NeoMap: Training-free Novel-View Synthesis from Single Images and Videos

Jinxi Li[1,2†], Tianyi Zhang[1,2†], Yafei Yang[1,2], Zihui Zhang[1,2], Peng Huang[1,2], Koon Wing Macgyver Lin[‡], and Bo Yang[1,2‡✉]

[1] Shenzhen Research Institute, The Hong Kong Polytechnic University
[2] vLAR Group, The Hong Kong Polytechnic University
[†] equal contribution [‡] joint supervision [✉] corresponding author
{jinxi.li,tonax.zhang}@connect.polyu.hk, bo.yang@polyu.edu.hk

Abstract. We study the challenging problem of novel view video synthesis from single images or monocular videos. Existing methods, which operate under the assumption that pre-trained video models lack native novel view synthesis capability and enforce view alignment via camera conditioning, task-specific fine-tuning, or stepwise hard denoising guidance, often suffer from artifacts and compromised global scene consistency. In this paper, we introduce **NeoMap**, a novel training-free framework designed to locate high-fidelity, view-consistent novel view solutions from general pre-trained video models. The key to our approach is the core insight that promising novel view solutions are inherently encoded within the natural video data manifold learned by pre-trained models, and the core challenge is simply to locate this optimal solution. We solve this via our core mechanism: convergent manifold alternating projection iterations that optimize the initial noise. Extensive experiments demonstrate that NeoMap significantly outperforms all existing methods across 3 standard novel view synthesis benchmarks, including the challenging Tanks-and-Temples, LLFF and DAVIS datasets, achieving state-of-the-art generation fidelity and top-tier view consistency. Our code and data are available at https://github.com/vLAR-group/NeoMap

Keywords: Novel-view synthesis · Video generation · Training-free

1 Introduction

Monocular novel view synthesis (NVS), which generates photorealistic unseen views from a single-view image or unconstrained monocular video, is a core task in computer vision with wide applications in AR/VR, 3D/4D content creation, and robotic perception. However, it remains an extremely challenging problem for traditional reconstruction-based methods [27, 39] due to inherent severe occlusions, incomplete scene information, and depth ambiguity from monocular inputs. Recently, the rapid development of powerful large-scale video generation models [4, 11, 31, 52, 61], which show a reasonable understanding of spatial-temporal relationships, has opened a promising new direction to address this long-standing challenge.

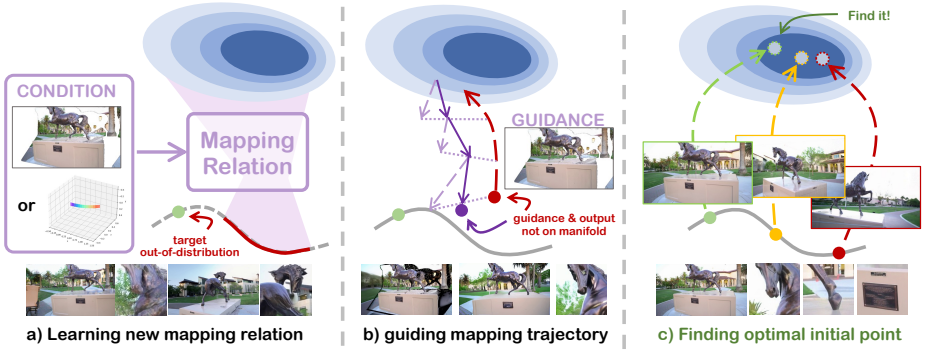

Fig. 1: Different NVS paradigms: a) Learning new mapping relation would shrink the output data manifold, leading to overfitting and OOD problem; b) Guiding generation trajectory with bad condition would lead the generation output away from data manifold; c) Different related output possibilities already exist in the data manifold.

One line of work learns a new conditional video generation mapping by training models on 3D/4D data with camera pose annotations. Either injecting camera trajectories [2,3,56,58] or taking depth-based warped view images [12,58,64] as additional generation conditions, these methods essentially constrain the denoising process with pose-related signals. Although these methods show promising performance, they inevitably suffer from overfitting to limited camera trajectory distributions and depth estimation noise seen during training, leading to poor generalization to in-the-wild scenes and unseen camera motions. Another line of work avoids task-specific retraining via training-free guidance strategies [49, 62]. They use depth-based warped images as reliable priors, and blend the latent features of priors with the generating latent to guide the generation denoising step-by-step. However, such hard step-wise constraints often disrupt the continuity of the native generation process of the pre-trained model, leading to severe artifacts including incomplete content, corrupted scene structure, and incorrect semantic understanding.

Despite their differences, as shown in Figure 1, both categories share a fundamental assumption: in order to obtain a valid and geometrically consistent NVS solution, *the mapping relation between the noise space and the video latent space should be modified or redefined in a task-specific way.* In this work, we propose to challenge this assumption with a key insight: a high-quality NVS solution for a given input and target pose should inherently exist in the native output latent data manifold of a pre-trained video generation model, which has been trained on massive diverse real-world videos, whose output data manifold thus already covers rich 3D scenes and natural camera motions. Therefore, **instead of focusing on the mapping relation by altering the denoising trajectory or re-training the model, we propose to solve this problem simply by locating this valid target results in the output space**.

Given a video generation model, it is extremely challenging to find a specific video in its intractable abstract output data manifold. Nevertheless, recent video diffusion methods can all be regarded as a deterministic mapping (ODE for flow-

matching methods [34, 35] or PF-ODE for diffusion methods [36]) from regular gaussian noise space to video data manifold, which means every video in the output space can be defined by a point in the random noise space. Thus, our challenge of locating the valid video data point in the data manifold can be simply framed as searching for the optimal initial noise latent corresponding to the target valid result.

With this motivation, we propose an elegant pipeline to discover the exact initial noise latent that naturally decodes into a high-fidelity, geometrically consistent novel view video. Our framework, **NeoMap** (**N**ois**E** initializati**O**n by **M**anifold **A**lternating **P**rojection), constructs this optimal initialization through three intuitive modules: 1) a **novel view prior module** establishes base geometric constraints via depth-based image warping, which inherently suffers from out-of-distribution artifacts and semantic emptiness in occlusions; 2) an **anchored manifold projection (AMP) module** addresses this by pushing the state toward the video data manifold to hallucinate plausible content in unobserved regions while adaptively anchoring reliable visible features; 3) a **pixel-constrained projection (PCP) module** rectifies the spatial bleeding caused by VAE latent blending by enforcing strict, immutable geometric boundaries back in the pixel space.

The core of our pipeline lies in the alternating projection between the AMP and PCP modules. By iteratively mapping the state between the unconstrained video data manifold and the strict NVS geometric constraints, **NeoMap** mathematically guaranties that our initial noise converges to the optimal intersection of global semantic realism and precise 3D view alignment. Coupled with an auxiliary trajectory re-anchoring strategy to prevent integration drift during generation, this enables us to synthesize photorealistic videos without any task-specific finetuning or additional guidance. Our main contributions are:

- We introduce a new training-free framework to solve the monocular novel view synthesis problem by finding a optimal initial noise, eliminating the need for any generation guidance or task-specific finetuning.
- We propose to initialize the noise via manifold alternating projection, simultaneously achieving high fidelity and geometric consistency by ensuring that the generation output is located on the NVS constrained data manifold.
- We demonstrate a significant improvement in visual quality and geometric consistency on three existing datasets for both static single-image and dynamic monocular video novel view synthesis.

2 Related Work

Novel View Synthesis by Camera-Controlled Video Generation. Driven by the rapid advancement of foundational video generation models [4, 11, 31, 52, 61], extensive research has explored conditional video generation to achieve precise camera control. A prominent line of work treats this as a conditional generation task. Early approaches in this domain [1, 17, 18, 32, 56, 59] primarily focused on incorporating camera extrinsics directly into the generation process

by fine-tuning diffusion models with specialized motion modules or parameter encoders to learn specific trajectory patterns. Building upon this, recent advancements introduce more explicit conditioning mechanisms to enhance both geometric alignment and visual quality. For instance, TrajectoryAttention [58] injects fine-grained motion control by aligning content along specified trajectories directly within the attention layers. TrajectoryCrafter [63] disentangles deterministic view transformations from stochastic generation by utilizing a dual-stream architecture conditioned on point cloud renders, while ReCamMaster [3] frames the problem as video-conditioned generative re-rendering to reproduce dynamic scenes under novel trajectories.

Novel View Synthesis by Warping and Inpainting. A parallel paradigm tackles NVS through a warping-and-inpainting formulation. This pipeline stems from the success of diffusion-based inpainting models [29, 37, 67], which demonstrate a strong ability to plausibly fill semantic blanks in unobserved regions. Methods like GenWarp [47] and VistaDream [53] synthesize novel views by first warping the source image using depth priors, followed by utilizing an inpainting network based on image generation to hallucinate the missing pixels caused by occlusion. Extending this to dynamic 4D scenes, recent video-based methods such as ViewCrafter [64] and FlexWorld [12] lift the source video into a 3D point cloud, render it from the target trajectory, and employ a video generation model to temporally inpaint the resulting visual holes. However, a major bottleneck of these approaches is the absolute requirement to learn or heavily fine-tune a task-specific video inpainting model to handle the severe warping artifacts and domain gaps. To alleviate this heavy training burden, recent works [21, 22, 33] like NVS-Solver [62] have explored training-free zero-shot novel view synthesis by adaptively modulating the score function of pre-trained models. In our pipeline, we align with this training-free warping and inpainting philosophy but strictly diverge by finding a good initialization rather than controlling the generation mapping relation.

Generation Control via Noise Handling. Controlling the generative process through explicit noise manipulation has emerged as a powerful paradigm, originating from image editing tasks where inverting the denoising trajectory successfully extracts the structural composition of a source image [19, 24, 40, 43, 50, 57]. Extending these principles to video generation introduces the critical challenge of temporal coherency [7,8,13,15,26,60]. More recently, explicit manipulation of the initial noise or intermediate latent space has proven highly effective for dictating motion dynamics without retraining. Techniques in this domain range from applying pre-specified translation vectors to inverted noise [28] to constructing correlated noise distributions [14] and aligning latent features [5, 41]. Pushing this concept further, noise warping [6, 9] demonstrates that spatially warping the initial noise map according to optical flow inherently establishes robust temporal consistency for zero-shot video editing. While these methods firmly establish that structural layout and motion can be dictated by carefully crafted initial noise, they have no constraints on the unseen regions and largely rely on the quality of the flow for the observed regions, which is quite low for an

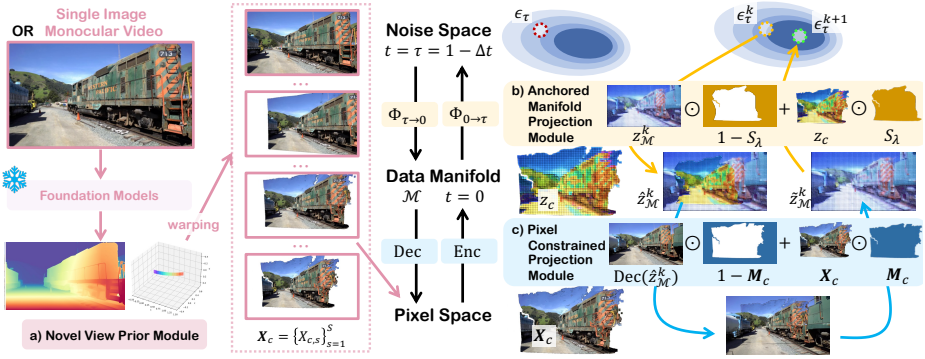

Fig. 2: An illustration of the three modules in our framework.

NVS problem due to the noise of the estimated depth and the incompleteness of the object. Our framework elegantly overcomes these limitations by propagating valid semantic priors to meaningfully constrain and hallucinate the unseen regions and adaptively trusting only the reliable visible regions.

3 NeoMap

We introduce a novel view synthesis (NVS) framework built on pre-trained flow-based video generation models, with a core manifold alternative projection scheme that unifies global semantic consistency from the video data manifold and pixel-level geometric constraints from the 3D warping prior, as shown in Figure 2. This section first covers the preliminaries of flow-based video generation (Section 3.1), then details our pipeline components: novel view prior construction (Section 3.2), anchored manifold projection (Section 3.3), pixel-constrained projection (Section 3.4), and auxiliary trajectory re-anchoring (Section 3.5).

3.1 Preliminaries on Flow-Based Video Generation

Our framework is compatible with deterministic flow-based video generation models, including Flow Matching [34, 35] and Probability Flow ODE (PF-ODE) [36]. We formalize core components with compact notation below.

A pre-trained video Variational Autoencoder (VAE) [46] maps an input video $\boldsymbol{X} \in \mathbb{R}^{F \times H \times W \times 3}$ to a low-dimensional latent space $\mathcal{Z} \in \mathbb{R}^{F_\mathcal{Z} \times H_\mathcal{Z} \times W_\mathcal{Z} \times d_\mathcal{Z}}$ via the encoder $\text{Enc}(\cdot)$, and reconstructs it via the decoder $\text{Dec}(\cdot)$. All semantically valid natural videos are mapped to a smooth low-dimensional data manifold $\mathcal{M} \in \mathcal{Z}$, which forms the support of the real video distribution $p_0(z)$. Latent codes on $\mathcal{M}$ are decoded to plausible, temporally consistent videos, while out-of-distribution (OOD) codes lie outside $\mathcal{M}$ and produce invalid content.

Flow-based video generation models learn a continuous vector field $v_t(z)$ governing latent code evolution over $t \in [0, 1]$ from the standard Gaussian prior

noise space $p_1(z) = \mathcal{N}(0, I)$ at $t = 1$ to the data space $z \in \mathcal{M}$ at $t = 0$, following the Ordinary Differential Equation (ODE):

$$\frac{dz}{dt} = v(z, t). \tag{1}$$

For simplicity, we denote the flow mapping $\Phi_{t\to s}(z)$ as the ODE solution mapping z_t from time t to s, with two core properties: **1) Forward (noising) flow**: $\Phi_{0\to t}(z_0)$ maps a data latent $z_o \in \mathcal{M}$ to a noised code z_t as time t; **2) Reverse (denoising) flow**: $\Phi_{t\to 0}(z_t)$ maps a noised code z_t back to the data manifold at $t = 0$. For a well-trained model, the reverse flow maps any Gaussian sample at $t = 1$ to a valid video latent. Our framework is agnostic to the vector field formulation, as long as the above flow properties hold.

3.2 Novel View Prior Module

Our goal is to generate a multi-frame novel view video from a single monocular reference video $\boldsymbol{X} = \{X_f\}_{f=1}^{F} \in \mathbb{R}^{F\times H\times W\times 3}$ (or a single image with $F = 1$). We use depth and pose estimation foundation models such as VGGT [54] and VIPE [23] to jointly recover per-frame depth maps $\boldsymbol{D} = \{D_f\}_{f=1}^{F}$ and the corresponding source camera trajectory $\mathcal{C}_{src} = \{\mathcal{C}_f = (K_f, [\boldsymbol{R}_f \mid \boldsymbol{T}_f] \in \mathrm{SE}(3))\}_{f=1}^{F}$. Then we can define the given user-specified target camera trajectory in the same world coordinate as $\mathcal{C} = \{c_s = (K_s, [\boldsymbol{R}_s \mid \boldsymbol{T}_s])\}_{s=1}^{S}$.

Following existing works [62], we use a warping operator $\mathcal{W}$ to project the source video content, along with the corresponding depth and source poses, to each target view as follows:

$$X_{c,s}, M_s = \mathcal{W}(X_s, D_s, \mathcal{C}_s, c_s), \tag{2}$$

and similarly for single-image input. This operation outputs a partial warped target frame $X_{c,s}$, along with a binary visibility mask M_s. Stacking the outputs across all target poses yields the full warped video prior $\boldsymbol{X}_c = \{X_{c,s}\}_{s=1}^{S}$ and its aligned visibility mask $\boldsymbol{M}_c = \{M_s\}_{s=1}^{S}$.

We map the warped prior $\boldsymbol{X}_c$ to the latent space via pre-trained VAE encoder to get the latent prior $z_c = \mathrm{Enc}(\boldsymbol{X}_c)$. In an ideal scenario where z_c is a clean and valid latent lying strictly on the natural video data manifold $\mathcal{M}$, one could simply obtain an optimal initial noise at $t = 1$ via the forward flow $z_1 = \Phi_{0\to 1}(z_c)$. However, our constructed latent prior z_c is inherently flawed: 1) visible regions of z_c contain out-of-distribution artifacts from depth noise, occlusions, and view-dependent appearance changes; 2) unobserved regions are uninformative and unconstrained, *i.e.*, black or empty in invisible pixel space, providing no valid semantic or geometric cues to align generated content with the source scene.

Consequently, directly utilizing this initial noise z_1 yields corrupted generation trajectories and is therefore insufficient. Instead, we define a near-noise starting state $z_\tau = \Phi_{0\to\tau}(z_c)$ at an early timestep $t = \tau = 1 - \Delta t$ via the forward flow, and then iteratively refine it to find an optimal initial noise that balances both global plausibility and precise view alignment.

3.3 Anchored Manifold Projection Module

To address this, we propose the **Anchored Manifold Projection (AMP)** module as the first component of our alternating projection framework. Intuitively, AMP serves two critical functions: 1) it projects plausible geometric and semantic information into the invisible unconstrained regions, thus enabling plausible completion; 2) it searches for a latent state with higher quality on the manifold $\mathcal{M}$ around the flowed state z_c, fixing the artifacts in the visible regions while being anchored by the global constraints. We achieve these two goals through the following procedures.

Given the current initial noise prior z_τ from Section 3.2, we denote it as our starting initial state z_τ^0 and iteratively update it to optimal points $z_\tau^* = z_\tau^K$ by K iterations. For each iteration k, we first use reverse flow to project the initial state z_τ^k toward the data manifold as $z_\mathcal{M}^k$. To bypass the computational bottleneck of full-step ODE integration, we employ a one-step large-stride Euler approximation of the generative reverse flow as follows:

$$z_\mathcal{M}^k = \Phi_{\tau\to 0}(z_\tau^k) \approx z_\tau^k - \tau v_\theta(z_\tau^k, \tau), \tag{3}$$

where $v_\theta(x,t)$ is the flow field function of the pre-trained video generation model. By passing the state through the pre-trained vector field v_θ, the model naturally leverages its learned spatial-temporal prior to broadcast the information from the visible region to the invisible region, pushing $z_\mathcal{M}^k$ one-step closer to $\mathcal{M}$.

However, since spatial-temporal attention in vector field v_θ also broadcasts noisy information from unobserved regions to the observed region, some of which is good and some of which is harmful, this unconstrained forward flow inevitably alters the highly confident visible regions. To strictly enforce the view prior, we need to anchor the estimated state $z_\mathcal{M}^k$ with the given prior constraint z_c. Intuitively, if one latent in $z_\mathcal{M}^k$ is similar to one in z_c, it means that the generation model v_θ tends not to fix it, which implies that it is a stable high-quality prior. Thus, we choose to believe this given prior. By contrast, if the latent is very different from z_c, it means the generation model tends to fix it significantly, which implies this latent tends to contain artifacts. Therefore, based on this intuition, we perform a latent blend operation to yield a harmonized state:

$$\hat{z}_\mathcal{M}^k = S_\lambda \odot z_c + (1 - S_\lambda) \odot z_\mathcal{M}^k, \tag{4}$$

where S_λ is the feature-level similarity mask with threshold λ, defined as

$$S_\lambda = \mathbb{1}(\mathrm{CosSim}(z_\mathcal{M}^k, z_c) > \lambda) \in \{0,1\}^{F_\mathcal{Z}\times H_\mathcal{Z}\times W_\mathcal{Z}}. \tag{5}$$

This dynamic latent blending preserves the model's high-quality hallucinations in artifact regions while forcefully overwriting the visible geometry back to the NVS prior z_c. Finally, we map this harmonized latent back to the near-noise timestamp τ again via the direct forward process:

$$z_\tau^{k+1} = \Phi_{0\to\tau}(\hat{z}_\mathcal{M}^k) = (1-\tau)\cdot \hat{z}_\mathcal{M}^k + \tau \cdot \epsilon, \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is a random standard Gaussian noise.

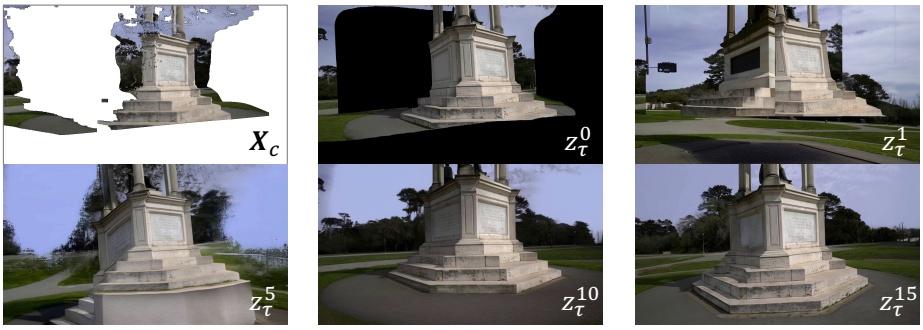

Fig. 3: Qualitative examples show that our alternative projections iteratively refine the noise quality, leading the generated results to higher-fidelity, better view consistency, and more continuous semantics.

3.4 Pixel-Constrained Projection Module

While the AMP module tries to project the state $\hat{z}^k_{\mathcal{M}}$ toward the data manifold, the similarity mask in the early iterations could be too sparse due to excessively strong noise from unconstrained parts, causing only a small fraction of the latent codes to be successfully anchored. Furthermore, even when latents are successfully anchored, the highly compressed and imperfectly disentangled nature of the VAE latent space $\mathcal{Z}$ means that latent-level blending cannot fully substitute for image-space blending. Latent blending inevitably induces spatial bleeding and global structural shifts, pushing the latent code slightly off the constraints. To solve these problems, we introduce the **Pixel-Constrained Projection (PCP)** module as the second component of our alternating projection framework.

After getting $\hat{z}^k_{\mathcal{M}}$ from Equation 4, instead of directly mapping it back to noise space using Equation 6, we decode it to the interpretable pixel space to blend the prior $\boldsymbol{X}_C$ according to our precise binary visibility mask $\boldsymbol{M}_c$ and then encode it back to the latent space via the pre-trained VAE. Mathematically, it is defined as

$$\tilde{z}^k_{\mathcal{M}} = \mathrm{Enc}(\boldsymbol{M}_c \odot \boldsymbol{X}_c + (1 - \boldsymbol{M}_c) \odot \mathrm{Dec}(\hat{\mathrm{z}}^{\mathrm{k}}_{\mathcal{M}})), \tag{7}$$

ensuring that every observable pixel from the source view remains geometrically immutable. Finally, we map $\tilde{z}^k_{\mathcal{M}}$ back to the noise space using Equation 6.

Alternating Projection Formulation: Our proposed pipeline can be theoretically formulated as Alternating Projections onto Sets. A single AMP pass pulls the state toward the natural video manifold $\mathcal{M}$, but its aggressive approximation relaxes spatial constraints; whereas a single PCP pass forcefully projects the state back onto the strict geometric constraint subspace $\mathcal{V}$, but introduces VAE encoding artifacts that push it away from $\mathcal{M}$. Since neither projection independently satisfies both conditions, as shown in Figure 3, using the manifold alternating projection iteratively between AMP and PCP as

$$\ldots \rightarrow z^k_\tau \xrightarrow{\mathrm{AMP}_{\mathrm{stage\ 1}}} \hat{z}^k_{\mathcal{M}} \xrightarrow{\mathrm{PCP}} \tilde{z}^k_{\mathcal{M}} \xrightarrow{\mathrm{AMP}_{\mathrm{stage\ 2}}} z^{k+1}_\tau \rightarrow \ldots \tag{8}$$

monotonically leads the optimal latent state towards $\mathcal{M} \cap \mathcal{V}$. The convergence proof is provided in Appendix A.

3.5 Auxiliary Trajectory Re-anchoring

While the alternating projection pipeline successfully yields an optimized initial noise z_τ^*, the usage of large-stride linear Euler approximation inevitably introduces integration errors. Because the learned vector field v_θ is highly non-linear, deploying a standard ODE solver initialized directly from z_τ^* may cause the generation trajectory to gradually drift away from the target geometric constraints.

Fortunately, extensive literature [10, 25, 48] on diffusion and flow-based generative models demonstrates that the global structure, semantic layout, and low-frequency content of a generated sample are predominantly determined during the earliest generation steps. Consequently, we only need to explicitly constrain the trajectory during these initial iterations to guarantee global view alignment. As discussed in Section 3.3, an AMP module ensuring the intermediate manifold prediction is dynamically anchored to the reliable geometric priors. Therefore, at a new denoise timestamp $t = \tau' < \tau$, we can simply using the AMP module on $z_{\tau'}$ by treating the current intermediate latent $z_{\tau'}$ as a new initialization point to re-anchor the denoising trajectory to our NVS constraints.

4 Experiments

Implementation details: we implement our method based on the recent video generation model Wan2.2-I2V-A14B [52]. The whole denoising steps are set to be 15, and we use time shift 3.5 to control the generation pace. All other configurations are kept as the original setting and its vanilla UNIPC solver [66] is used to perform denoising. All our hyperparameters are discussed in ablation study in Section 4.3 and more implementation details are in Appendix B.

Datasets: We evaluate our method for static single-image novel view synthesis on two public datasets: 1) **Tanks-and-Temples dataset** (TNT) [30]: This dataset contains both indoor and outdoor scenes undergoing large global view changes. Following [12], we randomly sampled valid video clips from 14 scenes of its test sets, in which we only pick out 49 valid scenes with reasonable camera trajectories for evaluation, while ignoring those scenes with totally unconstrained views, *i.e.*, there exist frames totally empty after warping the first frame. Notably, since there is no official camera pose given, we use VGGT [54] with calibration by COLMAP [42] to recover camera trajectories. 2) **LLFF dataset** [38]: this dataset contains 8 high-quality scenes, undergoing fast forward-facing camera motions. The high-frequency periodic appearance patterns and high-speed camera motions make it extremely challenging for NVS task.

We evaluate our method for monocular video novel view synthesis on the 3) **DAVIS dataset** [44]: this dataset contains 89 monocular videos, each featuring motions varying from human behaviors to natural environments. We use VIPE [23] to reconstruct the camera trajectory for each scene and compose it with a

Table 1: Quantitative results of all methods for visual quality and camera accuracy on Tanks-and-Temples and LLFF datasets.

	Tanks and Temples								
	Visual Quality						Camera Accuracy		
	PSNR↑	SSIM ↑	LPIPS↓	FID-192↓	FID-2048↓	CLIP-S↑	ATE↓	RPE-T↓	RPE-R↓
TrajectoryAttention [58]	13.064	0.577	0.522	14.377	151.418	0.899	0.061	0.007	0.010
ReCamMaster [3]	11.549	0.533	0.594	9.542	139.487	0.903	0.099	0.004	0.006
ViewCrafter [64]	13.675	0.590	0.564	13.748	133.219	0.902	0.010	0.004	0.003
FlexWorld [12]	14.655	0.601	0.531	4.941	92.686	0.918	0.005	0.002	0.002
NVS-Solver [62]	12.356	0.547	0.548	6.376	131.179	0.892	0.079	0.006	0.012
LanPaint [67]	13.299	0.544	0.575	13.254	196.341	0.834	0.006	0.002	0.001
Reversed Noise [35]	10.715	0.435	0.520	17.275	160.994	0.902	0.029	0.003	0.005
Flow Warped Noise [9]	11.353	0.510	0.552	6.671	130.484	0.909	0.080	0.007	0.013
NeoMap (Wan2.1)	14.876	0.591	0.481	4.735	89.527	0.922	0.010	0.002	0.004
NeoMap (Ours)	15.250	0.596	0.486	3.712	83.341	0.935	0.008	0.002	0.003
	LLFF								
	Visual Quality						Camera Accuracy		
	PSNR↑	SSIM ↑	LPIPS↓	FID-192↓	FID-2048↓	CLIP-S↑	ATE↓	RPE-T↓	RPE-R↓
TrajectoryAttention [58]	12.557	0.356	0.498	19.930	123.126	0.923	1.644	0.481	0.045
ReCamMaster [3]	10.928	0.312	0.561	10.994	173.737	0.885	1.300	0.500	0.042
ViewCrafter [64]	11.573	0.344	0.594	22.091	135.944	0.907	0.864	0.470	0.035
FlexWorld [12]	13.119	0.378	0.538	12.549	97.500	0.930	0.618	0.470	0.031
NVS-Solver [62]	11.418	0.304	0.531	6.739	150.368	0.895	1.925	0.483	0.054
LanPaint [67]	11.946	0.334	0.506	13.021	196.799	0.862	0.543	0.449	0.013
Reversed Noise [35]	10.866	0.287	0.505	7.221	148.187	0.923	1.142	0.489	0.040
Flow Warped Noise [9]	10.678	0.261	0.538	19.483	130.105	0.941	2.011	0.501	0.043
NeoMap (Ours)	13.385	0.360	0.432	3.586	80.888	0.953	0.551	0.476	0.031

manually designed relative trajectory (following the protocol of ReCamMaster [3]) as our target trajectory. More datasets details are provided in Appendix D.

Baselines: We compare with 3 groups of baselines: 1) camera-controlled video generation method TrajectoryAttention [58] and ReCamMaster [3]; 2) methods under warping and inpainting scheme, including task-specific-trained methods ViewCrafter [64] and FlexWorld [12], and training-free methods NVS-Solver [62]. In order to show the gap between NVS and traditional inpainting problem, we also include powerful training-free inpainting method LanPaint [67]; 3) noise handling methods, including the direct reversed noise via the rectified-flow forward process [35] (generating with z_τ directly) and flow warped noise following Go-With-The-Flow [6]. For all the baselines, we unify the same depth and camera priors and align the scale as required by different models. For training-free methods LanPaint, reversed noise, and flow warped noise, we adapt them to the same generation backbone Wan2.2-I2V-A14B [52] as ours. More implementation details about baselines are shown in Appendix E.

Metrics: Following [3, 12, 62], we evaluate NVS in both visual quality and camera accuracy. 1) For visual quality, we compare fidelity and coherence with text, quantified with PSNR, SSIM [55], LPIPS [65], Frechet Image Distance [20] (FID), Frechet Video Distance [51] (FVD), and CLIP Similarity (CLIP-S),

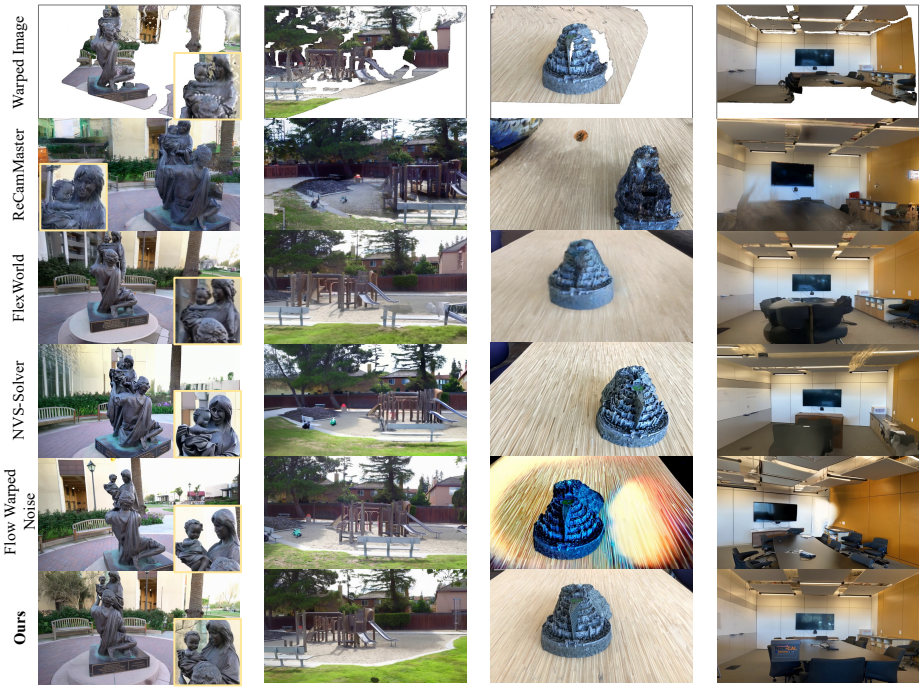

Fig. 4: Qualitative results of top 4 baselines and our methods on Tanks-and-Temples dataset and LLFF dataset.

respectively. We note that baselines perform differently with different feature dimensions for FID, so we report both FID-192 and FID-2048. CLIP Similarity refers to the average of per-frame CLIP [45] similarity between generated frames and ground-truth frames. 2) For camera accuracy, we report Absolute Trajectory Error (ATE) [16] and Relative Pose Error for translation (RPE-T) and rotation (RPE-R) [16]. More details about metrics implementation are in Appendix F.

4.1 Static Single-image Novel View Synthesis

As reported in Table 1, our proposed NeoMap framework achieves state-of-the-art performance in both visual quality and semantic consistency across the Tanks-and-Temples and LLFF datasets, securing the best scores in almost all perceptual metrics, including PSNR, LPIPS, FID, and CLIP-S. The quantitative comparisons clearly demonstrate that merely manipulating the initial noise for visible regions, as attempted by baselines such as Reversed Noise and Flow Warped Noise, is fundamentally insufficient to yield reasonable novel view synthesis, which is reflected in their substantial performance drops. Alongside these superior visual results, NeoMap maintains commendable camera control accuracy that rivals heavily constrained optimization methods. These quantitative gains are strongly corroborated by our qualitative results as shown in Figure 4: baseline models stubbornly retain warping noise due to an absolute over-reliance

Table 2: Quantitative results of all methods for dynamic monocular video novel view synthesis on DAVIS dataset.

	Visual Quality			Camera Accuracy		
	FID-192↓	FID-2048↓	FVD↓	ATE↓	RPE-T↓	RPE-R↓
TrajectoryAttention [58]	15.320	115.751	496.158	1.166	0.190	0.023
ReCamMaster [3]	3.772	115.584	532.946	0.709	0.078	0.012
ViewCrafter [64]	12.604	127.991	490.705	0.794	0.111	0.014
FlexWorld [12]	2.948	77.179	243.930	0.805	0.101	0.013
NVS-Solver [62]	2.654	81.553	341.127	0.996	0.159	0.021
LanPaint [67]	2.436	92.367	258.957	1.052	0.132	0.011
Reversed Noise [35]	3.821	92.280	321.535	0.798	0.090	0.013
Flow Warped Noise [9]	14.969	106.699	479.094	1.815	0.209	0.030
NeoMap (Ours)	1.544	68.864	234.634	0.534	0.067	0.011

on the flawed geometric prior, whereas our alternating projection approach robustly corrects visible artifacts while seamlessly hallucinating geometrically coherent content in unseen areas.

4.2 Dynamic Monocular Video Novel View Synthesis

As shown in Table 2, our NeoMap framework consistently achieves state-of-the-art performance in dynamic monocular video novel view synthesis on the DAVIS dataset, demonstrating optimal results across visual quality and camera control metrics. In complex dynamic scenes, the entanglement of global camera motion and independent object movement frequently pushes baseline models into out-of-distribution (OOD) states, allowing our method to exhibit an even more pronounced advantage here than in relatively static NVS scenarios. Furthermore, the intricate spatial-temporal correlations introduced by these complex motions often cause baseline models to severely misjudge camera trajectories, leading to a noticeable degradation in their camera control accuracy, whereas our method maintains highly robust and precise camera conditioning with significant leads in ATE and RPE metrics. As shown by further qualitative results in Figure 5, since depth estimation in dynamic scenes is notoriously inferior to that in static environments, frequently introducing severe warping deformations and noise, baseline methods are highly susceptible to pronounced object distortion and semantic deviation of the main subjects, whereas ours effectively mitigates these geometric artifacts and preserves the structural integrity of the moving subjects.

4.3 Ablation Study

To evaluate the effectiveness of each component and hyperparameter choices, we conduct the following ablation experiments on the Tanks-and-Temples dataset.

(1) Removing all the modules: We remove all the proposed modules to show the vanilla backbone performances. The method degenerates to generating with z_τ, which is equivalent to a reversed noise baseline.

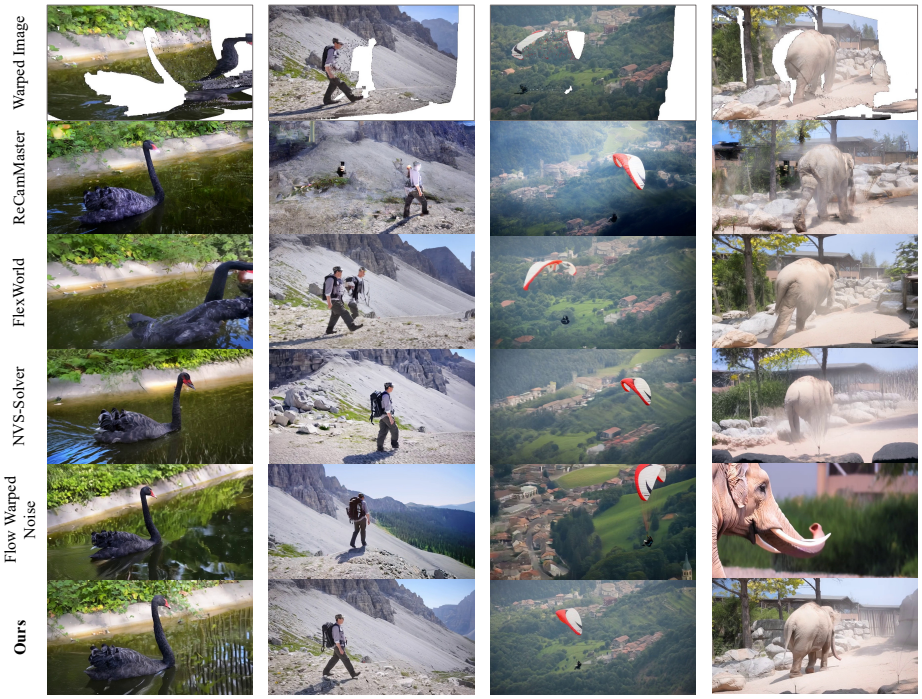

Fig. 5: Qualitative results of top 4 baselines and our methods on DAVIS dataset.

(2) The number of alternating projection iterations K: we choose $K = 15$ in our main experiments, and we ablate the performance with $K = 5$ and $K = 10$.

(3) The threshold used in the similarity mask in AMP: we choose $\lambda = 0.8$ in our main experiments, and we ablate the performance with $\lambda = 0.7$ and $\lambda = 0.9$.

(4) Removing AMP: we remove the AMP module.

(5) Removing PCP: we remove the PCP module.

(6) Applying PCP in latent space: we remove the VAE mappings in the PCP module. Instead, we interpolate the visibility mask $\boldsymbol{M}$ into the size of the latent space as $\boldsymbol{M_Z}$, and blend the latent state $\hat{z}^k_{\mathcal{M}}$ with z_c directly as $\hat{z}^k_{\mathcal{M}} = \boldsymbol{M_Z} \odot z_c + (1 - \boldsymbol{M_Z}) \odot \hat{z}^k_{\mathcal{M}}$.

(7) The number of denoising steps to apply auxiliary trajectory re-anchoring: we use the trajectory re-anchoring in the first 3 denoising steps in our main experiments, while we ablate in the first 0, 5, and 8 steps here.

(8) Adding PCP in the trajectory re-anchoring stage:

As detailed in Table 3, our proposed framework delivers massive performance gains over the naive baseline, with visual metrics steadily converging as the number of alternating projection iterations (K) increases. Our constraint mechanisms prove critical for maintaining geometric integrity: enforcing absolute pixel-space boundaries via PCP is irreplaceable, as removing it or applying it solely in the VAE latent space induces spatial bleeding and severe camera decoupling. Within

Table 3: Quantitative results of all ablation studies on Tanks-and-Temples datasets.

						Visual Quality			Camera Accuracy		
	K	λ	AMP	PCP	Re-anchor	FID-192↓	FID-2048↓	CLIP-S↑	ATE↓	RPE-T↓	RPE-R↓
1)	-	-	✗	✗	-	17.275	160.994	0.902	0.029	0.003	0.005
2)	5	0.8	✓	✓	3	4.864	96.359	0.923	0.010	0.002	0.003
2)	10	0.8	✓	✓	3	4.774	88.310	0.930	0.008	0.002	0.003
3)	15	0.7	✓	✓	3	5.240	92.013	0.926	0.010	0.002	0.003
3)	15	0.9	✓	✓	3	3.976	88.237	0.933	0.022	0.002	0.004
4)	15	0.8	✗	✓	3	4.943	91.215	0.929	0.012	0.002	0.003
5)	15	0.8	✓	✗	3	5.839	115.214	0.925	0.089	0.005	0.011
6)	15	0.8	✓	Latent	3	5.661	93.343	0.928	0.011	0.002	0.003
7)	15	0.8	✓	✓	0	2.773	83.750	0.940	0.017	0.002	0.005
7)	15	0.8	✓	✓	5	4.674	86.026	0.932	0.008	0.002	0.003
7)	15	0.8	✓	✓	8	6.326	96.721	0.926	0.007	0.002	0.002
8)	15	0.8	✓	✓	PCP	6.893	107.733	0.909	0.006	0.002	0.002
Full	15	0.8	✓	✓	3	3.712	83.341	0.935	0.008	0.002	0.003

AMP, the similarity threshold (λ) requires delicate balancing; overly low values admit low-quality warping artifacts, while excessively high values discard too many reliable priors, similarly leading to geometric decoupling and overall view shifts. Finally, the auxiliary trajectory re-anchoring introduces a crucial trade-off between raw visual fidelity and precise camera alignment. While omitting it entirely yields the highest raw image quality, it suffers from noticeable trajectory drift; applying an optimal 3 steps effectively calibrates the camera alignment with negligible visual loss, whereas excessive re-anchoring aggressively harms image quality without yielding further trajectory benefits.

4.4 Further Application

In addition to the above NVS application, we further show the ability of our methods on video-to-video (V2V) future extrapolation. Given a sequence of video frames, we directly mark them as visible part in our framework, while frames in future unobserved time as invisible part. Based on these given video priors, we directly run our method on it, which unifies training-free V2V future extrapolation task and NVS task. We show qualitative results in some selected scenes from DAVIS [44] and the training set of PhysInOne [68] as shown in Figure 6.

5 Conclusion

We present a novel entirely training-free framework that addresses the inherent challenges of severe occlusions and depth ambiguity in monocular novel view synthesis (NVS). Rather than relying on expensive task-specific fine-tuning or heuristic generation guidance, we reframed the NVS problem as discovering the optimal initial noise latent within the native output data manifold of

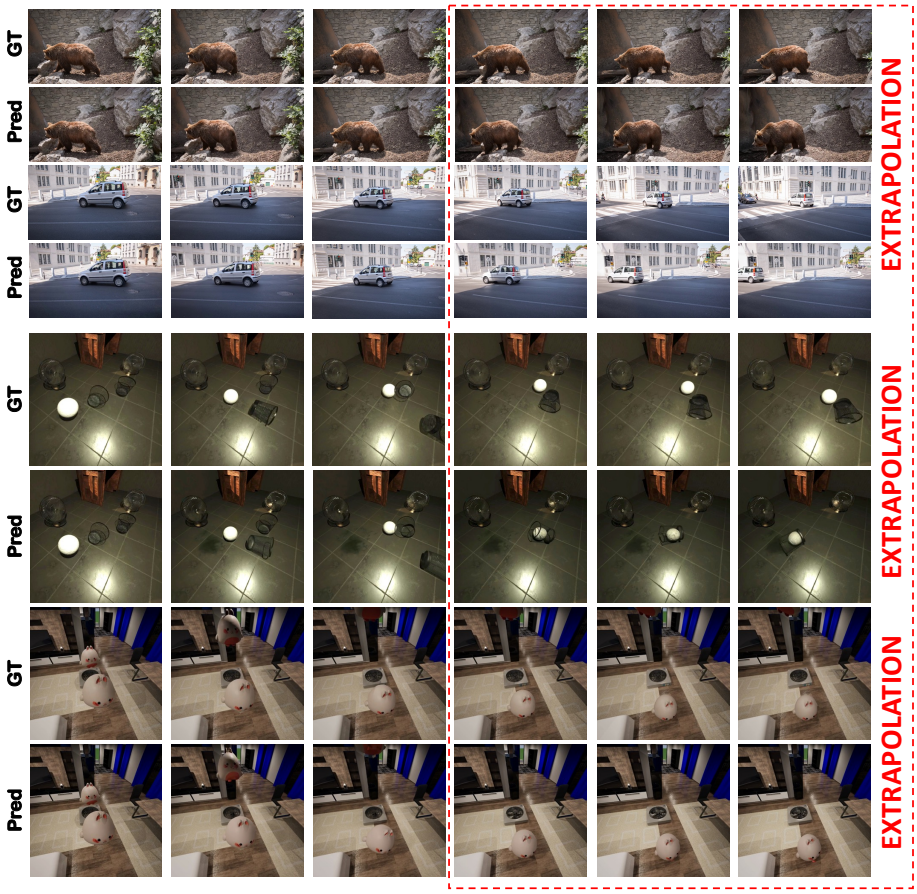

Fig. 6: Qualitative results of video-to-video future extrapolation on DAVIS dataset (upper two scenes) and PhysInOne dataset (bottom two scenes). Our method can provide smooth motion prediction based on the given videos.

a pre-trained video generation model. Our core contribution lies in the manifold alternating projection scheme: iterating between Anchored Manifold Projection (AMP) for plausible semantic completion and Pixel-Constrained Projection (PCP) for exact geometric boundary enforcement. This mathematically guaranties convergence to a solution that perfectly intersects global semantic realism with precise 3D view alignment. Extensive experiments demonstrate that our method significantly advances state-of-the-art visual quality and geometric consistency for both static image and dynamic video inputs, establishing a robust new paradigm for training-free novel view synthesis.

Acknowledgments: This work was supported in part by Research Grants Council of Hong Kong under Grants 15225522 & 15219125 &15228626, and in part by National Natural Science Foundation of China under Grant 62271431.

References

1. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. CVPR (2025)
2. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Vd3d: Taming large video diffusion transformers for 3d camera control. ICLR (2025)
3. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., Zhang, D.: Recammaster: Camera-controlled generative rendering from a single video. ICCV (2025)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv preprint arXiv:2311.15127 (2023)
5. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. CVPR (2023)
6. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., Ryoo, M., Debevec, P., Yu, N.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. CVPR (2025)
7. Ceylan, D., Huang, C.H., Mitra, N.J.: Pix2video: Video editing using image diffusion. ICCV (2023)
8. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. ICCV (2023)
9. Chang, P., Tang, J., Gross, M., Azevedo, V.C.: How i warped your noise: a temporally-correlated noise prior for diffusion models. ICLR (2024)
10. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. ICML (2025)
11. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. CVPR (2024)
12. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. arXiv preprint arXiv:2503.13265 (2025)
13. Couairon, P., Rambour, C., HAUGEARD, J.E., THOME, N.: Videdit: Zero-shot and spatially aware text-driven video editing. TMLR (2024)
14. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. ICCV (2023)
15. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. ICLR (2024)
16. Goel, P.K., Roumeliotis, S.I., Sukhatme, G.S.: Robust localization using relative and absolute position estimates. Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289) **2**, 1134–1140 vol.2 (1999), `https://api.semanticscholar.org/CorpusID:90589`

17. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
18. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
21. Hou, C., Chen, Z.: Training-free camera control for video generation. ICLR (2025)
22. Hu, T., Zhang, J., Yi, R., Wang, Y., Huang, H., Weng, J., Wang, Y., Ma, L.: Motionmaster: Training-free camera motion transfer for video generation. arXiv preprint arXiv:2404.15789 (2024)
23. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., Ren, J., Xie, K., Biswas, J., Leal-Taixe, L., Fidler, S.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
24. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly DDPM noise space: Inversion and manipulations. CVPR (2024)
25. Jang, S., Ki, T., Jo, J., Yoon, J., Kim, S.Y., Lin, Z., Hwang, S.J.: Frame guidance: Training-free guidance for frame-level control in video diffusion model. ICLR (2026)
26. Kasten, Y., Ofri, D., Wang, O., Dekel, T.: Layered neural atlases for consistent video editing. TOG **40**(6), 1–12 (2021)
27. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
28. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. ICCV (2023)
29. Kim, S., Suh, S., Lee, M.: Rad: Region-aware diffusion models for image inpainting. CVPR (2025)
30. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. TOG **36**(4) (2017)
31. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
32. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. arXiv preprint arXiv:2405.17414 (2024)
33. Ling, P., Bu, J., Zhang, P., Dong, X., Zang, Y., Wu, T., Chen, H., Wang, J., Jin, Y.: Motionclone: Training-free motion cloning for controllable video generation. arXiv preprint arXiv:2406.05338 (2024)
34. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2023)
35. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
36. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. NeurIPS (2022)

37. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models. CVPR (2022)
38. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. TOG (2019)
39. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. ECCV (2020)
40. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. arXiv preprint arXiv:2211.09794 (2022)
41. Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. CVPR (2023)
42. Pan, L., Barath, D., Pollefeys, M., Schönberger, J.L.: Global Structure-from-Motion Revisited. ECCV (2024)
43. Parmar, G., Singh, K.K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. SIGGRAPH (2023)
44. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L.V., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. CVPR (2016)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. ICML (2021)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. CVPR (2022)
47. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. arXiv preprint arXiv:2405.17251 (2024)
48. Shaulov, A., Hazan, I., Wolf, L., Chefer, H.: Flowmo: Variance-based flow guidance for coherent motion in video generation. arXiv preprint arXiv:2506.01144 (2025)
49. Song, C., Yang, Y., Zhao, T., Li, R., Zhang, C.: Taming video models for 3d and 4d generation via zero-shot camera control. CVPR (2026)
50. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. CVPR (2023)
51. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. Openreview (2019)
52. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
53. Wang, H., Liu, Y., Liu, Z., Wang, W., Dong, Z., Yang, B.: Vistadream: Sampling multiview consistent images for single-view scene reconstruction. ICCV (2025)
54. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. CVPR (2025)

55. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
56. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. SIGGRAPH pp. 1–11 (2024)
57. Wu, C.H., la Torre, F.D.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. ICCV (2023)
58. Xiao, Z., Ouyang, W., Zhou, Y., Yang, S., Yang, L., Si, J., Pan, X.: Trajectory attention for fine-grained video motion control. ICLR (2025)
59. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. SIGGRAPH pp. 1–12 (2024)
60. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. ACM SIGGRAPH Asia Conference Proceedings (2023)
61. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. ICLR (2025)
62. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. ICLR (2025)
63. Yu, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. ICCV (2025)
64. 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. arXiv preprint arXiv:2409.02048 (2024)
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CVPR (2018)
66. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. NeurIPS (2023)
67. Zheng, C., Lan, Y., Wang, Y.: Lanpaint: Training-free diffusion inpainting with asymptotically exact and fast conditional sampling. TMLR (2025)
68. Zhou, S., Wang, H., Cheng, H., Li, J., Wang, D., Jiang, J., Jin, Y., Huang, J., Mao, S., Liu, S., Yang, Y., Song, H., Wei, S., Zhang, Z., Huang, P., Liu, S., Hao, Z., Li, H., Li, Y., Zhou, W., Zhao, Z., He, Z., Wen, H., Huang, S., Yun, P., Cheng, B., Fu, P.K., Lai, W.K., Chen, J., Wang, K., Sun, Z., Li, Z., Hu, H., Zhang, D., Yuen, C.H., Wang, B., Wang, Z., Zou, C., Yang, B.: Physinone: Visual physics learning and reasoning in one suite. CVPR (2026)