

## 1 A Proof of Lemma 1

2 We introduce Lemma 1 to demonstrate the necessity of using base video as additional input, which is

3 **Lemma 1.** *Let  $x_A(t), x_B(t) \in \mathbb{R}^3$  be the positions of two 3D points at time  $t \in [0, T]$ , defined by a*  
 4 *shared time-independent motion coefficient vector  $\beta \in \mathbb{R}^K$ , and a set of time-dependent, spatially*  
 5 *smooth motion bases  $\{M_k(t)\}_{k=1}^K$ .*

6 *Assume the following:*

7 (i) *The scene is trained over two segments: an edited clip  $[0, t_1]$  without supervision, and an*  
 8 *original clip  $[t_1, t_2]$  with ground-truth geometry supervision;*

9 (ii) *A temporal smoothness regularization is applied on the motion bases  $M_k(t)$ , i.e.,*  
 10  $\sum_{k,t} \|M_k(t+1) - M_k(t)\|^2$ ;

11 (iii) *Each point's position at time  $t$  is given by a transformation composed from the bases and*  
 12 *coefficients:*

$$x(t) = \sum_{k=1}^K \beta_k \cdot M_k(t)(x_0),$$

13 where  $x_0$  is the canonical position.

14 Then, the variance of the pairwise distance  $d(t) = \|x_A(t) - x_B(t)\|$  within the edited clip satisfies:

$$\sigma_2 := \text{Var}_{t \in [0, t_1]}(d(t)) < \sigma_0,$$

15 where  $\sigma_0$  is the variance obtained by training only on the edited clip without supervision.

16 We can prove this lemma using

17 *Proof.* Let  $x_A^0, x_B^0 \in \mathbb{R}^3$  be fixed canonical positions of two points. At time  $t$ , their positions in  
 18 world coordinates are given by:

$$x_i(t) = \sum_{k=1}^K \beta_k \cdot M_k(t)(x_i^0), \quad \text{for } i \in \{A, B\},$$

19 where  $\beta \in \mathbb{R}^K$  is shared and fixed, and  $M_k(t) : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  are time-varying motion basis functions.

20 Define the pairwise squared distance:

$$d^2(t) = \|x_A(t) - x_B(t)\|^2 = \left\| \sum_{k=1}^K \beta_k \cdot (M_k(t)(x_A^0) - M_k(t)(x_B^0)) \right\|^2.$$

21 Let  $\Delta_k(t) := M_k(t)(x_A^0) - M_k(t)(x_B^0) \in \mathbb{R}^3$ , then:

$$d^2(t) = \left\| \sum_{k=1}^K \beta_k \cdot \Delta_k(t) \right\|^2.$$

22 This is a quadratic form in the temporal functions  $\Delta_k(t)$ , linearly combined by fixed weights  $\beta_k$ . Its  
 23 time-variance depends on the temporal variability of  $\Delta_k(t)$ .

24 Now, consider the regularized training setup:

25 (i) The motion bases  $\{M_k(t)\}$  are supervised only in  $[t_1, t_2]$ , enforcing accurate deformation  
 26 there;

27 (ii) A temporal smoothness regularization is imposed:

$$\mathcal{L}_{\text{smooth}} = \sum_{k=1}^K \sum_t \|M_k(t+1) - M_k(t)\|^2.$$

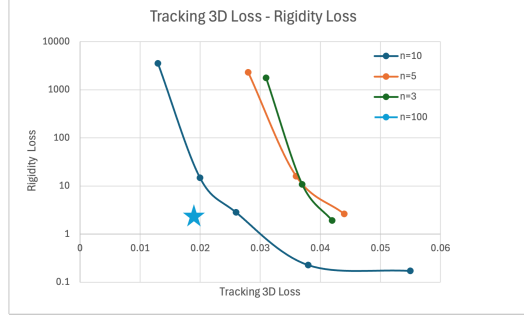


Figure 1: Effect of changing the number of bases.

Due to this regularization, each  $M_k(t)$  evolves smoothly over time. Let us denote the discrete temporal second difference of  $\Delta_k(t)$  as:

$$\delta_k(t) := \Delta_k(t+1) - \Delta_k(t).$$

Then, smoothness implies that  $\|\delta_k(t)\|$  is small, especially near the boundary  $t = t_1$ , where  $M_k(t)$  is influenced by supervision at  $t > t_1$ . Since the functions  $\Delta_k(t)$  are more constrained in this case (compared to unregularized training), their fluctuations in  $[0, t_1]$  are suppressed.

Let us define:

$$f(t) := \sum_{k=1}^K \beta_k \cdot \Delta_k(t), \quad \text{then} \quad d^2(t) = \|f(t)\|^2.$$

The temporal variance of  $d(t)$  satisfies:

$$\text{Var}_{t \in [0, t_1]}(d(t)) = \text{Var}_{t \in [0, t_1]}(\|f(t)\|).$$

Since  $f(t)$  is a fixed linear combination of smoother functions  $\{\Delta_k(t)\}$ , its temporal fluctuation is reduced by the regularization. That is:

$$\text{Var}_{t \in [0, t_1]}(\|f(t)\|) \downarrow \quad \text{as} \quad \mathcal{L}_{\text{smooth}} \downarrow.$$

In contrast, training on  $[0, t_1]$  alone without regularization leads to unconstrained  $M_k(t)$ , and thus larger temporal variability in  $f(t)$ . Therefore:

$$\sigma_2 := \text{Var}_{t \in [0, t_1]}(d(t)) < \sigma_0,$$

where  $\sigma_0$  is the variance from training without supervision or smoothness.

□

## B Training Details

### B.1 2D Priors

To enable 4D reconstruction from a monocular video, we leverage a set of 2D foundation models to provide reliable initialization. Following the pipeline of Shape-of-Motion, we use Segment Anything for foreground segmentation and Track Anything to propagate the masks across frames. For camera motion and point correspondence, we adopt Tapir for 2D point tracking and MegaSAM for estimating camera poses. Additionally, we incorporate dense depth maps from VideoDepthAnything to provide richer geometric details for the scene.

### B.2 Hyperparameters

To achieve geometry-preserving reconstruction while maintaining high-frequency motion details, we found it essential to carefully tune the rigidity loss. Since the overall motion is primarily determined during the motion initialization stage, we evaluate the effect of different combinations of rigidity loss weights and the number of motion bases.

54 As shown in Figure 1, each curve represents the trade-off between object rigidity and tracking  
55 accuracy for a given number of motion bases, as we vary the rigidity loss weight from 0 to  $10^{-1}$ . Our  
56 results show that increasing the number of motion bases can improve this trade-off curve, and that a  
57 loss weight of  $10^{-3}$  offers a good balance. Based on this analysis, we use 100 motion bases and set  
58 the rigidity loss weight to  $10^{-3}$  in all experiments.

## 59 C Samples from DAVIS Dataset

60 To qualitatively evaluate our 4D restaging pipeline, we present several representative samples from  
61 the DAVIS dataset in Figure 2. For each sequence, we show: (1) two keyframes from the base  
62 video (original monocular input); (2) the corresponding motion prompt describing the desired new  
63 motion; and (3) two keyframes from the generated driving video, synthesized using a pre-trained  
64 image-to-video model (*e.g.*, Sora) conditioned on the prompt and the first frame. These examples  
65 demonstrate the diversity of motion prompts supported by our pipeline, including changes in direction,  
66 speed, and posture. They also highlight the realism and coherence of the generated driving video,  
67 which serves as the motion source for our 4D reconstruction. For each example, we visualize the base  
68 and drive videos using two keyframes (top and bottom) to illustrate the overall temporal dynamics.

## 69 D More Results

70 To demonstrate generalization of the pipeline, we also provide visualization on some web-collected  
71 sequences shown in Figure 3.

## 72 E Video Demo

73 Please refer to the video demo in the attachment.






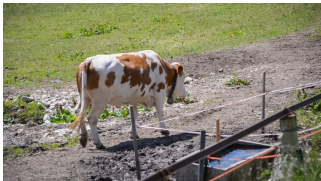

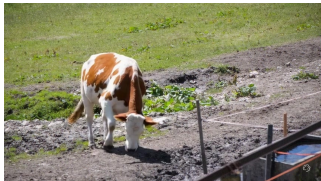




Base Video	Prompt	Drive Video
 	<p>"A small silver car slowly drives up onto the sidewalk from the street in a quiet urban area. The camera stays fixed in position but smoothly pans to follow the vehicle's motion as it mounts the curb and continues onto the pedestrian pavement. Surrounding buildings are modern and light-colored, with crosswalk lines, bollards, and a few pedestrians in the background. The motion feels slightly unusual yet calm, with no abrupt changes in camera movement."</p>	 
 	<p>"A brown and white cow slowly turns its body and walks toward the grassy field to graze. The camera remains fixed in physical position but gently pans to follow the cow's movement as it turns and begins eating grass. The scene takes place in a quiet rural pasture under bright daylight, with patches of soil, stones, and fencing visible in the foreground. The motion is natural and unhurried, with the cow staying centered in the frame."</p>	 
 	<p>"A light-colored dog wearing a harness suddenly dashes into the nearby bushes on the right side of the frame. The camera remains completely still, capturing the dog's quick movement as it disappears into the foliage. The setting is a dry grassy area with sparse leaves, and a fence and road are visible in the background. The scene feels spontaneous and natural, with no camera tracking or zooming."</p>	 

Figure 2: Examples of motion prompts and corresponding base/drive videos. Each column shows the input video, the motion prompt, and the generated video. Each video is visualized using two keyframes (top and bottom).

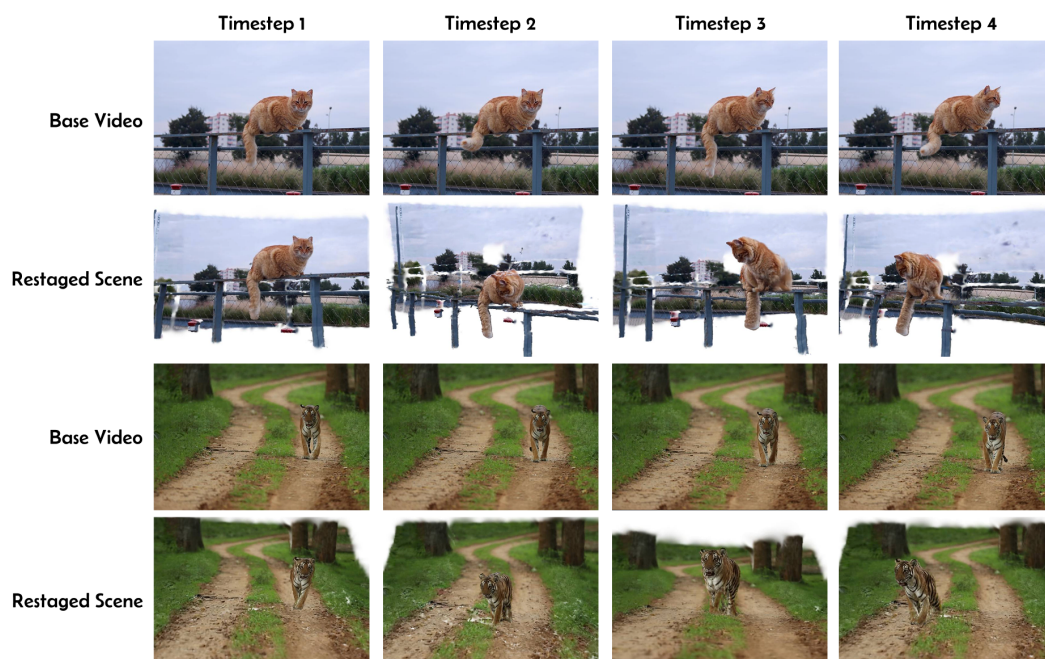


Figure 3: Some samples of 4D restaging on in-the-wild video, which are collected from Internet.