StereoEdit: A Diffusion-Based Framework for Stereo-Consistent Image Editing

Baolin Liu, Zongyuan Yang, Yingde Song, and Yongping Xiong*

The State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China
{baolin,yangzongyuan0,songyingde,ypxiong}@bupt.edu.cn

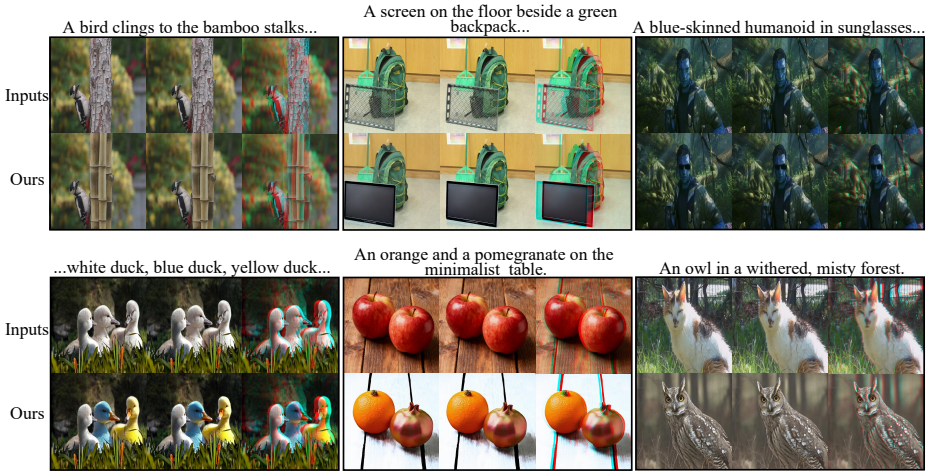

Fig. 1: The partial results of StereoEdit are presented, where the first two columns in each group correspond to the left and right views, and the third column shows the anaglyph stereo image. Our method demonstrates superior performance, highlighting high fidelity to the prompts, stereo consistency, and visual realism.

Abstract. While recent diffusion models have achieved remarkable success in visual content editing, extending them to stereo image editing remains highly challenging due to the lack of explicit stereo geometric consistency alignment, which leads to cross-view inconsistencies and geometric distortions. We propose StereoEdit, a training-free diffusion framework designed to achieve geometrically consistent and structurally coherent stereo editing. StereoEdit introduces two complementary modules that explicitly enforce geometric reliability during the denoising process. The Geometrically-Gated Attention Reference (GGAR) adaptively filters cross-view attention based on disparity-aware confidence, ensuring stable correspondences across epipolar regions, while the Geometry-

* Corresponding author.

Aware Latent Alignment (GALA) performs confidence-weighted, bidirectional latent corrections to maintain parallax alignment throughout the diffusion process. Together, these modules enable robust and consistent stereo edits. Extensive experiments on diverse stereo benchmarks demonstrate that StereoEdit significantly improves cross-view coherence and visual realism, establishing a new foundation for controllable and immersive stereoscopic content editing.

Keywords: Stereo Image Editing · Stereo Consistency · Diffusion Models

1 Introduction

Stereoscopic 3D has long been on the verge of becoming a mainstream consumer medium, encompassing 3D movies, television, and gaming. More recently, the emergence of head-mounted 3D displays (e.g., AR/VR headsets) and dual-lens smartphones has renewed interest in stereoscopic 3D [1, 29, 33, 40], spurring a variety of research directions, including stereoscopic inpainting [31, 39], video stabilization [10], and stereo generation and editing [41, 46]. Among these efforts, the creation and editing of stereoscopic 3D content remains a particularly challenging task. In this work, we focus on text-guided, stereo-consistent editing, a task that has not yet been explored in prior studies. Early research on stereoscopic image editing has explored geometry-preserving warping and distortion-aware optimization for depth adjustment and retargeting [4, 13, 41], as well as disparity-consistent stylization and appearance transfer [5, 17]. Although these methods maintain structural coherence across views in specific scenarios, they are generally limited to task-specific operations and lack semantic flexibility, making them unsuitable for modern content creation that demands free-form, text-driven editing. Moreover, research in this area has slowed in recent years, and many early methods remain unavailable to the public, further hindering practical progress.

The advent of diffusion models such as Stable Diffusion [24] has revolutionized 2D image editing, enabling photorealistic, text-guided manipulation. Extending these models to stereoscopic pairs, however, remains challenging. Naively applying diffusion to each view independently leads to severe texture mismatches and geometric distortions, breaking binocular coherence. This arises from a fundamental gap: 2D diffusion operates in an image-centric latent space, while stereo consistency depends on rigid geometric relationships that the models were never trained to enforce. One might consider adapting recent video editing frameworks [3, 9, 35]. However, these methods are intrinsically designed for temporally sequential frames that capture the scene from a single moving viewpoint. Their alignment objectives emphasize temporal continuity rather than spatial coherence. In contrast, stereoscopic image sets consist of simultaneous projections of the same 3D scene from different viewpoints along a fixed horizontal baseline. This setting is governed not by temporal dynamics but by strict epipolar geometry [14, 36, 45], which defines precise pixel-level correspondences across views.

Applying video-based alignment to such data ignores these geometric constraints, leading to view-dependent texture drift, structural inconsistencies, and a loss of cross-view coherence. Consequently, these methods struggle to maintain texture and structural consistency in the edited dual-view images, potentially resulting in visual inconsistencies during display. These limitations underscore the need for a specialized framework to address the unique challenges of stereoscopic views.

Existing explicit 3D editing methods, typically based on representations such as NeRF [42] and 3DGS [2], enable scene editing [6, 28] but typically require known camera poses or multi-view initialization, e.g., via COLMAP [27], which is challenging for stereo pairs. While recent variants, such as COLMAP-free [7] or transformer-based initialization (e.g., VGGT [30]), reduce this dependency, they remain fragile for small-baseline stereo pairs. In such cases, the combination of limited parallax and only two views provides weak geometric constraints, often leading to unstable reconstruction, geometry collapse, or texture loss (details in the Supplementary Materials). These limitations motivate a geometry-aware framework that enforces stereo consistency directly at the image level, without explicit modeling.

To this end, we propose StereoEdit, a zero-shot, geometry-aware diffusion framework that leverages the strong editing capabilities of pretrained text-to-image models, focusing solely on maintaining stereo consistency. Our key insight is that stereo consistency can be enforced by leveraging geometric priors such as disparity and occlusion cues to guide attention and latent alignment throughout denoising. Specifically, StereoEdit introduces a unified geometry-aware inference pipeline composed of two complementary components. (1) The Geometrically-Gated Attention Reference (GGAR) module integrates geometric priors directly into the diffusion model's attention mechanism. It warps key features across views using predicted disparity maps, applies a soft Gaussian mask for epipolar alignment, and introduces reliability-aware gating derived from forward–backward disparity consistency. These mechanisms ensure that cross-view attention focuses only on geometrically valid correspondences, preserving local coherence while suppressing unreliable or occluded regions. (2) The Geometry-Aware Latent Alignment (GALA) module operates at the latent level to correct residual misalignment between views. It estimates bidirectional warping errors and adaptively refines the less consistent view under weighting, gradually improving global geometric consistency throughout denoising. Together, GGAR and GALA embed geometric awareness at both the feature and latent levels, enabling stereo-consistent editing.

We validate StereoEdit on diverse stereo datasets and synthetic benchmarks, comparing it with state-of-the-art editing methods. Extensive qualitative and quantitative evaluations demonstrate that our method significantly outperforms existing approaches in stereo coherence, texture alignment. In summary, our key contributions are:

1. We propose StereoEdit, a zero-shot diffusion framework for stereo-consistent, text-guided image editing that requires no retraining.

2. We design a geometry-aware inference mechanism that combines Geometrically-Gated Attention Reference for attention-level geometric filtering and Geometry-Aware Latent Alignment for latent refinement, ensuring both local and global stereo coherence.
3. We conduct comprehensive experiments on StereoEdit, demonstrating significant improvements in both qualitative and quantitative evaluations over existing methods.

2 Related Work

2.1 Stereo Editing

Significant progress has recently been made in the field of stereoscopic image editing, encompassing both geometric and content-level manipulations. At the geometric level, research has primarily concentrated on depth adjustment and image retargeting. For instance, Yan et al. [41] extended the shift-map method to enable effective depth adjustment of stereoscopic images. Chai et al. [4] employed a roundness-preserving warping model to maintain a natural appearance during retargeting, while Imani et al. [13] introduced an unsupervised framework aimed at minimizing geometric distortion in salient regions. At the content level, research aims to consistently extend monocular appearance edits, like style transfer, to stereo pairs. Representative works include the disparity-constrained method introduced by Chen et al. [5] and the consistent stylization achieved by Kučera et al. [17] using seamless patches. However, existing methods lack flexible control over appearance and textures, which hinders free-form, semantic-level editing. Therefore, this work explores the potential of leveraging general-purpose image editing models for stereoscopic view editing.

2.2 Text-guided Video Editing

Recent training-free video editing methods use pre-trained diffusion models to adjust appearance, objects, motion and structure without losing temporal coherence. For precise control, VideoGrain [43] modulates space-time attention to enable multi-granular texture and semantic refinements, ReAtCo [34] integrates re-attentional mechanisms into diffusion pipelines for enhanced spatial controllability, and VideoDirector [35] employs spatial-temporal decoupled guidance and multi-frame null-text optimization for accurate editing. To mitigate motion inconsistencies and non-rigid deformations, TV-LiVE [15] injects layer-informed vitality into position embeddings for effective feature integration, TokenFlow [9] propagates diffusion features across frames for seamless trajectory alterations, and VidToMe [20] fuses inter-frame tokens to bolster zero-shot temporal consistency. Unfortunately, these video editing approaches predominantly emphasize temporal features, overlooking the unique properties of binocular disparity in stereo views, and thus yield suboptimal performance in handling stereoscopic consistency.

2.3 Text-guided 3D Editing

Recent 3D editing methods based on 3D Gaussian Splatting (3DGS) and NeRF can now achieve view-consistent modifications through neural representations. 3DGS approaches such as InterGSEdit [37], which uses interactive view selection with geometry-aware attention, DYG [23], which employs drag-based editing with triplane scaffolds, and EditSplat [18], which integrates multi-view fusion with attention-guided trimming, enable high-fidelity text-driven local edits. NeRF-based methods including Multi-StepNeRF [22], which supports sequential instruction-guided editing through iterative refinement, and Instruct-NeRF2NeRF [11], which applies image-conditioned diffusion models for text-instructed updates while preserving scene coherence, also demonstrate strong editing capabilities. Despite these advances, we observe in our experiments that such methods often exhibit instability or outright failure in stereoscopic editing scenarios, particularly on datasets with limited disparity. This instability arises not only from inaccurate or missing camera poses, but also from the inherently weak geometric constraints of stereo inputs, revealing a critical gap in robustness for real-world stereo editing tasks.

3 Preliminaries

Our framework builds upon three key components: stereo disparity, latent diffusion, and self-attention, which together provide the foundation for geometry-aware stereo editing. This section briefly reviews these components to establish the context for our stereo editing framework.

Stereo Disparity. A stereo image pair (I_L, I_R) represents the same 3D scene from two horizontally displaced viewpoints. Each pixel in I_L corresponds to a point along an epipolar line in I_R, satisfying the fundamental relation $x_R^\top F x_L = 0$, where F is the fundamental matrix. The disparity between corresponding pixels encodes scene depth, providing a strong geometric prior. Our framework leverages these disparities to guide both local feature interactions and global latent alignment, ensuring structural consistency across views.

Latent Diffusion Models. Latent diffusion models [12, 24] iteratively denoise a latent variable z_t to reconstruct a clean latent representation. At each step, the model predicts and removes noise conditioned on external guidance. The latent space captures both global structure and fine details, enabling geometry-aware latent refinement without additional training. This forms the basis of our Geometry-Aware Latent Alignment (GALA), which corrects accumulated cross-view drift during denoising.

Self-Attention. Self-attention allows the model to capture long-range dependencies within the latent space. Given queries Q, keys K, and values V, the output is computed as $\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$. In our framework, we modulate self-attention scores using a novel disparity-based method, selectively controlling the influence of cross-view features. This mechanism underlies the Geometrically-Gated Attention Reference (GGAR), which preserves local geometric coherence during diffusion.

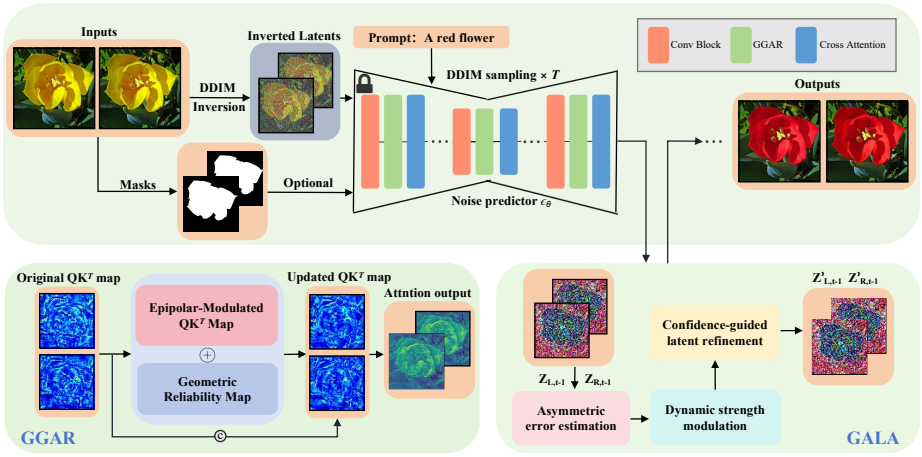

Fig. 2: Overview of StereoEdit. The GGAR (Geometrically-Gated Attention Reference) and GALA (Geometry-Aware Latent Alignment) jointly ensure geometry-aware and consistent stereo editing.

4 Method

Our goal is to enable stereo-consistent, text-guided editing within a training-free diffusion framework. As shown in Fig. 2, we propose a unified geometry-aware inference framework that incorporates explicit geometric reasoning into both feature interactions and latent evolution, allowing the diffusion model to maintain correspondence and structural coherence throughout the denoising process.

Given a stereoscopic pair (I_L, I_R), where I_L and I_R denote the left and right images of a stereo pair, respectively, along with a target text prompt, both images are first inverted into a shared latent z_T via DDIM inversion and then jointly denoised from step T to 0 under text guidance. During denoising, geometry-aware mechanisms operate at multiple levels to preserve stereo consistency. GGAR filters cross-view attention using disparity-based warping, epipolar alignment, and reliability gating, ensuring that each token attends only to geometrically plausible regions. GALA further mitigates residual misalignments by comparing bidirectional warping errors and adaptively refining the less consistent latent. Together, these modules enforce both local correspondence and global structural coherence across the entire denoising trajectory.

4.1 Geometrically-Gated Attention Reference

Existing methods emphasize temporal smoothness or independent 2D alignment, lacking explicit mechanisms to enforce the geometric correspondences required for binocular consistency. To overcome this limitation, we propose a geometry-guided attention refinement (GGAR) module that enforces local stereo feature coherence by embedding epipolar constraints directly into the attention process.

By integrating disparity-based warping and reliability-aware gating, GGAR enables geometry-consistent feature fusion and coherent generation across stereoscopic views. For clarity, we describe the formulation for the left-view–referenced branch; an identical formulation is applied symmetrically in the opposite direction (right-to-left), forming a bidirectional GGAR module.

Formally, within a self-attention layer of the left-view latent, we compute standard query, key, and value matrices (Q_L, K_L, V_L) and obtain (K_R, V_R) for the right view. Using the disparity map $D_{R\to L}$, predicted by a disparity estimation model, we warp them to the left-view space, producing $(K_{R\to L}, V_{R\to L})$. To incorporate epipolar alignment, we introduce an epipolar mask matrix $M \in \mathbb{R}^{N\times N}$, where (N) denotes the number of spatial tokens, each corresponding to a latent feature in the 2D latent grid of one view. The mask is defined as $M_{i,j} = \exp\left(-\frac{(row_i - row_j)^2}{2\sigma^2}\right)$ for a soft Gaussian decay based on row distance between tokens i (left) and j (warped right), with σ as a bandwidth hyperparameter to accommodate minor estimation errors. This mask suppresses attention across non-epipolar regions, focusing computation on geometrically plausible horizontal bands.

The GGAR module then modulates the cross-view attention scores with both the epipolar mask and a geometric reliability map $\mathbf{R}_{R\to L}$ as soft gates:

$$S'_{\text{cross}} = \left(Q_L K_{R\to L}^T \odot M\right) + \lambda \cdot \log(\mathbf{R}_{R\to L}), \tag{1}$$

where λ controls the gating strength. Here, $\mathbf{R}_{R\to L}$ measures the geometric consistency between the predicted left–right disparity pairs $(D_{R\to L}, D_{L\to R})$, assigning high confidence to consistent regions and low confidence to areas with disparity mismatch, occlusion, or geometric ambiguity. The logarithmic modulation sharpens this gating effect, such that reliable regions ($\mathbf{R}_{R\to L} \approx 1$) remain largely unaffected, while unreliable or occluded areas ($\mathbf{R}_{R\to L} \to 0$) are effectively suppressed after softmax normalization. A detailed formulation of $\mathbf{R}_{R\to L}$ is provided in the supplementary materials.

The final attention output concatenates the self-view and gated cross-view contributions:

$$\text{Attn} = \text{Softmax}\left(\frac{[S_{\text{self}}, S'_{\text{cross}}]}{\sqrt{d}}\right) \cdot \text{concat}(V_L, V_{R\to L}), \tag{2}$$

where $S_{\text{self}} = Q_L K_L^\top$ denotes the standard attention scores without geometric modulation. By incorporating geometry directly into the attention mechanism via disparity warping, reliability gating, and epipolar alignment, GGAR enables the diffusion model to attend selectively by trusting correspondences where stereo cues are strong and ignoring them where they are uncertain or epipolarly inconsistent. This ensures that local feature generation remains coherent across both views from the earliest stages of denoising.

4.2 Geometry-Aware Latent Alignment

We observe that residual global structural misalignments may still occur in the latent features, which can affect the structural distribution across views even

with GGAR enforcing geometry-aware coherence within the U-Net. To mitigate this, we introduce GALA, a lightweight and training-free module that performs geometry-guided correction after each denoising step.

Error estimation. Given the predicted latents $(z_{L,t-1}, z_{R,t-1}) \in \mathbb{R}^{C\times H\times W}$, we measure bidirectional inconsistencies by warping one view into the other:

$$\begin{aligned} E_{L\to R} &= \|\mathrm{Warp}(z_{L,t-1}, D_{L\to R}) - z_{R,t-1}\|_1, \\ E_{R\to L} &= \|\mathrm{Warp}(z_{R,t-1}, D_{R\to L}) - z_{L,t-1}\|_1, \end{aligned} \tag{3}$$

where $\mathrm{Warp}(\cdot)$ uses the the predicted disparity map . If $E_{L\to R} > E_{R\to L}$, we conclude that the right view is less geometrically consistent (as the mismatch is larger when projecting from left to right, implying the right deviates more).

Dynamic strength modulation. We adaptively assign correction strengths according to the error asymmetry:

$$\begin{aligned} \beta_{L,t} &= \beta_{\text{base}} \cdot \sigma(k(E_{R\to L} - E_{L\to R})), \\ \beta_{R,t} &= \beta_{\text{base}} \cdot \sigma(k(E_{L\to R} - E_{R\to L})), \end{aligned} \tag{4}$$

where σ is the sigmoid, $\beta_{\text{base}} \in [0,1]$ a base weight, and k controls sensitivity. Thus, the view with larger deviation receives stronger correction.

Confidence-guided refinement. Each view is then fused with the warped counterpart using reliability maps $\mathbf{R}_{L\to R}, \mathbf{R}_{R\to L} \in [0,1]^{1\times H\times W}$:

$$\begin{aligned} z'_{R,t-1} &= (1 - \mathbf{B}_{R,t}) \odot z_{R,t-1} \\ &\quad + \mathbf{B}_{R,t} \odot \mathrm{Warp}(z_{L,t-1}, D_{L\to R}), \\ \mathbf{B}_{R,t} &= \beta_{R,t} \cdot \mathbf{R}_{L\to R}. \end{aligned} \tag{5}$$

and symmetrically for $z'_{L,t-1}$. The refined latents $(z'_{L,t-1}, z'_{R,t-1})$ are passed to the next diffusion step. Although both updates use the pre-refined latents as sources for warping, the asymmetric strengths ensure the more deviant view is pulled stronger toward the projection of the consistent view, while the latter changes minimally. This partial alignment per step accumulates over multiple denoising iterations to achieve global consistency.

GGAR preserves local geometry within the network, while GALA explicitly realigns global geometry across sampling steps, yielding stereo-consistent, text-guided diffusion without retraining.

4.3 Mask-Guided Content Editing.

Users sometimes wish to edit only specific regions of interest, yet the model's fine-grained control may fall short of this requirement. To address this, we extend our method following [21] to support mask input, allowing the model to edit only the masked region while keeping other areas unchanged.

5 Experiments

5.1 Experimental setups

Datasets and Baselines. We selected 120 representative stereo image pairs from multiple sources, including NBU-SIRQA [8], Flickr1024 [32], and Middlebury 2014 [26], as well as public online repositories and synthetically generated images [44]. All images were resized to 512×512 for consistency. The dataset covers diverse categories such as human subjects, animals, and natural environments. The corresponding text prompts were either automatically generated using ChatGPT or manually authored and verified by the authors.

To the best of our knowledge, as there are no publicly available stereo-consistent editing methods for direct comparison, we evaluate our proposed approach, StereoEdit, against five state-of-the-art text-guided video editing methods built upon text-to-image (T2I) or text-to-video (T2V) diffusion models. These methods maintain temporal consistency, which conceptually aligns with our goal of achieving text-guided stereo-view consistency. Specifically, we compare with TokenFlow [9], VidToMe [20], VideoGrain [43], VideoDirector [35], and ReAtCo [34]. These methods, while originally designed for temporal coherence across frames, serve as reasonable baselines to assess our model's ability to preserve consistency across stereo views.

Metrics. We evaluate the edited stereo image pairs using both automatic metrics and human assessment. For automatic evaluation, we employ three metrics. CLIP-T [25] measures how well the edits match the prompt by averaging the cosine similarity between each image and the text in CLIP embedding space. CLIP-F [25] measures the average cosine similarity between the CLIP embeddings of the left and right stereo views, capturing cross-view semantic and textural consistency. Warp-Err [38] quantifies pixel-level discrepancies between the warped image and its corresponding target view, based on the disparity map estimated from the original stereo pair. For human evaluation, 25 participants rated 50 edited stereo pairs on a scale from 1 to 10 across three dimensions: Edit Accuracy, assessing the correctness of local edits with respect to the prompt; Stereo Consistency, evaluating the perceptual coherence between left and right views; and Overall Edit Quality.

Implementation Details. The experiments were conducted on a single NVIDIA A100 GPU. We implemented our method using Stable Diffusion v1.5 as the pre-trained text-to-image diffusion model, with 50 DDIM inversion and denoising steps. StereoEdit operates in a zero-shot manner, requiring no additional parameter tuning. Segmentation is performed using the Segment Anything Model [16], and disparity estimation is conducted using MASt3R [19]. We set $\sigma = 2$ and $\lambda = 0.5$ for GGAR, and $k = \beta_{\text{base}} = 0.5$ for GALA, providing stable and effective geometry-guided refinement across all experiments. Additional experiments are provided in the supplementary materials.

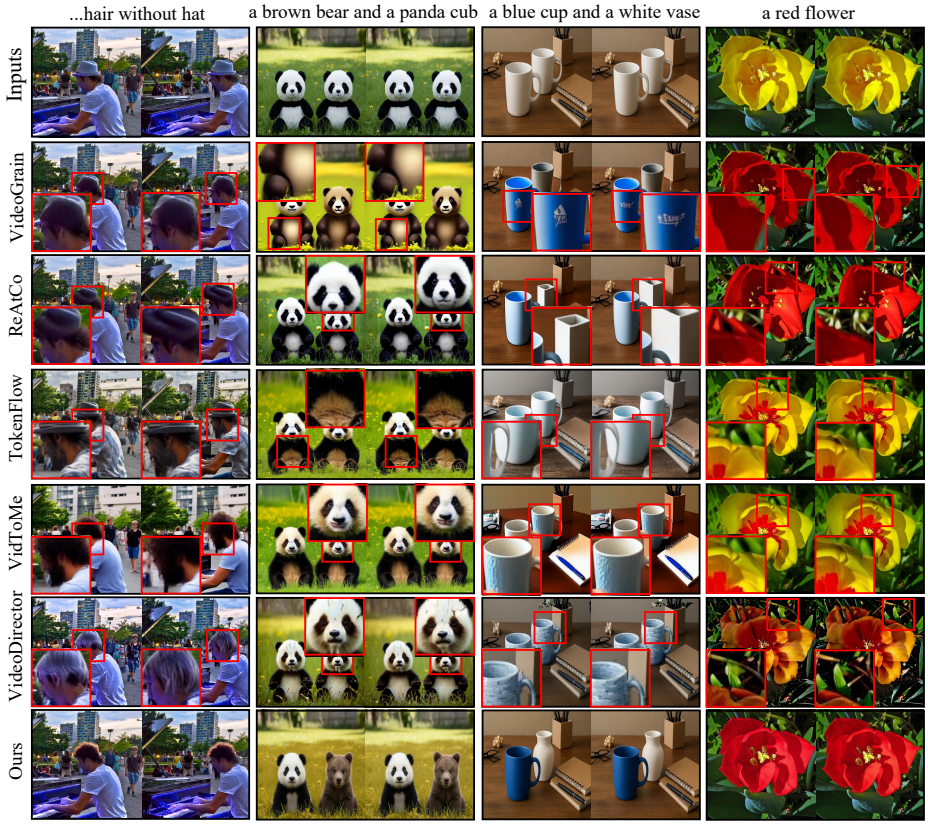

Fig. 3: Qualitative comparison of our method against state-of-the-art approaches. Our method outperforms previous methods across diverse editing tasks, demonstrating superior visual quality and stereo consistency. Regions with obvious differences are enclosed in rectangular boxes, emphasizing improvements in stereo consistency, and editing accuracy.

5.2 Qualitative and quantitative comparisons

Qualitative Evaluation. Fig. 1 shows some results of our method on diverse instances. The side-by-side stereo-view comparisons demonstrate strong consistency in both appearance and geometry across dual perspectives, a key requirement for stereo image editing. Our approach reliably supports a variety of edits, including color changes, geometric deformations, and texture modifications, while preserving multi-view coherence and producing high-quality outputs. These results highlight the robustness and versatility of our framework in preserving stereo-view coherence while enabling flexible and intuitive user-guided edits.

Fig. 3 presents a qualitative comparison with existing methods. Our approach consistently delivers more coherent and faithful edits across diverse scenarios. In

Table 1: Quantitative evaluation of various editing methods using automatic metrics.

Methods	Automatic Metric		
	CLIP-T ↑	CLIP-F ↑	Warp-Err ↓
TokenFlow	31.87	96.14	5.53
VidToMe	32.14	96.38	6.88
VideoGrain	33.77	94.62	6.51
VideoDirector	33.08	96.09	8.07
ReAtCo	33.72	95.04	7.20
StereoEdit(Ours)	**33.82**	**97.36**	**3.92**

Table 2: Quantitative evaluation of different editing methods using various human metrics.

Methods	Human Evaluation		
	Edit-Acc ↑	Stereo-Con ↑	Overall ↑
TokenFlow	6.52	7.31	6.19
VidToMe	6.30	7.64	6.64
VideoGrain	7.62	7.06	7.18
VideoDirector	7.08	6.96	6.89
ReAtCo	6.99	6.76	6.39
StereoEdit(Ours)	**7.65**	**8.01**	**8.09**

the two-cup scene, Videograin produces visible artifacts and inconsistent patterns on the blue cup between the left and right views, while ReAtCo suffers from geometric misalignment, most notably reflected in the mismatched sizes of the white vase across views. TokenFlow introduces extra geometric structures around the cup handle, and VidToMe produces view-inconsistent textures. VideoDirector further exhibits both texture inconsistency and geometric distortion. In the flower color editing task, Videograin shows shadow–color interference and noticeable inter-view inconsistency. ReAtCo, TokenFlow, and VidToMe exhibit geometric misalignment near the upper-right petal edges, while TokenFlow and VidToMe also fail to fully propagate the red color to the target flower. Additionally, VideoDirector unintentionally alters the background and introduces severe cross-view inconsistency. Conversely, our method accurately propagates edits and generates plausible, consistent textures and geometric structures across both stereo views, demonstrating superior stereo coherence and visual fidelity.

Quantitative Evaluation. As shown in Tab. 1, our method consistently achieves state-of-the-art results across all automatic metrics, demonstrating superior stereo-view coherence and visual fidelity. Specifically, StereoEdit attains the highest CLIP-T and CLIP-F scores, reflecting improved text-image alignment and vi-

sual realism, while achieving a 29.11% reduction in Warp-Err compared with the best prior method, indicating significantly enhanced cross-view consistency. These gains can be attributed to the proposed disparity-aware GGAR and GALA consistency modules, which jointly enhance edit fidelity and stereo alignment.

User Study. For human evaluation, we recruited 25 participants who are familiar with image editing and stereo vision concepts, including graduate students and researchers in computer vision as well as experienced visual practitioners. Before the study, all participants were given detailed evaluation guidelines and exemplar results to ensure a consistent understanding of the criteria. Each participant evaluated 50 randomly ordered and anonymized stereo pairs from all methods, using a 1–10 rating scale on three aspects (see details in Sec. 5.1). As shown in Tab. 2, StereoEdit achieved the best performance across all dimensions, with 7.65 in Edit Accuracy, 8.01 in Stereo Consistency, and an Overall Edit Quality score of 8.09, significantly surpassing all baselines. These results confirm that StereoEdit produces edits with highly consistent textures and geometric structures across stereo views, while maintaining superior visual realism and perceptual quality favored by users.

Additional analyses are in the supplementary materials.

5.3 Ablation study

Table 3: Quantitative ablation study. Here, "$\mathbf{W}_{\text{Attn}}$" denotes a attention alignment strategy commonly adopted in video editing [3] and "GGAR w/o R" indicates the variant of GGAR with the logarithmic modulation component removed. And "Base" refers to the Stable Diffusion model without any stereo adaptation.

Base	GGAR w/o R	GGAR	GALA	$\mathbf{W}_{\text{Attn}}$	CLIP-F ↑	CLIP-T ↑	Warp-Err ↓
✓					94.97	32.29	8.78
✓	✓				97.19	33.19	4.55
✓		✓			97.23	33.19	4.41
✓			✓		95.97	32.84	5.15
✓			✓	✓	96.31	33.25	5.04
✓		✓	✓		97.36	33.82	3.92

We conduct ablation studies to analyze the impact of individual modules in StereoEdit, as shown in Fig. 4 and Tab. 3.

Geometrically-Gated Attention Reference. As shown in Fig. 4 (see "only GGAR" and "Origin edited"), the proposed GGAR module markedly enhances local texture coherence, particularly improving the consistency of the white surface region of the egg compared with the original edited result. This qualitative improvement is consistent with the quantitative results reported in Tab. 3. To

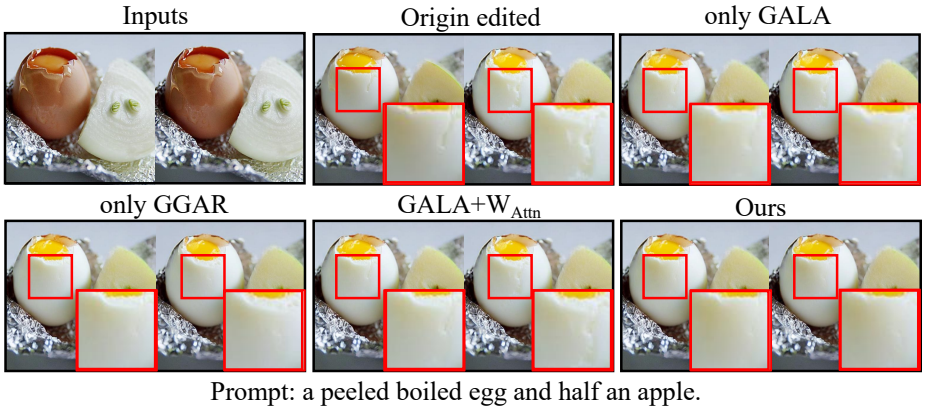

Fig. 4: Qualitative results of the ablation study on the components of our method. Regions with noticeable differences are highlighted by rectangular boxes.

further assess the stereo-specific effectiveness of GGAR, we replaced it with a disparity-agnostic attention alignment strategy commonly adopted in video editing [3]. As shown by both Fig. 4 (GALA+$\mathbf{W}_{\text{Attn}}$) and Tab. 3, this alternative yields inferior performance, exhibiting noticeable geometric distortions and textural inconsistencies across stereo views. These findings underscore the importance of explicitly modeling stereo geometry for achieving high-fidelity stereoscopic editing. Furthermore, the ablation of the logarithmic modulation component, also reported in Tab. 3, validates its additional contribution to enhancing inter-view consistency.

Geometry-Aware Latent Alignment. As shown in Fig. 4, the GALA module effectively preserves global structural consistency, as observed in the "only GALA" and "Origin edited" results. Removing the GALA module significantly degrades global geometry coherence during stereo editing, as illustrated by the "only GGAR" and "Ours" cases, which exhibit noticeable geometric inconsistencies between stereo views. These observations highlight the crucial role of GALA in maintaining coherent global geometry for stable and effective stereoscopic editing. The quantitative ablation results summarized in Tab. 3 further confirm the effectiveness of the proposed module.

5.4 Extended to stereo video editing

To further verify the scalability and effectiveness of our approach, we extend the proposed module to diffusion-based T2V frameworks for stereo video editing by inserting GGAR/GALA into a video diffusion editor and stacking the left/right images as a joint input. As shown in Fig. 5, our method demonstrates strong generalization and scalability, enabling consistent editing across stereo video frames. By leveraging the temporal coherence inherent in video diffusion models and enforcing stereo geometric consistency through our module, the approach

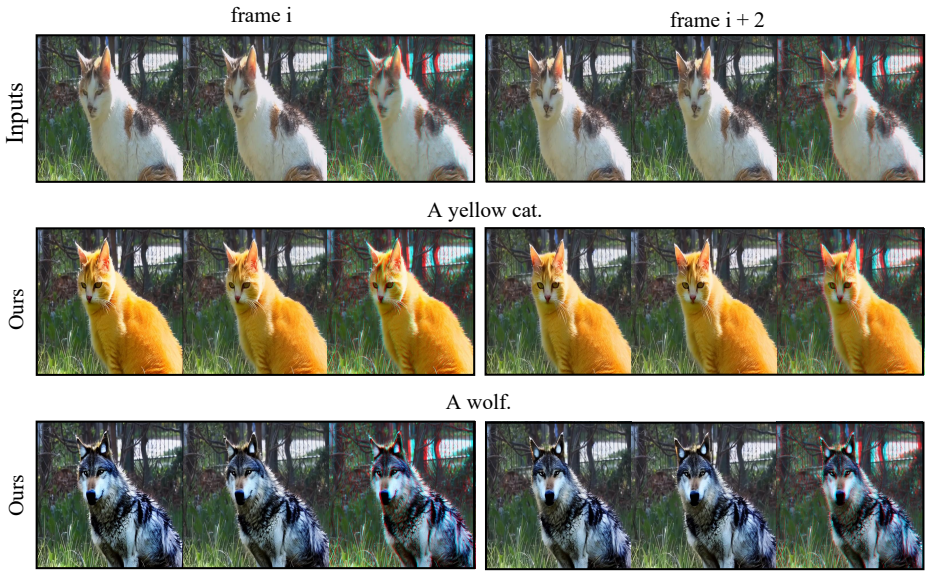

Fig. 5: StereoEdit can also be effectively applied to stereo video editing tasks, achieving promising results. Specifically, frame i denotes the i-th frame of the video, while frame i + 2 represents the frame two time steps later.

achieves accurate and stable results across both spatial and temporal dimensions. This extension broadens the applicability of our method and establishes a solid foundation for future stereo-aware video editing tasks.

6 Conclusion

We present StereoEdit, a training-free diffusion framework for consistent stereo image editing. By leveraging pre-trained text-to-image diffusion models without any fine-tuning, StereoEdit achieves structurally coherent and semantically faithful edits through two dedicated modules designed specifically with disparity geometry in mind. These modules jointly address cross-view inconsistency and geometric distortion, which are longstanding challenges in stereoscopic editing. Extensive experiments on diverse stereo benchmarks demonstrate that StereoEdit significantly improves geometric stability, cross-view coherence, and visual realism. The edited stereo image pairs can be directly used for immersive stereo displays, providing a coherent and controllable 3D visual experience. Furthermore, StereoEdit can be naturally extended to stereo video editing, maintaining stereo consistency across frames. Thus, our method offers a practical and effective solution for stereo content editing.

References

1. Al-Ansi, A.M., Jaboob, M., Garad, A., Al-Ansi, A.: Analyzing augmented reality (ar) and virtual reality (vr) recent development in education. Social Sciences & Humanities Open **8**(1), 100532 (2023)
2. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology **35**(7), 6832–6852 (2025)
3. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
4. Chai, X., Shao, F., Jiang, Q., Ho, Y.S.: Roundness-preserving warping for aesthetic enhancement-based stereoscopic image editing. IEEE Transactions on Circuits and Systems for Video Technology **31**(4), 1463–1477 (2020)
5. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stereoscopic neural style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6654–6663 (2018)
6. Chen, Y., Chen, Z., Zhang, C., Wang, F., Yang, X., Wang, Y., Cai, Z., Yang, L., Liu, H., Lin, G.: Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21476–21485 (2024)
7. Fu, Y., Liu, S., Kulkarni, A., Kautz, J., Efros, A.A., Wang, X.: Colmap-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20796–20805 (2024)
8. Fu, Z., Shao, F., Jiang, Q., Meng, X., Ho, Y.S.: Subjective and objective quality assessment for stereoscopic image retargeting. IEEE Transactions on Multimedia **23**, 2100–2113 (2020)
9. Geyer, M., Bar Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: International Conference on Learning Representations. vol. 2024, pp. 1608–1620 (2024)
10. Guo, H., Liu, S., Zhu, S., Zeng, B.: Joint bundled camera paths for stereoscopic video stabilization. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 1071–1075. IEEE (2016)
11. Haque, A., Tancik, M., Efros, A.A., Holynski, A., Kanazawa, A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19740–19750 (2023)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Imani, H., Islam, M.B., Wong, L.K.: Saliency-aware stereoscopic video retargeting. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 1230–1239 (2023)
14. Jiang, H., Lou, Z., Ding, L., Xu, R., Tan, M., Jiang, W., Huang, R.: Defom-stereo: Depth foundation model based stereo matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21857–21867 (2025)
15. Kim, M.J., Kim, D., Yun, S., Choo, J.: Tv-live: Training-free, text-guided video editing via layer informed vitality exploitation. arXiv preprint arXiv:2506.07205 (2025)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)

17. Kučera, M., Mould, D., Sỳkora, D.: Stylebin: Stylizing video by example in stereo. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022)
18. Lee, D.I., Park, H., Seo, J., Park, E., Park, H., Baek, H.D., Shin, S., Kim, S., Kim, S.: Editsplat: Multi-view fusion and attention-guided optimization for view-consistent 3d scene editing with 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11135–11145 (2025)
19. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024)
20. Li, X., Ma, C., Yang, X., Yang, M.H.: Vidtome: Video token merging for zero-shot video editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7486–7495 (2024)
21. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022)
22. Padmavathi, A., Skandha, V., et al.: Multi-step instruction-guided editing for neural radiance fields: Enhancing user control in 3d scene manipulation. In: 2024 First International Conference on Data, Computation and Communication (ICDCC). pp. 730–736. IEEE (2024)
23. Qu, Y., Chen, D., Li, X., Li, X., Zhang, S., Cao, L., Ji, R.: Drag your gaussian: Effective drag-based editing with score distillation for 3d gaussian splatting. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
25. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
26. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German conference on pattern recognition. pp. 31–42. Springer (2014)
27. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
28. Sun, C., Liu, Y., Han, J., Gould, S.: Nerfeditor: Differentiable style decomposition for 3d scene editing. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 7306–7315 (2024)
29. Van Der Hooft, J., Amirpour, H., Vega, M.T., Sanchez, Y., Schatz, R., Schierl, T., Timmerer, C.: A tutorial on immersive video delivery: From omnidirectional video to holography. IEEE Communications Surveys & Tutorials **25**(2), 1336–1375 (2023)
30. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
31. Wang, L., Jin, H., Yang, R., Gong, M.: Stereoscopic inpainting: Joint color and depth completion from stereo images. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)

32. Wang, Y., Wang, L., Yang, J., An, W., Guo, Y.: Flickr1024: A large-scale dataset for stereo image super-resolution. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019)
33. Wang, Y., Ying, X., Wang, L., Yang, J., An, W., Guo, Y.: Symmetric parallax attention for stereo image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 766–775 (2021)
34. Wang, Y., Li, Y., Liu, M., Zhang, X., Liu, X., Cui, Z., Chan, A.B.: Re-attentional controllable video diffusion editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8123–8131 (2025)
35. Wang, Y., Wang, L., Ma, Z., Hu, Q., Xu, K., Guo, Y.: Videodirector: Precise video editing via text-to-video models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2589–2598 (2025)
36. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5249–5260 (2025)
37. Wen, M., Wu, S., Wang, K., Liang, D.: Intergsedit: Interactive 3d gaussian splatting editing with 3d geometry-consistent attention prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26136–26145 (2025)
38. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)
39. Wu, Z., Sun, C., Xuan, H., Yan, Y.: Deep stereo video inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5702 (2023)
40. Xiong, J., Hsiang, E.L., He, Z., Zhan, T., Wu, S.T.: Augmented reality and virtual reality displays: emerging technologies and future perspectives. Light: Science & Applications **10**(1), 216 (2021)
41. Yan, T., He, S., Lau, R.W., Xu, Y.: Consistent stereo image editing. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 677–680 (2013)
42. Yan, Z., Li, C., Lee, G.H.: Nerf-ds: Neural radiance fields for dynamic specular objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8285–8295 (2023)
43. Yang, X., Zhu, L., Fan, H., Yang, Y.: Videograin: Modulating space-time attention for multi-grained video editing. In: The Thirteenth International Conference on Learning Representations (2025)
44. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–18 (2025). https://doi.org/10.1109/TPAMI.2025.3613256
45. Yuan, Z., Luo, J., Shen, F., Li, Z., Liu, C., Mao, T., Wang, Z.: Dvp-mvs: Synergize depth-edge and visibility prior for multi-view stereo. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9743–9752 (2025)
46. Zhao, S., Hu, W., Cun, X., Zhang, Y., Li, X., Kong, Z., Gao, X., Niu, M., Shan, Y.: Stereocrafter: Diffusion-based generation of long and high-fidelity stereoscopic 3d from monocular videos. arXiv preprint arXiv:2409.07447 (2024)