images with the best trade off between stereo and prompt consistency.

6. Conclusion, Limitations, and Future Work

We have presented a novel method for generating wide-baseline stereo images by adapting a pre-trained diffusion model to this task. Specifically, we first fine-tune a Stable Diffusion model on a stereo dataset to produce vertically

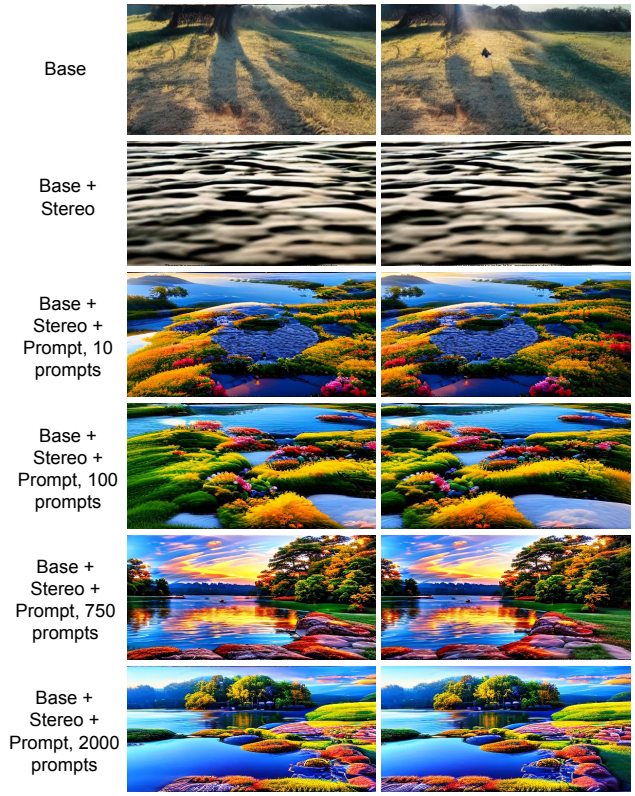

Figure 9. We show the impact of different component of our system visually. The test prompt is: "There is a serene sunrise over a calm lake, promising a day filled with wonder."

stacked left and right images. Moreover, to improve stereo consistency and prompt alignment, we propose specific reward functions used to further tune the model. In particular, we introduce a stereo consistency reward that calculates the similarity of monocular and stereo disparities using Pearson correlation. Through experimental results, we demonstrate that our approach outperforms existing methods.

Despite producing high-quality results, our approach has a few limitations. For example, it currently does not provide a mechanism to control the baseline of the generated stereo images. In the future, it would be interesting to investigate a way to use the baseline as an input to the diffusion process to enhance controllability. Additionally, our approach can generate stereo images only from a text prompt and cannot reconstruct stereo images from a single image. One potential solution is to invert the image into our diffusion process to reconstruct the other view. We leave the investigation of this strategy to future work. Finally, we observed that the captions generated by the BLIP model are short and could sometimes be inaccurate. In the future, it would be interesting to utilize more descriptive image captioning approaches such as LLaVA [23], combined with human verification of the captions, as they have been shown to improve the image generation quality [36].

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[3] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 2

[4] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 3

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2

[6] Ying Fan and Kangwook Lee. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023. 3

[7] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[8] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[10] Audrunas Gruslys, Rémi Munos, Ivo Danihelka, Marc Lanctot, and Alex Graves. Memory-efficient backpropagation through time. *Advances in neural information processing systems*, 29, 2016. 3

[11] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4

[14] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *CVPR*, 2020. 3

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 4, 6, 7

[16] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[17] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1169–1178, 2023. 1

[18] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 4

[20] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 5

[21] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4

[22] Qinbo Li and Nima Khademi Kalantari. Synthesizing light field from a single image with variable mpi and two network fusion. *ACM Transactions on Graphics*, 39(6), 2020. 2

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 8

[24] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 3

[25] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8446–8455, 2023. 2

[26] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[27] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[28] Hao Ouyang, Kathryn Heal, Stephen Lombardi, and Tiancheng Sun. Text2immersion: Generative immersive scene with 3d gaussians. *arXiv preprint arXiv:2312.09242*, 2023. 2

[29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[30] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 2, 3, 4, 5

[31] Guo Pu, Peng-Shuai Wang, and Zhouhui Lian. Sinmpi: Novel view synthesis from a single image with expanded multiplane images. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 2

[32] 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, 2021. 5

[33] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021. 1

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 4

[35] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[36] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation, 2023. 8

[37] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242, 1998. 2

[38] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[39] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8028–8038, 2020. 1, 2, 6, 7

[40] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibarated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 6, 7

[41] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 1

[42] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[43] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 855–866, 2023. 2, 3, 4

[44] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 551–560, 2020. 1, 2

[45] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–317, 2018. 2

[46] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. Stereodiffusion: Training-free stereo image generation using latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7416–7425, 2024. 1, 2, 6, 7

[47] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2025. 5

[48] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[49] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020. 1

[50] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 5

[51] Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, and Arie E. Kaufman. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6369–6379, 2024. 2

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 5

[53] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 2

[54] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 2, 4

[55] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[56] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2

[57] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European Conference on Computer Vision*, pages 145–163. Springer, 2024. 4