

# Zero-shot Customized Video Editing with Diffusion Feature Transfer

## Supplementary Material

In this supplementary material, we provide more experimental setting details, and results.

### 6. Experimental Settings

**Diffusion Feature Visualization.** In Section 3, to visualize those high-dimensional SD features and understand how semantic spatial information is internally encoded within diffusion features, we use Principal Component Analysis (PCA) to interpret the dominant visual properties within these high-dimensional features. Specifically, we select the denoising step  $t = 401$  and extract features  $\mathbf{O}$  from each block of the UNet decoder. We represent these features using the first three leading components from PCA, and visualize them as RGB images.

**Diffusion Feature Injection.** The Target Branch in Fig. 5 starts from the same DDIM inverted noise as the Source Branch. In the denoising steps, the shallow features from the Shallow Block of the Source Branch, and the deep features from the Intermediate Block and Deep Block of the Reference Branch are injected into the Target Branch. We inject features from both branches during the first 60% denoising steps, *i.e.*,  $t_{\max} = (1 - 0.6)T$ , where  $T$  is the number of total denoising time steps. Finally, the Target Branch produces a synthesized video.

**Evaluation Metrics.** We employ the adapted CLIP-Score [18] to evaluate text alignment and temporal consistency. Specifically, we calculate the CLIP similarity between each frame of the synthesized video and the target text prompt, then report the average CLIP similarity across all frames. For temporal consistency, we assess the CLIP similarity between consecutive frames of the synthesized video, and calculate the average value. To evaluate concept alignment, we follow Custom Diffusion [25] to measure the CLIP similarity between each frame of the synthesized video and all reference images, and report the average score across all video frames.

**Details of Masks.** In our experiments, the DAVIS dataset contains masks for each video. For videos without masks, we use the TRACER [26] to extract masks. In practice, any off-the-shelf segmentation tool, such as the Grounded-SAM-2 [28, 37], can be used to produce image and video masks.

Table 4. Human preferences.

Win Rate	Motion fidelity	Temporal consistency
Ours vs. MotionDirector	<b>93%</b> vs. 7%	<b>69%</b> vs. 31%
Ours vs. VideoSwap	<b>56%</b> vs. 44%	30% vs. <b>70%</b>

### 7. Experiment Results

#### 7.1. Human Preferences

We provide human evaluation results on synthesized videos in Tab. 4. Specifically, we visually evaluate the temporal consistency and motion fidelity of the videos generated by different methods, where motion fidelity means whether the motion of object in the synthesized video is the same as that of the object in the source video. Tab. 4 presents the win rates of our FreeMix against VideoSwap and MotionDirector. The results demonstrate that our method outperforms both approaches in motion fidelity and exhibits higher temporal consistency compared to MotionDirector.

#### 7.2. Ablation Study

**Different Feature Injection Combinations.** We present various feature injection strategies in Tab. 5. Specifically, we evaluate various injection strategies by testing different combinations of injecting source video or reference image features into shallow or deep UNet blocks: (1) Injecting features from the source video solely into either the Shallow Block or Deep Block; (2) Injecting features from the reference images solely into either the Shallow Block or Deep Block; (3) Injecting features from both the source video and reference images into the Shallow Block; (4) Injecting features from both the source video and reference images into the Deep Block; (5) Injecting features from the source video into the Deep Block, and injecting features from the reference images into the Shallow Block; (6) Injecting features from the source video into the Shallow Block, and injecting features from the reference images into the Deep Block. The results align with our observations: injecting source video features into shallow layers ensures good temporal consistency, while injecting both source video and reference image features into shallow layers compromises temporal consistency. Injecting reference image features into either shallow or deep layers achieves good concept alignment, but mixing source video and reference image features reduces concept alignment. Injecting source video features into shallow layers and reference image features into deeper layers, as configured in our paper, achieves both strong concept alignment and temporal consistency.

Table 5. Different feature injection combinations.

	Src(shallow)	Src(deep)	Ref(shallow)	Ref(deep)
Concept Alignment	71.44	71.48	75.17	75.88
Temporal Consistency	95.67	90.71	87.53	90.28

---

	Src(s) Ref(s)	Src(d) Ref(d)	Src(d) Ref(s)	Src(s) Ref(d)
Concept Alignment	73.43	71.8	70.66	<b>75.79</b>
Temporal Consistency	92.77	92.1	91.37	<b>95.71</b>

Table 6. Number of reference images.

Number of References	1	2	3	4	5
Concept Alignment	74.97	74.78	75.17	<b>75.92</b>	75.79
Temporal Consistency	93.76	94.35	95.11	95.44	<b>95.71</b>

Table 7. Influence of timesteps.

$t_{\max}$	$0.2T$	$0.4T$	$0.6T$	$0.8T$	$T$
Concept Alignment	74.29	75.6	<b>75.79</b>	75.73	74.23
Temporal Consistency	93.63	93.87	<b>95.71</b>	94.73	94.41

Table 8. GPU memory consumption and running time.

	FateZero	CCEdit	ControlVideo	VideoSwap	Ours
GPU (GB)	24	26	10	16	<b>8</b>
Time (s)	246	137	73	144	<b>50</b>

**Number of Reference Images.** Tab. 6 illustrates the impact of the number of reference images on temporal consistency and concept alignment. Using only 1-2 reference images results in lower scores, while using 4-5 reference images significantly improves both temporal consistency and concept alignment. Based on our experiments, we use 5 reference images.

**Influence of Timesteps.** We compare different values of  $t_{\max}$ , the stopping timestep for feature injection, ranging from  $0.2T$  to  $T$ , where  $T$  denotes the total number of timesteps. The results are shown in Tab. 7. Based on our experiments, we select  $t_{\max} = 0.6T$ , as it achieves the best concept alignment and temporal consistency.

**Computational Complexity.** We present the GPU memory requirements and video generation times for different methods in Tab. 8. Among all methods, our approach requires the least memory and achieves the fastest speed. In our implementation, OT is computed once for each video and takes approximately 5 seconds.

**Additional Evaluation Metrics.** Tab. 9 extends the comparison with baseline methods using additional evaluation metrics, complementing the primary results presented in Tabs. 1 and 2. We evaluate image quality using three standard metrics. Peak Signal-to-Noise Ratio (PSNR) quantifies reconstruction fidelity based on the pixel-wise mean squared error. The Structural Similarity Index Measure

Table 9. Video quality and consistency with additional metrics. PSNR, SSIM, LPIPS are metrics for video equality, Warp error is the metric for temporal consistency.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS (AlexNet) $\downarrow$	LPIPS (Vgg) $\downarrow$	Warp Error $\downarrow$
FateZero	15.88	0.58	0.38	0.42	0.08
FLATTEN	16.23	0.53	0.28	0.36	0.11
TokenFlow	18.46	0.6	0.33	0.39	0.05
ControlVideo	9.7	0.24	0.74	0.73	0.33
CCEdit	13.94	0.32	0.52	0.56	0.12
FreeMix (ours)	<b>19.87</b>	<b>0.68</b>	<b>0.19</b>	<b>0.28</b>	<b>0.03</b>

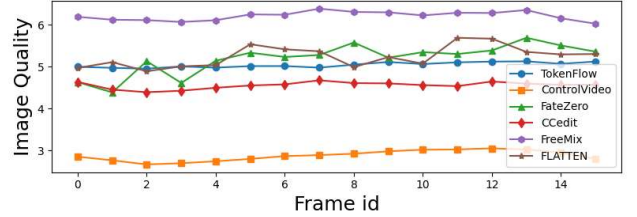


Figure 9. Image quality vs. each frame

(SSIM) [49] provides a perceptually-motivated assessment by comparing local patterns of luminance, contrast, and structure. Finally, the Learned Perceptual Image Patch Similarity (LPIPS) [55] leverages a pre-trained deep neural network to compute the distance between image patches in a feature space, which correlates closely with human perceptual judgment. We also use warp error to quantify the temporal consistency, which measures the geometric misalignment between a point’s true position in a target image and its predicted position after being transformed from a source image by an estimated warp field. FreeMix outperforms baseline methods with better video quality and temporal consistency.

**Evaluate Image Quality Over Time.** To assess temporal quality consistency, we evaluate each video frame using the pre-trained NIMA model [42], which predicts perceptual image quality. As illustrated in Fig. 9, the per-frame analysis demonstrates that FreeMix consistently outperforms all baseline methods in terms of overall image quality.

### 7.3. Qualitative Results

We show more qualitative results of FreeMix in Figs 10-13. As can be seen from those results, FreeMix successfully transfers the motion of the object from the source video, and the appearance of the object in the reference images, to the synthesized video. It is worth noting that FreeMix is flexible to deal with source video with different aspect ratios.

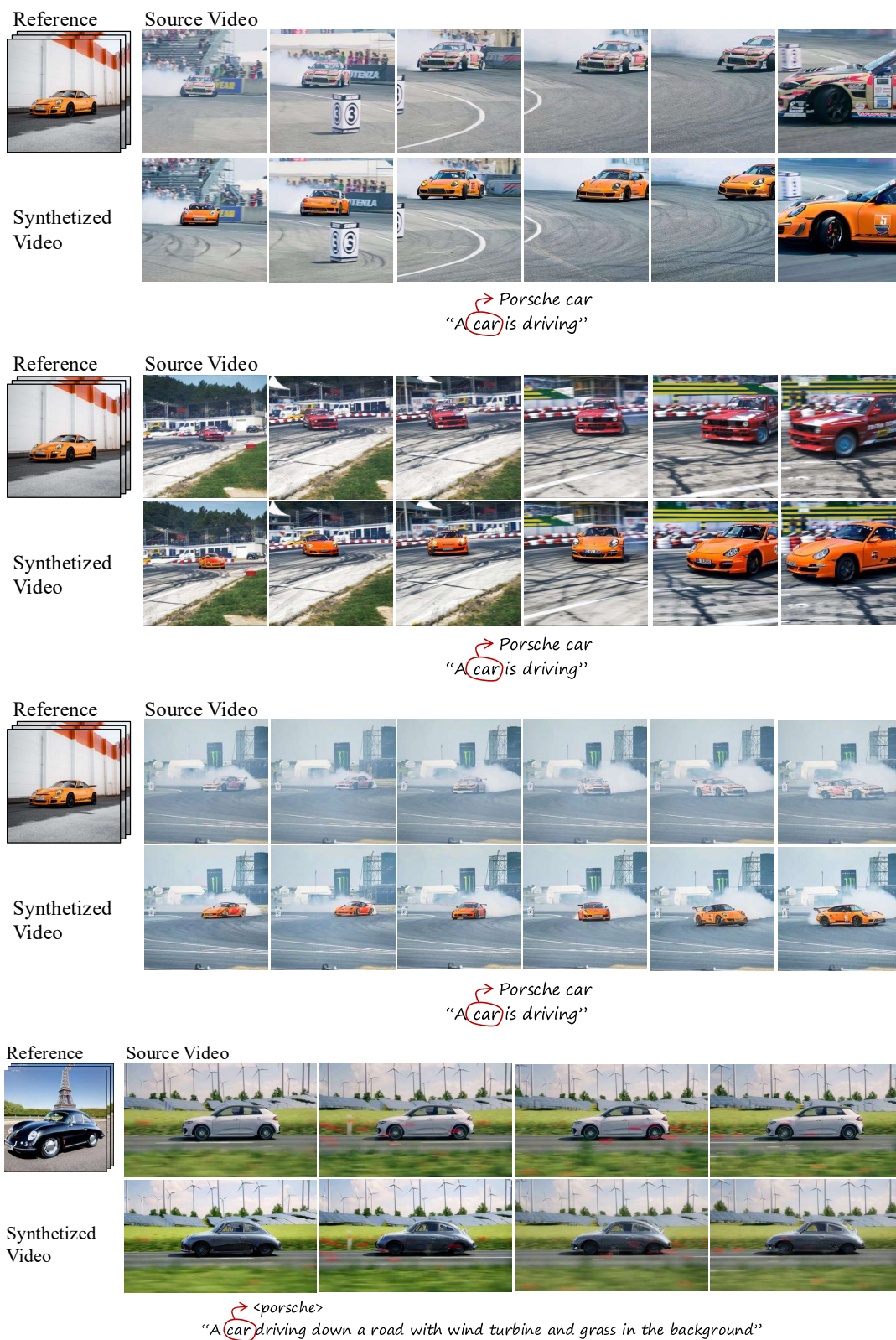


Figure 10. Sample results of FreeMix.



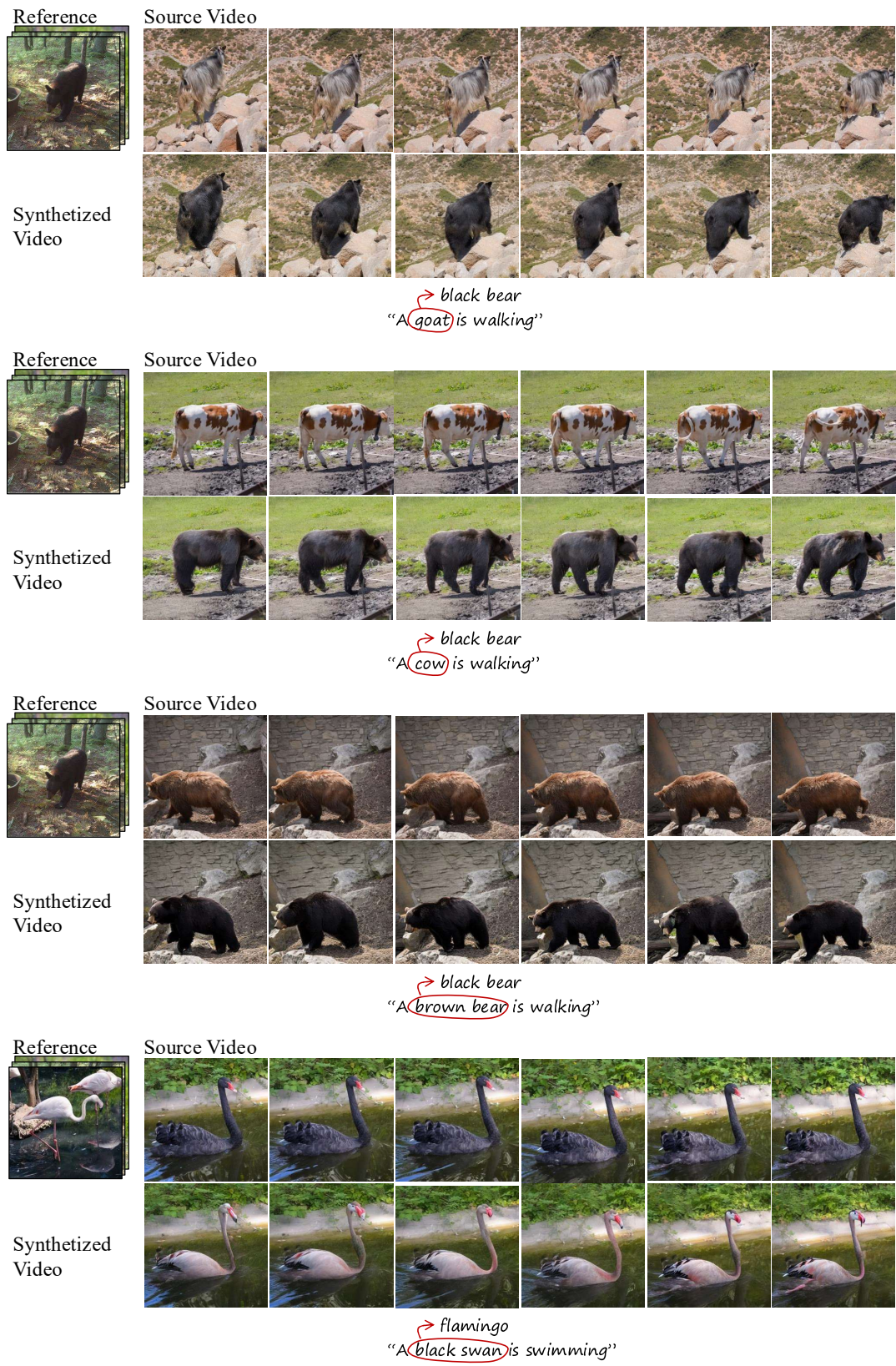


Figure 11. Sample results of FreeMix.

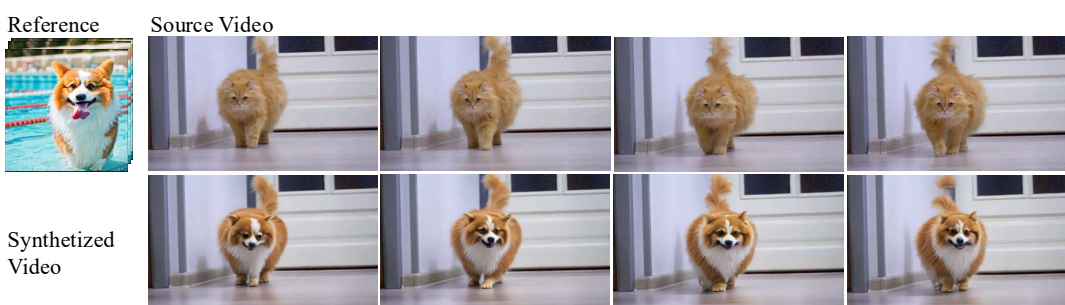


Figure 12. Sample results of FreeMix.

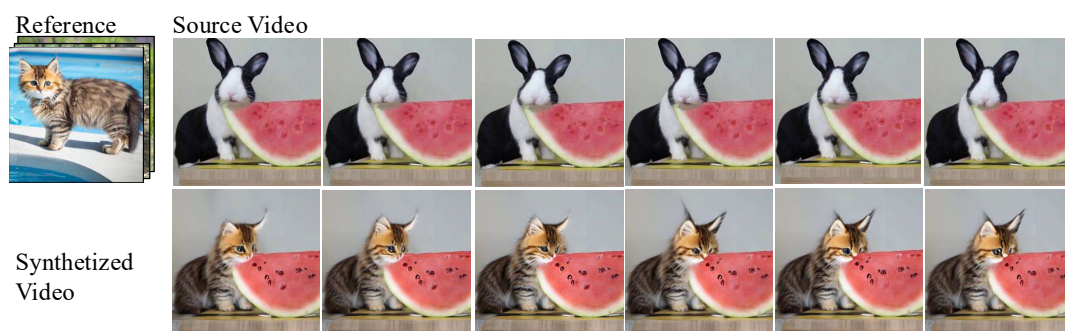




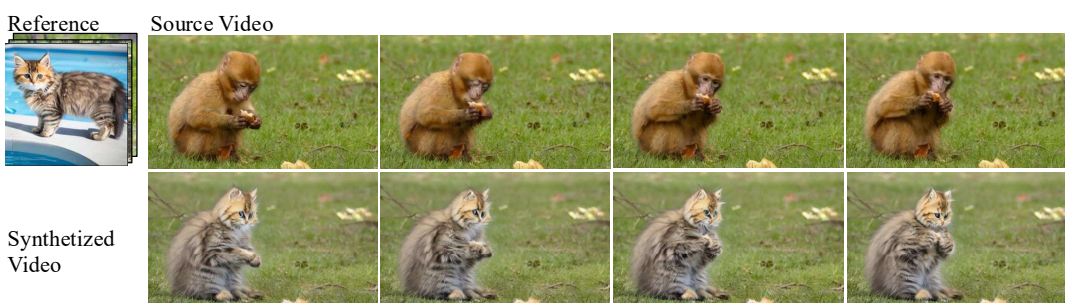
→ <cat>  
 "A cat walking on a piano keyboard near a potted plant and a mirror in a room with a plant"



→ <dog>  
 "A cat is walking on the floor at a room"



→ <cat>  
 "A rabbit eating a piece of watermelon on the table"



→ <cat>  
 "A monkey sitting on the ground eating something in its hands"

Figure 13. Sample results of FreeMix.