A. Appendix

A.1. User study.

To evaluate our approach, we conducted a user study comparing Ours, TrajectoryCrafter [38], TrajectoryAttention [31], and CameraCtrl [8] across four key metrics: View Angle, General Quality, Smoothness, and Background Quality. Participants viewed generated videos and selected the most visually appealing results for each criterion, providing subjective feedback on the overall quality and realism. As shown in Table 5, our method consistently achieved the highest user preference, particularly excelling in General Quality (36%) and Background Quality (39%), which highlights its superior fidelity and ability to preserve scene details. The View Angle metric (30%) confirms accurate and convincing novel-view synthesis, while Smoothness (33%) indicates our approach produces fluid transitions with minimal distortion or artifacts. These results collectively demonstrate that our method offers a more immersive and visually coherent experience compared to competing techniques.

Table 5. **User study.** Winning rates across four evaluation metrics. Our method consistently outperforms the baselines, particularly in General Quality and Background Quality.

Method	View Angle	General Quality	Smoothness	BG Quality
Ours	30%	**36%**	33%	**39%**
TrajectoryCrafter	**32%**	30%	27%	28%
TrajectoryAttention	27%	26%	**34%**	23%
CameraCtrl	11%	8%	6%	10%

A.2. Pre-trained model checkpoints

Zero4D is developed based on publicly available, pre-trained generative models for both images and videos. For transparency and reproducibility, we specify below the exact versions of each model employed in our framework:

- Depth estimation model: Depthcrafter
- Image-to-Video generation model: stable-video-diffusion-img2vid-xt

A.3. Camera trajectory control

We support various camera motions for novel view synthesis, leveraging depth information for realistic scene transformation:

Camera orbit rotation: Horizontal camera movement around the subject, creating a side-to-side viewing effect. The depth map guides proper parallax by determining each pixel's displacement based on its relative depth.

Dolly movement: Forward/backward camera translation that adjusts focal length to maintain subject size. For dolly-in, foreground elements remain stable while the background compresses; for dolly-out, the background expands naturally.

Elevation transition: Vertical camera movement that rotates the viewpoint up or down. Depth information ensures accurate perspective shifts as the camera changes height, maintaining geometric consistency.

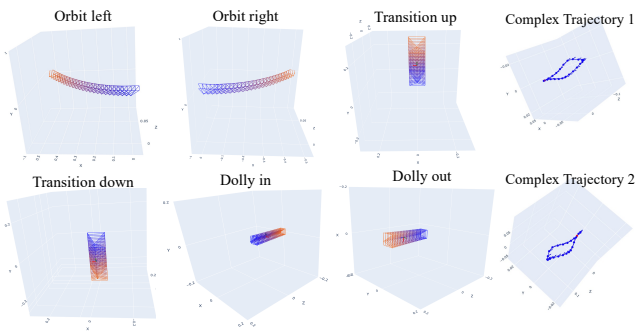

Figure 5. **Camera trajectory visualization.** With a monocular depth estimation model, our approach can generate various novel view videos with spatio-temporal synchronized videos.

Complex trajectory: We also conducted experiments on complex camera trajectories. In this setting, the camera moves along a combined path in the x, y, and z axes, first moving inward toward the subject and then moving outward, forming a complex trajectory. In addition, we generated another variant, referred to as complex trajectory 2, where the camera first moves outward and then moves back inward.

Our system utilizes monocular video depth estimation to construct a pseudo-3D dynamic representation of the scene. This depth map is crucial for maintaining geometric consistency during novel view synthesis, allowing for convincing parallax effects and occlusion handling. By projecting pixels according to their estimated depth values, we achieve realistic scene transformations without explicit 3D reconstruction.

Algorithm 2: I_θ: A sampling step of extended ViBiDSampler for bidirectional interpolation

Function $I_\theta\,(x_t, \sigma_t, c_{start}, c_{end}, x_w)$ **:**

$\hat{x}_{c_{start}} \leftarrow D_\theta(x_t;\, \sigma_t, c_{start})$ `// EDM denoising`

$\bar{x}_{c_{start}} \leftarrow \hat{x}_{c_{start}} \cdot m + x_w \cdot (1-m)$

$x_{t-1,c_{start}} \leftarrow \bar{x}_{c_{start}} + \frac{\sigma_{t-1}}{\sigma_t}(x_t - \hat{x}_\emptyset)$

$(x_{t,c_{start}}) \leftarrow x_{t-1,c_{start}} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\,\epsilon$ `// Re-noise`

$(x'_{t,c_{start}}) \leftarrow \text{flip}(x_{t,c_{start}})$ `// Time reverse`

$\hat{x}'_{c_{end}} \leftarrow D_\theta(x'_{t,c_{start}};\, \sigma_t, c_{end})$ `// EDM denoising`

$\bar{x}'_{c_{end}} \leftarrow \hat{x}'_{c_{end}} \cdot m + x_w \cdot (1-m)$

$x'_{t-1} \leftarrow \bar{x}'_{c_{end}} + \frac{\sigma_{t-1}}{\sigma_t}(x'_t - \hat{x}'_\emptyset)$

$x'_{t-1} \leftarrow \text{flip}(x'_{t-1})$ `// Time reverse`

return x_{t-1}

Algorithm 3: Novel view synthesis and end-view video generation algorithm from [15]

Input: Warped frames x_w, opacity mask m

Output: Input video x_0

$x_T \sim \mathcal{N}(0, 1)$

for $t \leftarrow T$ **to** 1 **do**

  if $t > T - T^{guide}$ **then**

    for $r \leftarrow 1$ **to** R **do**

      $\hat{x}_0 \leftarrow \text{Predict}(x_t)$

      if $r \leq R^{guide}$ **then**

        $\hat{x}_0 \leftarrow D_\theta(x_t;\, \sigma_t, c_{x_0})$

        $\bar{x}_0 \leftarrow \hat{x}_0 \cdot m + x_w \cdot (1-m)$

      else

        $\bar{x}_0 \leftarrow \hat{x}_0$

      end

      $x_{t-1} \leftarrow \bar{x}_0 + \frac{\sigma_{t-1}}{\sigma_t}(x_t - \hat{x}_0)$

      if $r < R$ **then**

        $x_t \sim \mathcal{N}(\bar{x}_0, \sigma_t)$

      end

    end

  else

    $\hat{x}_{t-1} \leftarrow D_\theta(x_t;\, \sigma_t, c_{x_0})$

    $x_{t-1} \leftarrow \bar{x}_0 + \frac{\sigma_{t-1}}{\sigma_t}(x_t - \hat{x}_0)$

  end

end

return x_0

A.4. Details of Zero4D Implementation

Details of interpolation. To generate globally consistent 4D videos, we adapt the interpolation strategy during spatio-temporal video generation. Specifically, we leverage ViBiDSampler [35] as the interpolator I_θ. ViBiDSampler is a state-of-the-art training-free video interpolation method designed for image-to-video diffusion models. Given two conditioning frames, it alternates denoising along the temporal axis to synthesize intermediate frames. In our framework, we extend this process by incorporating warped-frame guidance (see Algorithm 2), which provides additional geometric cues. This modification refines the interpolation process, leading to more faithful structure preservation and improved global spatio-temporal coherence across the generated 4D video grid.

Novle-view synthesis. Algorithm 3 outlines the process for generating novel-view videos from a single monocular video. We first apply novel view synthesis to the initial frame using an I2V diffusion model [5] to produce the novel view $x[:, 1]$. For this,

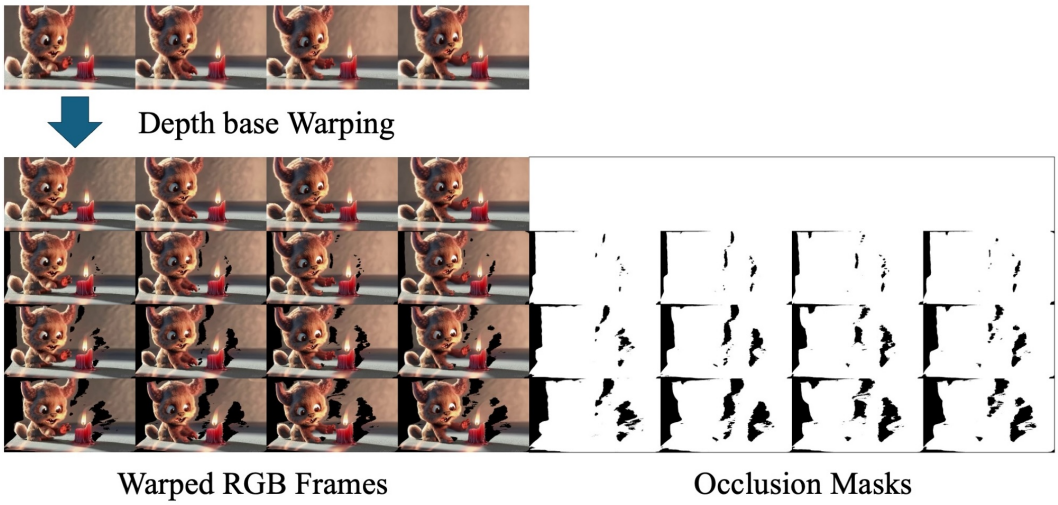

Figure 6. **Input Video Warping.** Given a single video, we utilize an off-the-shelf depth estimation model to generate warped frames from novel viewpoints.

depth-based warping priors from the input video are incorporated to enable inpainting-based synthesis. Specifically, using an off-the-shelf depth estimation model [10], we warp the original frame to novel viewpoints, as illustrated in Figure 5. As shown in Fig. 6, occluded regions from the warp operation appear black, allowing us to extract an opacity mask. Inspired by [15, 17, 36], we adopt a mask inpainting approach, where inpainting is performed on the estimated noisy frame $\hat{x}_0[:, 1]$. Rather than applying inpainting at every denoising step, as in [15], we utilize a re-noising process within the diffusion model's denoising step to refine the final synthesis by reducing artifacts and enhancing structural coherence. A detailed description is provided in Algorithm 3.

A.5. Additional Results

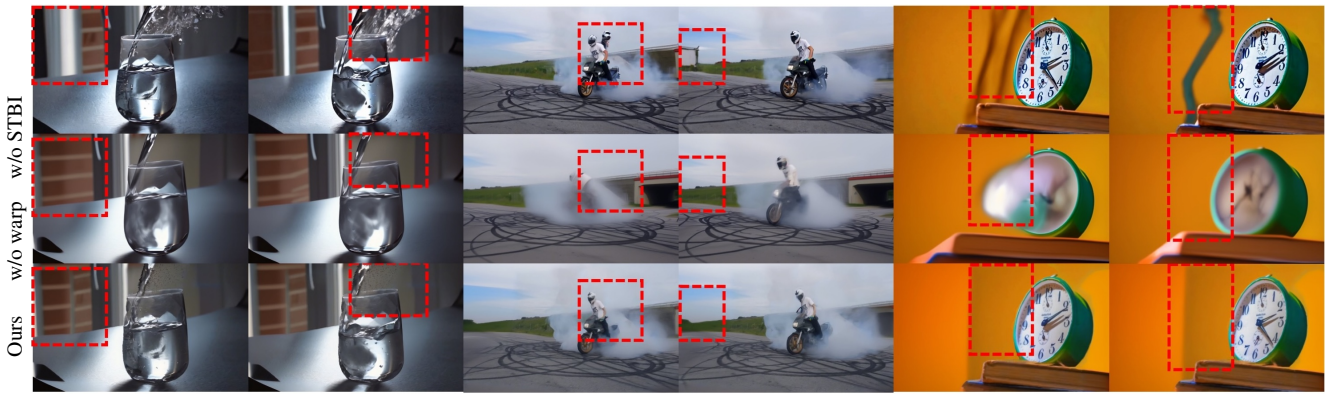

Figure 7. **Ablation results.** Removing spatio-temporal bidirectional interpolation (STBI) or warping guidance leads to broken consistency and geometric artifacts (red boxes). In contrast, our full method preserves spatial structure and temporal coherence across views.

Ablation (detailed analysis). Figure 7 qualitatively illustrates the role of each component in maintaining global consistency. Without spatio-temporal bidirectional interpolation (STBI), each frame is synthesized independently, which causes temporal flickering and background inconsistencies across views. For example, in the water-pouring sequence (left), the liquid surface fails to remain temporally stable, as highlighted by the red boxes. Similarly, without warping guidance, the model struggles with geometric alignment. In the motorcycle example (middle), artifacts appear in the generated human figure, leading to distorted or incomplete shapes. Finally, in the clock sequence (right), the absence of warping or spatio-temporal interpolation leads to visible structural mismatches and background inconsistencies. In contrast, our full model effectively aggregates global information through STBI and enforces geometric consistency via warped-frame guidance, resulting in coherent and high-quality multi-view videos across both spatial and temporal dimensions.

Table 6. **Quantitative results under low-light and camera-motion blur.** We evaluate our method in challenging scenarios where the depth estimation model may fail. We perturb the input videos using motion-blur and low-light filters. The top three rows show bullet-time videos(*), while the bottom three rows present novel-view videos.

Method	Subject Consistency ↑	Background Consistency ↑	Temporal Flickering ↑	Motion Smoothness ↑	Dynamic Degree ↓	Image Quality ↑	Aesthetic Quality ↑
Ours (Low-light)*	95.03%	**95.41%**	**98.80%**	**99.31%**	2.00%	34.12%	33.46%
Ours (Motion blur)*	94.62%	92.88%	94.28%	94.32%	**2.22%**	37.75%	28.81
Ours *	**95.73%**	94.81%	96.88%	98.76%	1.00%	**38.81%**	**38.14%**
Ours (Low-light)	94.28%	96.03%	95.98%	**99.22%**	30.13%	36.11%	33.76%
Ours (Motion blur)	94.76%	94.68%	95.49%	95.99%	32.35%	39.11%	29.21
Ours	**95.55%**	**95.75%**	**97.48%**	98.34%	**27.50%**	**51.12%**	**38.22%**

Quantitative comparison under difficult scenarios. We evaluated our method under difficult scenarios (low-light, blurred camera) to evaluate the robustness of our method. We evaluated the robustness of our proposed method under challenging scenarios, including low-light conditions and camera motion-blurred settings. We use 50 WebVid-10M videos with a low-light filter(30% brightness) and a motion blur filter (20px, 45°). As shown in Table 6, despite minor drops in aesthetic quality, performance remains stable, especially in bullet-time, where static objects are well preserved.