PASTEL: Panoramic Alignment for Monocular 4D Scene Reconstruction

Yuankun Yang[1], Yi Wei[2], Bo Bai[2], Wenyang Zhou[2], and Li Zhang[1†]

[1]School of Data Science, Fudan University
[2]Central Media Technology Institute, Huawei

github.com/LogosRoboticsGroup/PASTEL

Abstract. Reconstructing 4D scenes from casually captured monocular video is vital for applications in virtual reality (VR) and embodied AI. Recent advances in 4D reconstruction and novel view synthesis have substantially propelled this capability. However, existing reconstruction methods generally cannot recover regions beyond visible camera limits. Consequently, we introduce a new paradigm that achieves 4D scene synthesis by combining visible-region reconstruction from monocular input with invisible-region generation beyond observable camera boundaries. A straightforward approach is to leverage video generation models as "generative priors" for invisible-region exploration. However, their inherent stochasticity and large solution space prevent stable and consistent view synthesis. As a result, naively incorporating generative content into the reconstruction process often causes artifacts, especially when camera trajectories deviate significantly from the original video. To overcome these challenges, we present Panoramic Alignment for Strategic Exploitation of Generative Priors (**PASTEL**). Specifically, PASTEL proposes panoramic scene alignment, a novel representation that reformulates the intractable 3D "invisible region" exploration into a tractable 2D directional trajectory planning. This is achieved by reducing the viewpoint planning from 6-DoF search to a 2D directional search with explicit visibility boundaries. By operating within this panoramic space, our method strategically identifies camera trajectories that maximize exploration beyond observable boundaries while minimizing viewpoint deviation. Experimental results show that PASTEL can not only extrapolate plausible scene content beyond the observable boundaries of input monocular videos, but also substantially boost monocular 4D reconstruction performance. PASTEL outperforms the previous state-of-the-art method by 0.9dB in full-image PSNR on the DyCheck IPhone dataset.

Keywords: Monocular 4D Scene Synthesis · Diffusion Models · Panorama · Gaussian Splatting

† Corresponding author: lizhangfd@fudan.edu.cn

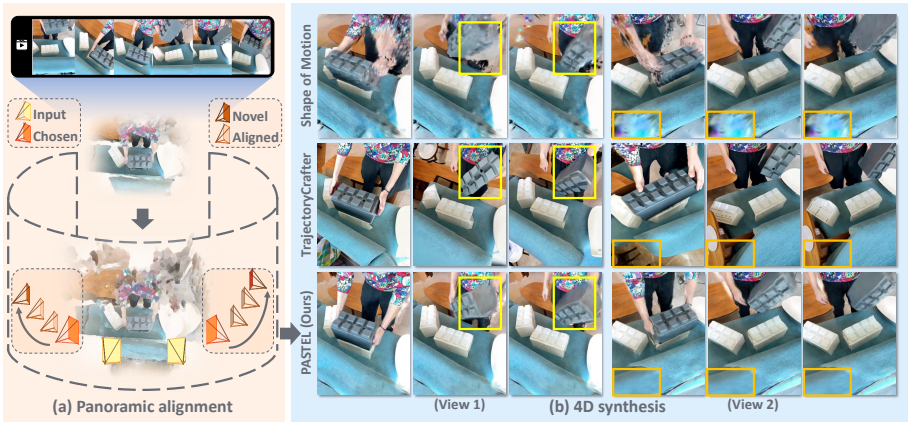

Fig. 1: Overview of PASTEL. (a) Given monocular videos, PASTEL proposes panoramic alignment to reformulate scene exploration. (b) PASTEL shows seamless extension of rendering beyond observable boundaries in orange boxed areas. PASTEL also preserves high fidelity and consistency in yellow boxed areas.

1 Introduction

Reconstructing geometry-consistent and motion-coherent 4D scenes from monocular video [18, 19] remains a fundamental and unsolved challenge in computer vision. Recent advances in novel view synthesis and dynamic scene reconstruction [20, 21, 30] have enabled progress in modeling motion dynamics within multi-view settings [16, 28]. However, these approaches face substantial limitations when extended to monocular inputs. This is because monocular videos typically exhibit limited fields of view and severe occlusions. Meanwhile, reconstruction-based methods rely on pixel-level supervision, which cannot infer regions beyond visible camera limits, as shown in the first row of Fig. 1 (b). The absence of such "invisible regions" introduces a significant gap for novel view synthesis, which limits the immersive quality and user experience of the reconstructed scenes.

To tackle these "invisible regions", an intuitive insight is to effectively leverage camera-controlled generative models [1, 2, 34, 35] as "generative priors" to infer regions beyond observable camera boundaries. However, directly applying these methods to reconstruction introduces notable limitations. Specifically, the inherent randomness within the diffusion denoising process often introduces deviations in the detailed texture and even the overall color. This deviation often results in inconsistencies and artifacts when directly applied to the 4D scene reconstruction, as exhibited in the second row in Fig. 1 (b). The greater the trajectory deviates from the original video, the more content needs to be inpainted by the generative prior, resulting in more severe artifacts. Therefore, designing effective camera trajectories is crucial. Optimal camera trajectories should maximize exploration beyond observable regions while minimizing both trajectory deviation and overlap.

However, searching for an optimal trajectory of L camera positions entails navigating a parameter space of $L \times 4 \times 4$ entries, since each camera is defined by a 4×4 extrinsic matrix. This high-dimensional Cartesian search makes the direct optimization of generative trajectories computationally challenging and prone to suboptimal, disjointed viewpoints.

To address this intractable high-dimensional search, we propose Panoramic Alignment for Strategic Exploitation of Generative Priors (PASTEL). Instead of struggling within the unconstrained 3D Cartesian space, our core idea is to reformulate scene exploration by anchoring it within a panoramic domain. Specifically, we mathematically reformulate the observable 4D scene into a spherical panoramic space. This paradigm shift offers a notable advantage: it simplifies the difficulty of high-dimensional camera viewpoint planning. This design reduces the search space from a $L \times 4 \times 4$ Cartesian domain into a 2D directional continuous expansion. Within this panoramic space, the highly complex 6-DoF search is substantially simplified. The boundaries of "invisible regions" become explicitly quantifiable and globally consistent. This elegant formulation enables us to derive expansion trajectories that maximize spatial coverage beyond observable limits while minimizing viewpoint distortion and overlaps. Subsequently, these naturally continuous panoramic trajectories provide highly stable, geometrically constrained conditioning inputs to any generative prior. By aligning the generative inpainting process within this panoramic space, our framework provides geometrically consistent conditioning that effectively constrains the diffusion randomness. This empirically yields comprehensive and geometrically consistent scene expansions. These generated components then serve as targeted pseudo-supervision to guide the 4D scene reconstruction, successfully mitigating cross-view inconsistencies.

To summarize, we make the following contributions: (1) We identify and address a fundamental bottleneck in monocular 4D synthesis: the inherent intractability of view expansions in high-dimensional Cartesian space. To address this, we introduce PASTEL, a conceptually novel framework to promote globally consistent novel view synthesis beyond observable boundaries. (2) Our core contribution is the panoramic scene alignment, a paradigm shift that reformulates the 4D scene exploration from a $L \times 4 \times 4$ Cartesian search into a constrained, continuous 2D directional field expansion. This formulation elegantly integrates the utilization of generative priors into a continuous space. (3) Extensive experiments demonstrate that our panoramic alignment seamlessly extends rendering capabilities beyond monocular video coverage while maintaining high fidelity and consistency, establishing new standards in 4D synthesis.

2 Related works

Reconstruction from monocular video Video constitutes a primary source of visual data for immersive content creation. Creating immersive experiences and constructing virtual worlds from monocular video holds significant value for applications in VR, robotics, and content creation. To address this challenge, [16]

integrates prior knowledge from multiple 2D foundation models to optimize the proposed motion scaffold representation. [28] exploits the low-dimensional structure inherent in scene motion to constrain the full-sequence-long 3D motion of each point. [19] combines Transformer architectures and image-based rendering to achieve photorealistic novel view synthesis from monocular videos. [9, 24, 39] employ a data-driven approach, training models to predict the geometry and motions from stereo image pairs. Some other works extend Gaussian Splatting into dynamic view synthesis [14, 17, 29]. They often incorporate regularization through MLPs [32] or low-rank motion representations [15]. However, these methods rely on pixel-level supervision to reconstruct 3D or 4D scenes. Consequently, they cannot infer "invisible regions" where no ground truth is available for supervision.

Camera-controlled video generation Driven by recent advances in video diffusion models, researchers have begun to explore using video generative models for directly synthesizing observations from the altered camera trajectories. Among them, [3, 34, 35, 37, 41] condition the video diffusion models with novel viewpoint cloud rendering. In contrast, [10, 11, 42] discard the explicit geometry estimation by only conditioning the model with the Plücker embedding of target views. [1, 2] adopt the Diffusion Transformer (DiT) [22] architecture with view or frame attention, generating videos from multiple fixed views or a specified trajectory. Despite the remarkable success in view synthesis, these methods typically exploit denoising process from diffusion models for video generation, which exhibit substantial randomness. This randomness usually leads to severe conflict and inconsistency when directly applied to 3D/4D scene reconstruction.

Panoramic representations Classical approaches assume the panoramic scene as a static image plane. [25, 43] propose cylindrical warping and seam matching to assemble multiple frames onto a unified cylindrical panoramic surface. [36, 38] introduced parallax-tolerant image stitching processes, relaxing the strict planar assumption but still producing static panoramas. To further improve reconstruction of object surfaces, [40] developed layered depth panoramas, and [12, 13] introduced mesh-based representations for interactive rendering. Recently, [5] proposed spherical neural light fields for implicit panoramic image stitching and re-rendering. These advances support vivid wide-angle rendering, but they remain limited by the lack of explicit geometry and physical constraints, making it difficult to represent dynamic object motion and interactions. Unlike panorama stitching for visualization, our work leverages panoramic representations for adaptive trajectory planning and for conditioning off-the-shelf generative priors.

3 Methods

We propose PASTEL for high-quality, unbounded 4D scene synthesis from casually captured monocular videos. Instead of relying on a fragmented pipeline of empirical heuristics, PASTEL is built upon a novel panoramic alignment. As

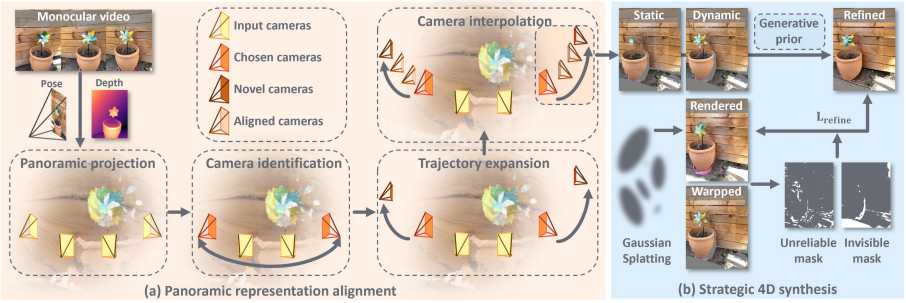

Fig. 2: Architecture overview of PASTEL. (a) Given a monocular video with predicted camera poses and depth, PASTEL designs panoramic representation alignment by projecting all video frames into a unified panoramic space. PASTEL then devises chosen cameras serving as the beginning of trajectories. (b) Along the designated trajectory, both static and dynamic components are utilized to achieve a comprehensive view expansion, which feeds into the diffusion prior for refinement. Finally, we compare warped and rendered images from Gaussian Splatting to identify unreliable and invisible regions, which are used to strategically supervise the 4D scene.

illustrated in Fig. 2, the core of our approach is the panoramic reformulation of the 4D scene (Sec. 3.1). By transforming the Cartesian exploration into a continuous panoramic domain, we can effectively derive expansion trajectories (Sec. 3.2). This formulation provides a principled structural constraint for external generative priors, ensuring that the synthesized unobservable regions are intrinsically consistent with the observable scene. These geometrically bounded priors then guide the holistic 4D synthesis, yielding high fidelity and cross-view consistency without relying on post-hoc corrections (Sec. 3.3).

3.1 Panoramic scene alignment

Conventionally, trajectory discovery is performed in full 3D Cartesian space. For M trajectories, each with L viewpoints and full 4×4 extrinsic matrices, the system must optimize $M \times L \times 4 \times 4$ parameters. In practice, such empirical optimization often leads to suboptimal scene coverage and visible inconsistencies due to redundant or poorly conditioned viewpoints. To address this fundamental bottleneck, PASTEL abandons the Cartesian search space entirely. Instead, we introduce a panoramic scene reformulation [25, 43] that inherently enforces global geometric consistency and provides an overall view of the scene.

Given a monocular video of T frames, we first back-project each frame into 3D space using our precomputed depth map, camera extrinsics, and intrinsics. This step produces a sequence of 3D point clouds, each aligned to the world coordinate system. We then compute the spherical center of the panorama $P^c = (x^c, y^c, z^c)$. Specifically, to make P^c robust to foreground dynamic objects, we apply depth-percentile filtering to remove the bottom 30% depth points before computing

the mean of the remaining 3D points. We then reproject all 3D points into the panoramic space. Concretely, for each point in point cloud $p_t^j = (x_t^j, y_t^j, z_t^j) \in P_t$, we first compute the polar coordinates as:

$$\theta_t^j = \text{atan2}\left(y_t^j - y^c, x_t^j - x^c\right), \tag{1}$$

$$\phi_t^j = \text{atan2}\left(z_t^j - z^c, \sqrt{(x_t^j - x^c)^2 + (y_t^j - y^c)^2}\right). \tag{2}$$

Finally, we compute the pixel coordinates within the panorama:

$$u_t^j = \left(\frac{\theta_t^j}{2\pi} + 0.5\right) \times w_p, \quad v_t^j = \left(0.5 - \frac{\phi_t^j}{\pi}\right) \times h_p, \tag{3}$$

where w_p and h_p are width and height of 2D panorama. This produces a dense 2D panorama that encodes the global scene. Meanwhile, for each image frame captured by the camera in timestamp t, we can also track all its image pixels through the projection. This could yield projected point sets in the panorama for each image, which is denoted as $M_t = \{(u_t^j, v_t^j) \mid j = 1, ..., n_t\}$.

Compared with Cartesian coordinates, the panoramic alignment collapses the high-dimensional trajectory parameterization into a compact spherical coordinate system. This is not merely a coordinate transformation, but a fundamental conceptual shift that makes trajectory optimization mathematically better constrained and explicitly exposes the global topological structure of the scene. Within this domain, PASTEL identifies expansion trajectories that enable consistent and efficient novel-view expansion.

3.2 Adaptive trajectory identification under panorama

After projecting the video into panoramic space, PASTEL seeks to determine the optimal camera trajectory under the panoramic space. Empowered by the panoramic formulation, the 6-DoF camera placement is reparameterized as a guided 2D directional ray-casting. This structural simplification enables an interpretable search for the outermost visible boundaries. The optimal trajectory should balance two goals: minimal deviation from the original cameras and maximal spatial coverage beyond the observed region. Achieving this balance ensures that generative priors can produce high-quality and consistent inference of regions beyond visible camera limits.

To ensure uniform exploration with minimum overlap, we determine our trajectory direction d_m through the evenly expanded angle. This is defined as:

$$d_m = (d_{m,u}, d_{m,v}) = \left(\cos\left(\frac{2\pi \cdot m}{M}\right), \sin\left(\frac{2\pi \cdot m}{M}\right)\right), \tag{4}$$

where $m \in \{1, ..., M\}$, M is the total number of evenly expanded directions. Each d_m represents a unit vector in the panoramic plane corresponding to the angle $\frac{2\pi \cdot m}{M}$. For each direction, we choose a starting camera by finding the point in the

panorama that extends furthest along d_m. In Sec. 3.1, we have achieved the point sets $M_t = \{(u_t^j, v_t^j) \mid j = 1, ..., n_t\}$ for each image frame captured by the camera in timestamp t under the panorama coordinate. Consequently, we can identify the chosen camera c_m by finding the point with the maximum projection onto direction d_m within each point set M_t. To prevent the identified trajectory from starting excessively close to or inside a foreground dynamic object which could lead to blocked views or geometrical failures, we introduce a valid search radius constraint R based on the 3D scene depth, which is defined as:

$$R = \text{median}(\|p - P_c\|_2) \cdot \tan(\alpha/2), \tag{5}$$

where $\|p - P_c\|_2$ is the Euclidean distance from each 3D point p to the spherical center P_c, and α is the angular step of trajectory expansion. This constraint ensures the selected starting point maintains a safe distance from foreground objects. This is done by maximizing the inner product between the relative 2D panorama point sets and the direction d_m within the valid region:

$$c_m = \arg\max_t \left\{ \max_{j \in \Omega_R} \left(\left(u_t^j - \frac{w_p}{2} \right) d_{m,u} + \left(v_t^j - \frac{h_p}{2} \right) d_{m,v} \right) \right\}, \tag{6}$$

where $\frac{w_p}{2}, \frac{h_p}{2}$ is the center of the 2D panorama. Ω_R denotes the set of points satisfying the depth constraint R. It ensures that the trajectory begins at the edge of the visible region and naturally expands outward without intersecting foreground objects.

Then, we expand the camera trajectory from the chosen camera c_m through the direction d_m in the panoramic space. The shifted location is then mapped back into Cartesian space, yielding the position of the novel camera. The orientation of the novel camera is defined as the direction pointing from the novel camera position toward the spherical center. This design ensures that new viewpoints remain consistent with the scene structure while allowing efficient exploration beyond the observed region. After that, we create a trajectory by interpolating between the novel camera $(\tilde{R}_{t,m}, \tilde{\mathbf{t}}_{t,m})$ and the original camera $(R_t, \mathbf{t}_t)$. We apply spherical linear interpolation (SLERP) for rotations and linear interpolation for translations:

$$(R_{t,m}^{\theta_l}, \mathbf{t}_{t,m}^{\theta_l}) = \text{Interpolate}\big((R_t, \mathbf{t}_t), (\tilde{R}_{t,m}, \tilde{\mathbf{t}}_{t,m}), \theta_l\big), \tag{7}$$

where $\{\theta_l\}_{l=1}^{L} \in [0, 1]$ denotes the interpolation factor.

By calculating trajectory directions through evenly expanded angles and interpolating between original and new camera positions, we create trajectories with minimized deviation and maximum exploration of areas beyond observable boundaries. This is exclusively enabled by the panoramic representation, which reduces the target trajectory's degrees of freedom from a $L \times 4 \times 4$ space to a directional search. Crucially, this trajectory design is not a heuristic post-processing step, but a principled structural prior for stabilizing external generative models. It avoids abrupt viewpoint jumps, maintains global scene coherence, and structurally enforces that novel views align smoothly with the physical bounds of the observable scene. As a result, this progression directly suppresses the generative drift inherent to diffusion models.

3.3 Comprehensive view expansion

After deriving the camera trajectory, we leverage it to guide generative priors for 4D synthesis beyond visible camera limits. To rigorously bound the stochasticity of the generative prior, we propose a holistic view expansion that naturally integrates static and dynamic elements within the derived panoramic trajectories. This alignment preserves coherence with the original video, thereby supporting the generation of more consistent videos.

Specifically, we first reproject the static regions of each frame into 3D point clouds P_t^{static}. The static areas are obtained by masking out moving objects using pre-computed dynamic masks from optical flow estimation. Next, we merge all static point clouds across time to form a denser and wider 3D coverage $\bigcup_{t=1}^{T} P_t^{\text{static}}$. This aggregated point cloud is then projected into each target trajectory view, producing static background images I_t^{static}. We then process the moving regions separately. Each frame, together with its depth map, is reprojected onto the target views as I_t^{moving}. This provides the dynamic components that align with the designed camera trajectory. Finally, we combine the static background and the reprojected moving objects as:

$$I_t^{\text{warped}} = I_t^{\text{static}} \cup I_t^{\text{moving}}, \tag{8}$$

where t denotes different timestamps of the frame. In this view expansion design, static content from all timestamps can be merged into a unified, viewpoint-consistent background. It prevents the diffusion model from hallucinating inconsistent geometry and provides a stable global reference across the entire 4D sequence. The resulting warped frames form a trajectory-aligned video $\mathcal{V}^{\text{warped}} = \{I_t^{\text{warped}}\}_{t=1}^{T}$, which is applied as guidance for camera-controlled video re-generation. The generative prior then refines this warped video to produce the refined outputs $\tilde{\mathcal{V}}'$.

PASTEL further designs a strategic supervision approach to minimize the adverse effects of RGB estimation deviations from generative priors. The key for strategic supervision is to exploit synthesized videos $\tilde{\mathcal{V}}'$ and the previously warped video $\mathcal{V}^{\text{warped}} = \{I_t^{\text{warped}}\}_{t=1}^{T}$ to optimize unreliable and invisible regions of the target 4D scene. Specifically, we identify invisible areas as areas not covered by $\{I_t^{\text{warped}}\}_{t=1}^{T}$. Note that we store invalid pixels as -1 during warping, this design could achieve the invisible mask as:

$$M_t^{\text{invisible}} = \mathbf{1}\left(\{I_t^{\text{warped}}\}_{t=1}^{T} < 0\right). \tag{9}$$

We further detect unreliable areas by comparing the Structural Similarity Index (SSIM) between the rendered RGB images $\{I_t^{\text{render}}\}_{t=1}^{T}$ and the warped RGB images $\{I_t^{\text{warped}}\}_{t=1}^{T}$. We set areas where the SSIM falls below a threshold ϵ as unreliable to achieve the unreliable mask:

$$M_t^{\text{unreliable}} = \mathbf{1}\left(\text{SSIM}(I_t^{\text{render}}, I_t^{\text{warped}}) < \epsilon\right). \tag{10}$$

Finally, we optimize our target scene representation through both the invisible mask and the unreliable mask. The refinement loss is thus defined as:

$$L_{\text{refine}} = L_{\text{rgb}}(M_t^{\text{refine}} \cdot I_t^{\text{render}}, M_t^{\text{refine}} \cdot I_t^{\text{refined}}), \tag{11}$$

where $M_t^{\text{refine}} = M_t^{\text{invisible}} \cup M_t^{\text{unreliable}}$ denotes regions for refinement.

By introducing these invisible and unreliable masks, we establish a mutually beneficial collaboration between the reconstruction model and the generative prior. The generative prior is strictly confined to inpainting unobserved or severely degraded regions, effectively preventing the diffusion model's stochasticity from overwriting high-fidelity, well-reconstructed areas. Consequently, this strategic supervision effectively mitigates the pervasive artifacts typically caused by unconstrained generative priors in 4D synthesis. Specifically, our pipeline (Fig. 2) depicts step-by-step panoramic projection, trajectory expansion, and supervised distillation with panoramic visualizations. In our implementation, MoSca serves as the underlying 4D representation, while TrajectoryCrafter acts as the generative prior to guide the 4DGS rendering and optimization. We also provide more panoramic visualizations in the supplementary material to clarify the pipeline.

4 Experiments

4.1 Experimental setup

Datasets. To quantitatively assess our rendering outcomes both within and beyond the camera's line of sight, we compare our approach to existing methods using the currently most challenging dataset—the DyCheck IPhone [19]. This dataset provides covisibility masks that encompass regions outside the range of all training cameras, making it ideally suited for a comprehensive evaluation of our performance. We additionally evaluate on Kubric-4D [4, 7] to assess view expansion under large camera displacements. We further conduct experiments on the NSFF dataset [33] to measure the ability of our method with fewer frames, and evaluate the generalization of PASTEL through reconstruction from real-world videos.

Metrics. Previous methods evaluated on the DyCheck IPhone dataset [19] often focus only on areas within covisibility masks, utilizing covisibility masked metrics such as mPSNR, mSSIM, and mLPIPS. However, examining occluded and out-of-view regions is crucial for enhancing user or robotic inference experiences in virtual reality and embodied AI contexts. Therefore, we also evaluate all methods, including areas beyond covisibility masks to assess their ability to extend views outside the camera boundaries and ensure the completeness of 4D scenes, which are denoted as full-image PSNR, SSIM, and LPIPS. Note that standard benchmarks do not provide pixel-level ground truth for purely invisible regions. We thus report metrics as in prior work [8, 16].

Fig. 3: Qualitative comparison on DyCheck [19]. The yellow boxed areas highlight synthesis beyond covisible regions. Our method achieves superior overall quality and consistency with the ground truth, excelling in both reconstruction fidelity and the ability to induce details in areas beyond covisible regions.

Implementation details. Our panoramic alignment is parameterized in a spherical coordinate system. The width and height of 2D panorama are empirically set as $w_p = 2048$ and $h_p = 2048$. Each pixel corresponds to a 3D direction vector defined by the azimuthal angle $\theta \in [0, 2\pi)$ and polar angle $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. This mapping adheres to the standard spherical-to-Cartesian conversion. Our pipeline builds upon standard estimators: UniDepth [23] for depth, BootsTAPIR [6] for pose, and RAFT [26] for dynamic masks when not precomputed. Nevertheless, our core contribution lies in panoramic alignment rather than these modules. Our generative prior is TrajectoryCrafter [35] (50 inference steps). The 4D scene is represented via Motion Scaffolds [16], optimized for 6000 steps. A detailed robustness analysis of depth, pose, and generative prior variations is provided in the supplementary material. We set $\epsilon = 0.3$ for the unreliable mask. The frame length $L = 49$ follows [35]. The interpolation factors θ_l are evenly distributed in $[0, 1]$. We set $M = 8$ expanded directions to cover extended regions with minimal overlap. Sensitivity to M, L, and ϵ is analyzed in the supplementary.

Table 1: Comparison with reconstruction-based and generation-based methods on DyCheck [19]. "Rec." denotes reconstruction time in hours. "FPS" denotes rendering speed. Our model significantly outperforms all baselines in PSNR, SSIM, and LPIPS, with comparable time cost.

Method	PSNR↑	SSIM↑	LPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓	Rec. (h)	FPS
4D GS [30]	15.71	0.450	0.398	16.54	0.594	0.347	1.2	44
Shape-of-Motion [28]	16.79	0.510	0.391	17.32	0.598	0.296	2.0	40
MoSca [16]	17.33	0.572	0.355	19.32	0.706	0.264	1.3	38
USPLAT4D [8]	17.63	0.583	0.341	19.63	0.716	0.250	-	-
Cat4D [31]	15.91	0.427	0.398	17.39	0.607	0.341	-	-
TrajectoryCrafter [35]	12.71	0.324	0.519	14.24	0.417	0.519	-	-
CogNVS [4]	15.42	0.362	0.689	16.94	0.449	0.598	-	-
PASTEL (Ours)	**18.54**	**0.582**	**0.335**	**19.75**	**0.739**	**0.247**	1.5	38

Performance remains stable within a broad range around these defaults. The preprocessing (including depth extraction, dense optical flow, and pose estimation) takes about 5min. The trajectory expansion and TrajectoryCrafter generation take approximately 10min. The final 4D GS reconstruction requires about 1.5h, and the rendering speed reaches 38FPS.

4.2 Comparison with state of the art

We compare the performance with both reconstruction-based methods and generative-based methods on the DyCheck IPhone dataset [19], the NSFF dataset [33], the Kubric-4D dataset [7], and other real-world monocular videos.

On the DyCheck IPhone dataset [19], the quantitative results are reported in Tab. 1, and qualitative results are shown in Fig. 3. Reconstruction-based techniques typically excel in regions within the camera's field of view, showing competitive performance in covisibility masked metrics, such as mPSNR, mSSIM, and mLPIPS. However, these methods often lack the capability to interpret occluded regions or areas that fall outside the coverage of covisibility masks, with no prediction of areas beyond the covisibility masks, as illustrated in the red box parts in Fig. 3. On the other hand, generation-based methods possess the ability to induce regions beyond the covisibility masks but may suffer from inconsistencies in illumination and color fidelity. This is attributed to the denoising processes inherent in generative models, which can introduce dataset-specific color biases that distort the original hues. In comparison, PASTEL demonstrates consistency with monocular video inputs while effectively extending its rendering capabilities to regions that fall outside the coverage of monocular videos in Fig. 3. PASTEL thus achieves state-of-the-art performance across areas both within and beyond the covisibility mask in Tab. 1, outperforming the previous best reconstruction method [8] by 0.9dB in PSNR. Moreover, PASTEL achieves the best overall quality on the NSFF dataset [33], see supplementary for details.

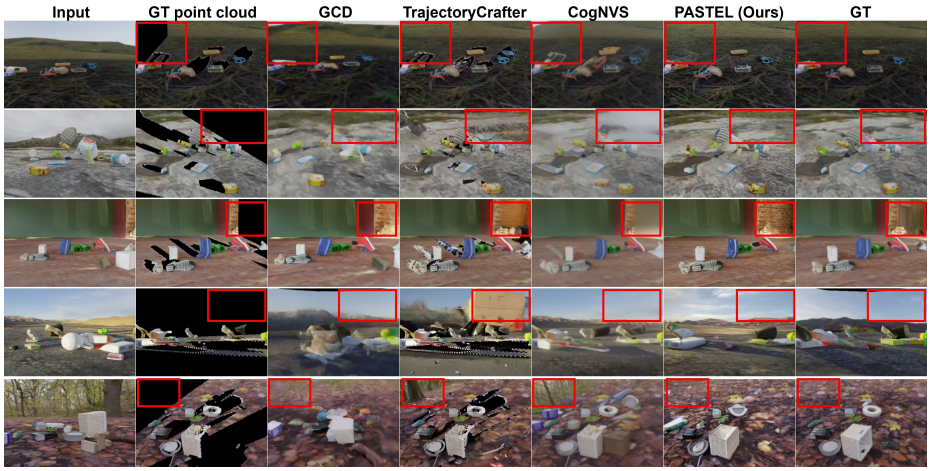

Fig. 4: Qualitative comparison on Kubric-4D [7]. The red boxed areas highlight synthesis in previously unobserved regions under large camera displacements. Our method achieves superior overall quality and consistency with the ground truth.

Table 2: Comparison with generation-based methods on Kubric-4D [7]. Our model achieves the best full-image PSNR, LPIPS under large camera displacements, while remaining competitive on the remaining metrics.

Method	PSNR↑	SSIM↑	LPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓
GCD [27]	18.59	0.555	0.383	18.68	0.560	0.312
TrajectoryCrafter [35]	20.93	0.730	0.257	23.44	0.781	0.186
CogNVS [4]	22.63	**0.760**	0.232	**24.85**	**0.799**	0.162
PASTEL (Ours)	**22.75**	0.733	**0.212**	24.25	0.781	**0.149**

On Kubric-4D [7], the quantitative results are reported in Tab. 2, and qualitative comparisons are shown in Fig. 4. This benchmark involves substantially larger camera displacements than DyCheck, providing a more challenging test of view expansion under extreme novel trajectories. Compared with prior methods in Fig. 4, existing approaches often fail to faithfully complete the red boxed invisible regions under large camera motion, leaving missing geometry, blur, or temporal inconsistency. Generation-based methods, including GCD [27], TrajectoryCrafter [35], and CogNVS [4], can synthesize previously unobserved content, but still struggle to balance perceptual quality with cross-view consistency across the full image. In comparison, PASTEL combines the cross-view consistency of 4D reconstruction with the inpainting capability of generative priors, enabling coherent synthesis in the red boxed invisible regions while preserving fidelity in observed areas. As shown in Tab. 2, PASTEL achieves the best full-image

Fig. 5: Reconstruction on monocular videos with complex motion and heavy occlusions. PASTEL captures highly dynamic scenes with irregular motion, self-occlusion, and dis-occlusion. Compared with MoSca [16], PASTEL complements geometry of dynamic content in unseen and occluded regions while improving static backgrounds at the same time.

Table 3: Ablation study on different components of the system. Panoramic representation, trajectory alignment, and view expansion are crucial for views beyond the camera range. Strategic supervision markedly improves performance.

Components	PSNR↑	SSIM↑	LPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓
Full model	**18.54**	**0.582**	**0.335**	**19.75**	**0.739**	**0.247**
w/o panorama (linear)	15.73	0.497	0.370	19.42	0.728	0.249
Cartesian-TopK	17.31	0.548	0.352	19.55	0.733	0.251
w/o trajectory alignment	17.87	0.572	0.342	19.57	0.736	0.247
w/o view expansion	17.44	0.564	0.357	19.45	0.733	0.250
w/o strategic supervision	14.09	0.306	0.548	14.80	0.506	0.445

PSNR, while remaining competitive with CogNVS on the remaining metrics. This demonstrates that our panoramic trajectory planning generalizes effectively beyond the small-baseline setting of DyCheck and supports complete 4D scene synthesis under large camera displacements. We further evaluate PASTEL on highly dynamic scenes with significant occlusions and dis-occlusions. As shown in Fig. 5, previous state-of-the-art method [16] often produce missing or broken geometry in the occluded moving body parts. In contrast, PASTEL maintains coherent motion trajectories and complements geometry under irregular human motion and heavy occlusions. PASTEL also delivers high-fidelity synthesis on other real-world monocular videos with crowded scenes, heavy occlusion, and complex motion. Please refer to the supplementary material for details.

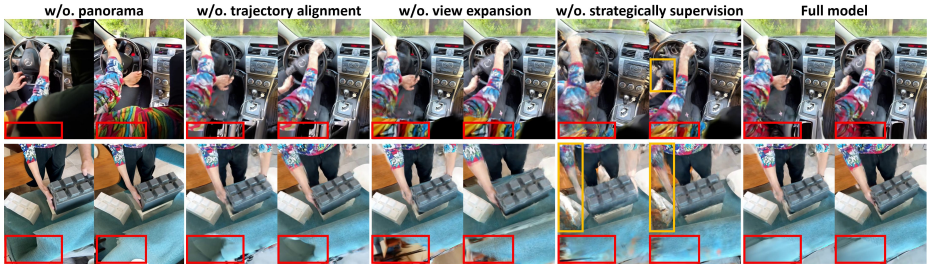

Fig. 6: Visual comparison of ablations. As exhibited in the red-boxed areas, panoramic representation facilitates the effective extension of views beyond camera boundaries. Trajectory alignment and view expansion contribute to the consistent synthesis of extended views of red-boxed areas. Strategic supervision is pivotal in harmonizing the generative content and the original monocular video, culminating in coherent 4D renderings in the orange-boxed areas.

4.3 Ablation study

To demonstrate that our performance gains fundamentally stem from our core conceptual contribution—the panoramic reformulation—rather than merely relying on external generative priors or depth/pose estimators, we systematically ablate the key components of PASTEL on the DyCheck dataset [19] in Tab. 3 and Fig. 6. We further include more ablations on hyper-parameters in the supplementary.

Panorama. Our core contribution is reformulating the 4D scene reconstruction into a panoramic space. This formulation simplifies the Cartesian camera search into a 2D directional expansion. To isolate the contribution of our panoramic formulation, we compare against a linear Cartesian baseline ("w/o panorama (linear)" in Tab. 3) and a stronger Cartesian baseline ("Cartesian-TopK") that selects coverage-maximizing TopK views from 100 random trajectories. Without panoramic projection, trajectory planning must operate in full 3D Cartesian space. We extend trajectories linearly from the first frame for the linear baseline, as no existing method provides structured directional planning for this task. As shown in Fig. 6 and Tab. 3, without our panoramic reformulation, the identical generative prior collapses under the unconstrained Cartesian search space, leading to suboptimal coverage and a drop in all quantitative metrics. Even with the stronger Cartesian-TopK sampling strategy, PASTEL still outperforms it significantly, confirming the panoramic formulation's advantage beyond sampling strategy. The panorama provides structured trajectory planning and geometric warping priors, not just camera placement. The substantial performance gap between our full model and these Cartesian baselines strongly supports that the strong performance of PASTEL is unlocked by our panoramic framework.

Trajectory alignment. Rather than being an isolated engineering step, trajectory alignment is a natural extension derived from our panoramic formulation. It ensures that the newly synthesized viewpoints evolve continuously within the spherical space. This continuous alignment intrinsically bounds the generative prior by avoiding abrupt viewpoint jumps, which improves PSNR, SSIM, and LPIPS across the entire image independently of the prior's specific architecture.

View expansion. The view-expansion mechanism in Eq. (8) leverages the globally consistent nature of our panoramic domain to merge static content from all timestamps into a unified background. This cohesive representation naturally ensures consistent inference for extended views, yielding substantial improvements in all metrics and further demonstrating the intrinsic value of operating in the panoramic space.

Strategic supervision. Our framework does not indiscriminately absorb the outputs of external generative modules. Instead, the strategic-supervision design in Eq. (11) acts as a rigorous filter, selectively applying generative priors only to invisible and unreliable regions based on geometric confidence. This principled approach is pivotal in reconciling the inherent hallucinations of diffusion models with the physics of the original monocular video. As demonstrated in Fig. 6 and Tab. 3, this strategy effectively quarantines generative errors, explicitly proving that our carefully designed framework is responsible for maintaining the physical integrity and geometric coherence of the final 4D rendering.

5 Conclusion

In this paper, we introduce PASTEL, a framework for monocular 4D scene synthesis that combines visible-region reconstruction with invisible-region generation beyond observable camera boundaries. Our approach employs Panoramic Alignment for Strategic Exploitation of Generative Priors, overcoming the cross-view inconsistency and stochasticity typical of generative models while extending rendering beyond monocular coverage. Central to our approach is the panoramic representation, which simplifies unconstrained 3D Cartesian scene exploration into a structured 2D search. This enables the effective planning of target cameras and trajectories. It optimizes the conditioning for generative priors and ensures high-fidelity inference beyond observable boundaries. We anticipate that the unbounded 4D scene synthesis achieved by PASTEL will unlock significant potential for immersive Virtual Reality experiences and embodied AI.

Acknowledgement.

This work was supported in part by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123004), Ningbo grant (2025Z038) and National Natural Science Foundation of China (Grant No. 62376060).

References

1. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint (2025)
2. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In: ICLR (2025)
3. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: SIGGRAPH (2025)
4. Chen, K., Khurana, T., Ramanan, D.: Reconstruct, inpaint, finetune: Dynamic novel-view synthesis from monocular videos. In: NeurIPS (2025)
5. Chugunov, I., Joshi, A., Murthy, K., Bleibel, F., Heide, F.: Neural light spheres for implicit image stitching and view synthesis. In: SIGGRAPH (2024)
6. Doersch, C., Luc, P., Yang, Y., Gokay, D., Koppula, S., Gupta, A., Heyward, J., Rocco, I., Goroshin, R., Carreira, J., Zisserman, A.: BootsTAP: Bootstrapped training for tracking-any-point. In: ACCV (2024)
7. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanasekar, D., Goleminov, S., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: CVPR (2022)
8. Guo, F., Hsu, C.C., Ding, S., Zhang, C.: Uncertainty matters in dynamic gaussian splatting for monocular 4d reconstruction. ICLR (2026)
9. Han, J., An, H., Jung, J., Narihira, T., Seo, J., Fukuda, K., Kim, C., Hong, S., Mitsufuji, Y., Kim, S.: Dˆ 2ust3r: Enhancing 3d reconstruction with 4d pointmaps for dynamic scenes. arXiv preprint (2025)
10. 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 (2024)
11. He, H., Yang, C., Lin, S., Xu, Y., Wei, M., Gui, L., Zhao, Q., Wetzstein, G., Jiang, L., Li, H.: Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. arXiv preprint (2025)
12. Hedman, P., Alsisan, S., Szeliski, R., Kopf, J.: Casual 3d photography. ACM TOG (2017)
13. Hedman, P., Kopf, J.: Instant 3d photography. ACM TOG (2018)
14. Huang, J., Miao, S., Yang, B., Ma, Y., Liao, Y.: Vivid4d: Improving 4d reconstruction from monocular video by video inpainting. In: ICCV (2025)
15. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: CVPR (2024)
16. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: CVPR (2025)
17. Li, C., Gu, J., Chen, J., Qiu, F., Sun, L.: Gaussian sequences with multi-scale dynamics for 4d reconstruction from monocular casual videos. arXiv preprint (2026)
18. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: CVPR (2021)
19. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: Dynibar: Neural dynamic image-based rendering. In: CVPR (2023)
20. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 3DV (2024)
21. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: ICCV (2021)

22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
23. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. arXiv preprint (2024)
24. Sucar, E., Lai, Z., Insafutdinov, E., Vedaldi, A.: Dynamic point maps: A versatile representation for dynamic 3d reconstruction. arXiv preprint (2025)
25. Szeliski, R., Shum, H.Y.: Creating full view panoramic image mosaics and environment maps. Computer graphics and interactive techniques (1997)
26. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020)
27. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: ECCV (2024)
28. Wang, Q., Ye, V., Gao, H., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. arXiv preprint (2024)
29. Wu, D., Liu, F., Hung, Y.H., Qian, Y., Zhan, X., Duan, Y.: 4d-fly: Fast 4d reconstruction from a single monocular video. In: CVPR (2025)
30. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: CVPR (2024)
31. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. arXiv preprint (2024)
32. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: CVPR (2024)
33. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: CVPR (2020)
34. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In: ICLR (2025)
35. YU, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. arXiv preprint (2025)
36. Zaragoza, J., Chin, T.J., Brown, M.S., Suter, D.: As-projective-as-possible image stitching with moving dlt. In: CVPR (2013)
37. Zhang, D.J., Paiss, R., Zada, S., Karnad, N., Jacobs, D.E., Pritch, Y., Mosseri, I., Shou, M.Z., Wadhwa, N., Ruiz, N.: Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. arXiv preprint (2024)
38. Zhang, F., Liu, F.: Parallax-tolerant image stitching. In: CVPR (2014)
39. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. In: ICLR (2025)
40. Zheng, K.C., Kang, S.B., Cohen, M.F., Szeliski, R.: Layered depth panoramas. In: CVPR (2007)
41. Zheng, S., Yin, M., Hu, W., Li, X., Shan, Y., Fu, Y.: Versecrafter: Dynamic realistic video world model with 4d geometric control. In: CVPR (2026)
42. Zhou, J.J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint (2025)
43. Zomet, A., Levin, A., Peleg, S., Weiss, Y.: Seamless image stitching by minimizing false edges. IEEE TIP (2006)