

A Video is Worth 256 Bases: Spatial-Temporal Expectation-Maximization Inversion for Zero-Shot Video Editing

Maomao Li¹, Yu Li², Tianyu Yang², Yunfei Liu², Dongxu Yue³, Zhihui Lin⁴, and Dong Xu¹

¹The University of Hong Kong ²International Digital Economy Academy (IDEA)

³Peking University ⁴Tsinghua University



Figure 1. We propose STEM inversion as an alternative approach to zero-shot video editing, which offers several advantages over the commonly employed DDIM inversion technique. STEM inversion achieves superior temporal consistency in video reconstruction while preserving intricate details. Moreover, it seamlessly integrates with contemporary video editing methods, such as TokenFlow (TF) [9] and FateZero (FZ) [30], enhancing their editing capabilities. Best viewed with zoom-in.

Abstract

This paper presents a video inversion approach for zero-shot video editing, which aims to model the input video with low-rank representation during the inversion process. The existing video editing methods usually apply the typical 2D DDIM inversion or naïve spatial-temporal DDIM inversion before editing, which leverages time-varying representation for each frame to derive noisy latent. Unlike most existing approaches, we propose a Spatial-Temporal Expectation-Maximization (STEM) inversion, which formulates the dense video feature under an expectation-maximization manner and iteratively estimates a more compact basis set to represent the whole video. Each frame applies the fixed and global representation for inversion, which is more friendly for temporal consistency during reconstruction and editing. Extensive qualitative and quantitative experiments demonstrate that our STEM inversion can achieve consistent improvement on two state-of-the-art video editing methods. Project page: <https://stem-inv.github.io/page/>.

1. Introduction

Recent years have witnessed a surge of interest in using diffusion models [13, 40] for text-to-image (T2I) generation. There are fruitful endeavours that have been pursued, such as DALLE-2 [33], Imagen [37], and Stable Diffusion [35]. Following this line, a group of methods finetune the general T2I model [35] for personalization usage [5, 8, 22, 36, 47, 50]. In view of amazing results in the image domain, it is natural to leverage a pre-trained large-scale T2I model [35] for video editing. However, frame-wise editing inevitably brings an unacceptable flickering effect. Thus, how to effectively model 3D correspondence in a 2D model poses the main obstacle in the current video editing task.

There are three schools of research for diffusion-based text-driven video editing. 1) Finetune or train addition modules [11, 25] for the image diffusion model on massive videos to learn video motion prior. 2) Tune the image model on each video to be edited to capture the temporal consistency in the input video. 3) Design frame-wise attention mechanisms to capture temporal cues, such as FateZero [30], Pix2Video [3], and TokenFlow [9]. The

last setting requires no training and reaches impressive results, which is more convenient for users and communities to use. Therefore, we focus on the third setting in this paper.

Although DDIM inversion struggles for precise reconstruction when classifier-free guidance [14] is applied, most of the existing zero-shot video editing approaches only focus on making an improvement on the editing process itself to eliminate the limited editing ability brought by inaccurate inversion and reconstruction. For example, Pix2Video [3] and TokenFlow [9] use the typical DDIM image inversion to invert a video without any temporal modeling. Tune-A-Video [46] and FateZero [30] adopt a naïve spatial-temporal DDIM inversion, which inflates the 2D UNet of the Stable Diffusion to incorporate multiple frames for each frame inversion. To achieve acceptable complexity, each frame only explores the spatial-temporal context of two frames in total, providing rough temporal modeling.

In contrast to previous methods, this paper aims at an accurate inversion for better reconstruction and editing ability. Specifically, we propose an efficient video inversion method, dubbed **Spatial-Temporal Expectation-Maximization** (STEM) inversion. Instead of regarding all pixels in a video as the reconstruction bases [24], we leverage the EM [4] algorithm to find a more compact basis set (e.g., 256 bases). Here, we treat the low-rank bases as the parameter to learn and the responsibility of each base as the latent variables in the EM algorithm. During each iteration, the E step in the proposed STEM inversion tries to calculate the expectation of latent variables (responsibility) while the M step updates the parameters (bases). The algorithm would achieve convergence by conducting E and M steps alternately for several iterations.

As for using the low-rank representation (bases) for each-frame inversion, we apply the obtained bases to generate the Key and Value embeddings in the Self-attention module, rather than applying the time-varying one frame or two frames in the sequence like previous techniques [3, 9, 30, 46]. As a result, the computational complexity of the Self-attention layer is greatly reduced. Furthermore, we simply replace the DDIM inversion of two zero-shot video editing methods TokenFlow [9] and FateZero [30] with the proposed STEM inversion. As seen in Fig. 1, our method not only achieves more accurate reconstruction but also improves editing performance, e.g., the advertising banner and man’s leg. To sum up, our main contributions are:

- We propose STEM inversion for diffusion-based video editing, which reformulates a dense video under an expectation-maximization iteration manner and evaluates a more compact basis set.
- Our STEM inversion can reduce the complexity and improve the reconstruction quality compared with DDIM inversion for videos even though it explores the global spatial-temporal context.

- Extensive qualitative and quantitative experiments demonstrate that our STEM inversion can improve video editing performance painlessly by replacing DDIM inversion in the existing video editing pipeline.

2. Related Work

Text-dirven image generation and editing. Image generation is a fundamental task in computer vision. There have been many studies on image generation based on GANs [10, 17], VAE [21, 42], auto-regressive Transformers [6, 7, 34, 45, 49], and flow [20]. Recently, diffusion models [13, 40] have merged as a popular choice for text-to-image (T2I) generation since their stunning results. DALLE-2 [33] and Imagen [37] can generate realistic images by conditioning in text embeddings from large language models [31, 32], where cascaded diffusion models are used to scale up image resolution gradually. Stable Diffusion [35] applies diffusion models in the latent space of powerful pre-trained auto-encoders in VQ-GAN [6], achieving training and inference on limited resources.

For text-driven image editing, ControlNet [51] enhances large pre-trained text-to-image diffusion models with task-specific image conditions, such as depth map, canny edge [2], and human pose. Besides, a number of studies explore text-based interfaces for content manipulation [1, 53], style transfer [23, 52], and generator domain adaption [19].

Text-dirven video generation and editing. Text-to-video (T2V) turns to generate corresponding videos based on the given prompt. Imagen-Video [12] proposes a cascaded video diffusion model that sequentially increases the resolution of generated videos via a series of super-resolution diffusion models. Make-A-Video [39] proposes to fine-tune the T2I model for video generation, where a spatiotemporally factorized diffusion model is designed.

In terms of text-driven video editing, Tune-A-Video [46] proposes to conduct one-shot video tuning for videos, while inflating 2D-UNet and replacing self-attention as sparse causal attention for temporal modeling. Fatezero [30] and Pix2Video [3] apply attention maps to model temporal consistency, which can be regarded as an implicit manipulation way. Besides, Rerender-A-Video [48] and Video-ControlNet [15] use optical flow for frame-wise constraints. Further, several methods [26, 38, 43] have been aware that the inaccurate DDIM inversion would lead to accumulated errors during the denoising process, and thus turn to Null-text inversion [28]. However, this optimization-based inversion takes 2 minutes per frame using an A100 GPU, which is unfavourable in real applications.

In contrast to previous methods, we propose to overhaul DDIM inversion into STEM inversion, where we reformulate the input video representation as a low-rank basis set by the EM algorithm, which harvests a fixed spatial-temporal context from the entire video for each frame, rather than

exploring time-varying one like existing DDIM does. Besides, the row-rank representation reduces the complexity in self-attention layers.

3. Preliminary

3.1. Stable Diffusion and DDIM Inversion

Stable Diffusion [35] (SD) applies diffusion models in the latent space of powerful autoencoders, where an encoder \mathcal{E} learns to encode the image x into latent representations $\mathbf{z} = \mathcal{E}(x)$, and a decoder \mathcal{D} take the latent back to the pixel space $\mathcal{D}(\mathcal{E}(x)) \approx x$. Then, in the forward process, the model iteratively adds noise to the latent code \mathbf{z}_0 , resulting in a perturbed \mathbf{z}_t :

$$q(\mathbf{z}_t | \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta_t} \mathbf{z}_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

where β_t denotes a hyperparameter chosen ahead of model training. In the backward process, a UNet ϵ_θ is trained to predict the noise with L_2 objective, where the architecture consists of residual, self-attention, and cross-attention blocks. Once trained, \mathbf{z}_0 can be sampled with the deterministic DDIM sampling [40]:

$$\mathbf{z}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \mathbf{z}_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \epsilon_\theta(\mathbf{z}_t, t, \mathcal{C}), \quad (2)$$

where $\alpha_t = \prod_{i=1}^t (1 - \beta_i)$ and \mathcal{C} is the embedding of the text condition.

DDIM Inversion aims to map an image into a known latent space (*i.e.*, the domain of model output) before reconstruction or editing. The deterministic DDIM inversion can be formulated as:

$$\mathbf{z}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} \mathbf{z}_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \epsilon_\theta(\mathbf{z}_t, t, \mathcal{C}), \quad (3)$$

which is based on the assumption that the ODE process can be reversed in the limit of small steps. It works well for unconditional generation, while amplifies errors when classifier-free guidance [14] is applied. Thus, it struggles for accurate reconstruction for text-driven editing [28].

3.2. Expectation-Maximization Algorithm

Formally, we denote \mathbf{X} as the observed date and \mathbf{Z} as the corresponding unobservable latent variables, where $\{\mathbf{X}, \mathbf{Z}\}$ represents the complete data. As an iterative method, the expectation-maximization (EM) [4] algorithm estimates the model parameter θ by finding the maximum likelihood of the complete data.

Each iteration in the EM algorithm consists of an expectation step (E step) and a maximization step (M step). Specifically, during r -th iteration, the E step first finds the

posterior distribution of the unobservable variables \mathbf{Z} (*i.e.*, $P(\mathbf{Z}|\mathbf{X}, \theta^{r-1})$), and then calculate the expectation of the complete data likelihood:

$$\mathcal{Q}(\theta, \theta^{r-1}) = \sum_{\mathbf{z}} P(\mathbf{Z}|\mathbf{X}, \theta^{r-1}) \ln P(\mathbf{X}, \mathbf{Z}|\theta). \quad (4)$$

Then M step updates the parameter θ in r -th iteration by maximizing the above likelihood:

$$\theta^r = \arg \max_{\theta} \mathcal{Q}(\theta, \theta^{r-1}). \quad (5)$$

The algorithm carries out the E step and M step for R times to make the model converge.

4. Methodology

In this section, we first revisit existing methods using DDIM inversion for video editing in Sec. 4.1. Then, the overview of the proposed STEM inversion is in Sec. 4.2. Next, we elaborate on the E step and M step in Sec. 4.3 and Sec. 4.4. Finally, we introduce how to apply our STEM inversion for zero-shot video editing in Sec. 4.5.

4.1. Using DDIM inversion For Video Editing

Given a source video sequence $\mathcal{I} = [I^1, \dots, I^N]$ and its corresponding source prompt \mathcal{P}_{src} , a zero-shot text-driven video editing system aims to generate a new video $\mathcal{Y} = [Y^1, \dots, Y^N]$ based on the target prompt \mathcal{P}_{tgt} without any training or finetuning. The edited video should fully reflect the \mathcal{P}_{tgt} while maintaining the temporal consistency.

To conduct video editing, the popular one-shot video editing method [46] and zero-shot ones [3, 9, 30] first conduct DDIM inversion on each frame before the denoising process. Pix2Video [3] and TokenFlow [9] use the typical 2D DDIM inversion, which inverts each frame with its own spatial context. Besides, although pre-trained SD is trained without temporal constraints, FateZero [30] and Tune-A-Video [46] inflate the 2D UNet ϵ_θ and transform self-attention as spatial-temporal self-attention without changing pre-trained SD weights. Concretely, when inverting a frame, the naive spatial-temporal DDIM Inversion would exploit context from its own and another frame in the sequence. The spatial-temporal self-attention process $\text{ATTENTION}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}) \cdot \mathbf{V}$ for the feature \mathbf{X}^n of n -th frame I^n can be described as using of both h -th and l -th frame:

$$\mathbf{Q} = \mathbf{X}^n \mathbf{W}_Q, \mathbf{K} = [\mathbf{X}^h, \mathbf{X}^l] \mathbf{W}_K, \mathbf{V} = [\mathbf{X}^h, \mathbf{X}^l] \mathbf{W}_V, \quad (6)$$

where $n, h, l \in [1, 2, \dots, N]$, $[.]$ indicates concatenation, and d is the channel number of embeddings. Fatezero employs $h = n$ and $l = \text{Round}[\frac{N}{2}]$ while Tune-A-Video adopts $h = 1$ and $l = n-1$. $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are projection metrics and shared across space and time.

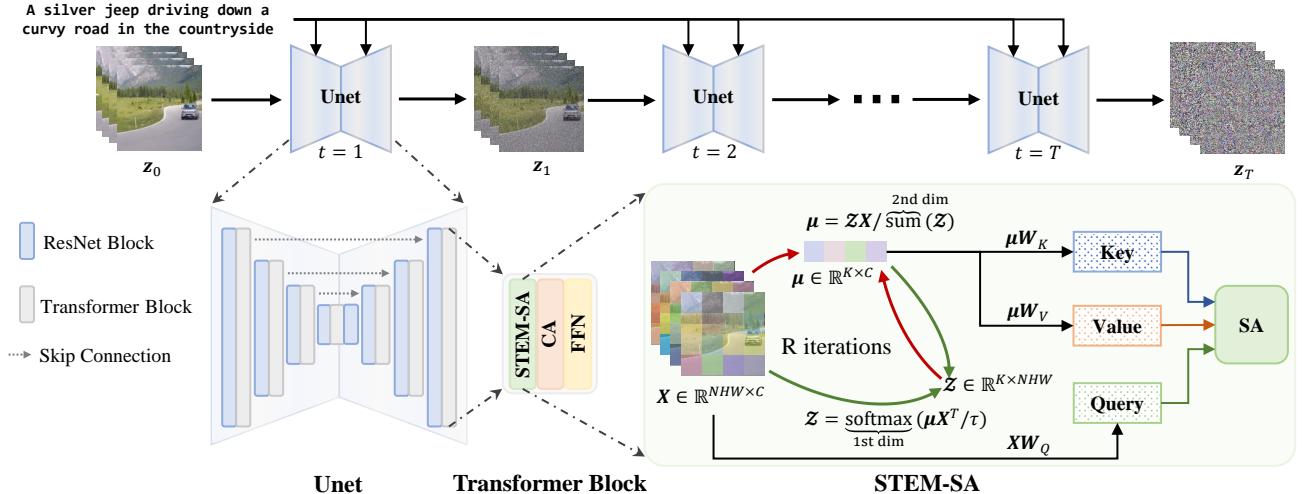


Figure 2. The illustration of the proposed STEM inversion method. We estimate a more compact representation (bases μ) for the input video via the EM algorithm. The ST-E step and ST-M step are executed alternately for R times until convergence. The Self-attention (SA) in our STEM inversion are denoted as STEM-SA, where the Key and Value embeddings are derived by projections of the converged μ .

4.2. The Overview of STEM Inversion

DDIM inversion provides initial noise for video reconstruction or editing, whose performance would directly influence final results. For the typical 2D DDIM inversion, each frame only considers its own spatial context without any temporal consideration, which poses a challenge to video reconstruction and even editing. For naive spatial-temporal DDIM inversion, each frame uses a two-frame context in the sequence to help counter the instability of temporal modeling. However, the benefit is limited.

In contrast, we argue that frames over a larger range should be considered to execute DDIM inversion. However, using all video frames directly will bring unacceptable complexity. To deal with this, as shown in Fig. 2, we propose a Spatial-Temporal Expectation-Maximization (STEM) inversion method. The insight behind this is massive intra-frame and inter-frame redundancy lie in a video, thus there is no need to treat every pixel in the video as reconstruction bases [24]. Then, we use the EM algorithm to find a more compact basis set, which can be regarded as a general and low-rank representation for the entire video.

Formally, we denote the feature map of the input video is $\mathbf{X} = [\mathbf{X}^1, \dots, \mathbf{X}^N] \in \mathbb{R}^{N \times H \times W \times C}$, where N, C, H, W represent the frame number, channel number, the height, and width of the feature. Besides, we express the initial basis set as $\mu \in \mathbb{R}^{K \times C}$, where K is the number of bases. For simplicity, we make $M = NHW$. Since $K \ll M$, the proposed compactness is non-trivial. Specifically, we make the bases μ as the parameters to be learned in the EM algorithm. The E step calculates the expectation of latent variables (responsibility) $\mathcal{Z} \in \mathbb{R}^{K \times M}$. The M step maximizes the likelihood of complete data to update bases μ . These

two steps would be executed R times before convergence.

Then, as shown in Fig. 2, we use the converged low-rank bases μ for all self-attention calculations in the pre-trained SD, which is denoted as STEM-SA. Consequently, the way of calculating self-attention in Eq. (6) becomes,

$$\mathbf{Q} = \mathbf{X}^n \mathbf{W}_Q, \mathbf{K} = \mu \mathbf{W}_K, \mathbf{V} = \mu \mathbf{W}_V, \quad (7)$$

where $\mathbf{Q} \in \mathbb{R}^{HW \times C}$, \mathbf{K} and $\mathbf{V} \in \mathbb{R}^{K \times C}$. Note that we perform STEM-SA in all time steps and self-attention layers of UNet. For simplicity, we omit the notations of the timestep t and layer l .

The STEM-SA mechanism in our STEM inversion would produce low-rank features and reduce complexity. For the typical 2D DDIM inversion, each self-attention layer will take $\mathcal{O}(N(HW)^2)$ computation for a N -frame video. The complexity of the self-attention layer in naive spatial-temporal DDIM inversion is $\mathcal{O}(2N(HW)^2)$, which can be considered approximately the same as that of the typical one. In contrast, our STEM inversion reduces the complexity to $\mathcal{O}(NHWK(R + 1))$. Since the iteration number R is a small constant, our complexity is only $\mathcal{O}(NHWK)$. In general, we set $K \ll HW$, and the computation complexity of the proposed STEM Inversion is much less than that of DDIM inversion. Moreover, the higher the resolution of the video, the more computation our STEM can reduce.

4.3. ST-E Step: Responsibility Estimation

The E step in our STEM inversion conducts responsibility estimation. This step computes the expected value of unobservable latent variables $\mathcal{Z} = \{\mathcal{Z}_{km}\}_{k=1, m=1}^{K, M} \in \mathbb{R}^{K \times M}$, where $M = NHW$, which corresponds to the responsibil-

Algorithm 1: Zero-shot Video Editing Pipeline With STEM Inversion

Input:

feature of frames:
 $\mathbf{X} \in \mathbb{R}^{M \times C}$, $M = N \times H \times W$
 source prompt \mathcal{P}_{src} , target prompt \mathcal{P}_{tgt}

Output:

bases $\boldsymbol{\mu} \in \mathbb{R}^{K \times C}$
 the edited video $\mathcal{Y} = [Y^1, \dots, Y^N]$

1 Part I: STEM Inversion

```

/* Setting  $\tau$  as a constant. */ 
2 for  $r = 1$  to  $R$  do
    // ST-E step, estimate
    // responsibilities:
3    $\mathcal{Z}_{km} \leftarrow \frac{\exp(\boldsymbol{\mu}_k \mathbf{X}_m^T / \tau)}{\sum_{j=1}^K \exp(\boldsymbol{\mu}_j \mathbf{X}_m^T / \tau)}$ 
    // ST-M step, update bases:
4    $\boldsymbol{\mu}_k \leftarrow \sum_{i=1}^M \frac{\mathcal{Z}_{ki} \mathbf{X}_i}{\sum_{m=1}^M \mathcal{Z}_{km}}$ 
    //  $m = \{1, 2, \dots, M\}; k = \{1, 2, \dots, K\}$ 
5 Performing STEM-SA in DDIM-Inv (Eq.(3)) with
 $\mathcal{P}_{src}$  as  $\mathcal{C}$ :
6    $\mathbf{Q} \leftarrow \mathbf{X}^n \mathbf{W}^Q, \mathbf{K} \leftarrow \boldsymbol{\mu} \mathbf{W}^K, \mathbf{V} \leftarrow \boldsymbol{\mu} \mathbf{W}^V$ 


---


7 Part II: Video Editing with STEM Inversion


---


8 for  $t = T$  to  $1$  do
9    $\mathcal{Y}_{t-1} = [Y^1_{t-1}, \dots, Y^N_{t-1}] \leftarrow$  editing with
    // spatial-temporal correspondence and  $\mathcal{P}_{tgt}$ .

```

ity of the k -th basis $\boldsymbol{\mu}_k$ to \mathbf{X}_m :

$$\mathcal{Z}_{km} = \underbrace{\frac{\exp(\boldsymbol{\mu}_k \mathbf{X}_m^T / \tau)}{\sum_{j=1}^K \exp(\boldsymbol{\mu}_j \mathbf{X}_m^T / \tau)}}_{\text{1st dim}}. \quad (8)$$

Here, τ is a temperature hyper-parameter controlling the shape of distribution \mathcal{Z} . We also provide a more brief perspective to understand the E step in our STEM inversion. As K-means clustering [27] is a special case of EM algorithm, step E can be regarded as calculating the responsibility of k -th clustering center for each pixel of the video feature \mathbf{X} . That is, \mathcal{Z}_{mk} represents the possibility of the m -th pixel belonging to the k -th basis. Summing up, the E step in STEM inversion can be described as:

$$\mathcal{Z} = \underbrace{\text{softmax}(\boldsymbol{\mu} \mathbf{X}^T / \tau)}_{\text{1st dim}}. \quad (9)$$

4.4. ST-M Step: Likelihood Maximization

The M step in the proposed STEM inversion aims to update the bases $\boldsymbol{\mu}$ by maximizing the likelihood of the com-

plete data (i.e., given \mathbf{X} and estimated \mathcal{Z}). During the r -th iteration, the bases $\boldsymbol{\mu}$ is updated via the following weighted summation formulation:

$$\boldsymbol{\mu}_k = \sum_{i=1}^M \frac{\mathcal{Z}_{ki} \mathbf{X}_i}{\sum_{m=1}^M \mathcal{Z}_{km}}. \quad (10)$$

Here, the weighted summation of \mathbf{X} encourages the bases $\boldsymbol{\mu}$ to share the same embedding space with the input feature \mathbf{X} . To put it briefly, the M step can be described as:

$$\boldsymbol{\mu} = \mathcal{Z} \mathbf{X} / \overbrace{\sum}^{\text{2nd dim}} (\mathcal{Z}). \quad (11)$$

When $\tau \rightarrow 0$ in Eq.(8) and Eq.(9), for any $m \in \{1, 2, \dots, M\}$, $\mathcal{Z}_{\cdot, m}$ turns to a one-hot vector. In this way, each pixel in a video only belongs to one basis. Then, the basis would be updated by the mean of those pixels that belong to it. In this situation, EM algorithm degenerates into K-means clustering.

4.5. Using STEM Inversion for Video Editing

We show the pseudo algorithm of performing zero-shot video editing with our STEM inversion in Alg. 1. First, the video feature $\mathbf{X} \in \mathbb{R}^{NHW \times C}$ are extracted from the input $\mathcal{I} = [I^1, \dots, I^N]$. Next, each frame I^n undergoes Eq.(3) to derive the noisy latent \mathbf{z}_t^n , where the STEM-SA process of \mathbf{X}^n applies Eq.(7). Then, we denoise the latent with the editing process of the existing zero-shot video editing methods, such as TokenFlow [9] and Fatezero [30], where the former is based on the typical 2D DDIM inversion and the latter on the naive spatial-temporal DDIM inversion.

Besides computational complexity, our STEM inversion has a unique advantage over the typical DDIM inversion and the inflated one. Ours explores a fixed global representation rather than a time-varying one for each frame, which is more stable for reconstruction and editing.

5. Experiment

5.1. Experimental Settings

Dataset. We evaluate our methods on the videos from DAVIS dataset [29] and the Internet, which contains various objects and motions. Each video consists of $40 \sim 200$ frames, where the size is 512×512 or 360×640 .

Implementation Details. We use the official code of Stable Diffusion [35], where the pre-trained weights are from version 1.5. In all experiments, we apply DDIM deterministic sampling with 50 steps. Following [30, 46], we use 50 forward steps and the classifier-free guidance of 7.5 when inverting a video. Besides, we set the hyper-parameter τ in Eq.(8) as 0.05, and the iteration number R as 3. We perform all experiments on a single NVIDIA Tesla A100 GPU.

Evaluation Metrics of Inversion. To demonstrate the effectiveness and efficiency of the proposed STEM inversion,



Figure 3. Ablation with different basis number K . Left: The reconstruction results of DDIM and our STEM inversion. Right: The corresponding editing results of various inversion settings, where TokenFlow [9] editing process is used. Best viewed with zoom-in.

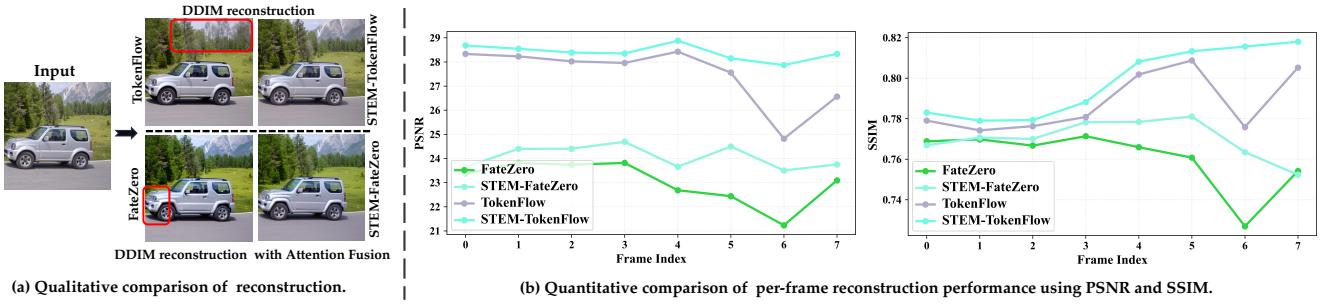


Figure 4. Qualitative and quantitative comparison of the reconstruction with DDIM and STEM inversion, where two reconstruction fashions are applied: (i) DDIM reconstruction, (ii) DDIM reconstruction with additional attention fusion (i.e., Fatezero [30] reconstruction).

we use *PSNR* [16] (\uparrow) and *SSIM* [44] (\uparrow) to evaluate the quality of each frame in the reconstructed video.

Evaluation Metrics of Edited Videos. Following [3, 9], our first metric *CLIP score*(\uparrow) measures edit fidelity, which calculates the similarity of CLIP embedding [31] of each edited frame and the target prompt. The second metric *warp error* (\downarrow) reflects temporal consistency. We first estimate optical flow [41] of the input video, which is used to warp the edited frames. Then, we compute the average MSE between each warped frames and the target ones.

5.2. Ablation Study

To investigate the influence of basis number K , we report the reconstruction and editing performance when K is 128, 256, and 512 respectively. Here, the video resolution is 512×512 , and the frame number N is 120. As shown in Fig. 3, $K = 256$ achieves a similar reconstruction and editing performance with $K = 512$. It shows that 256 bases are sufficient to represent the entire video. However, when $K = 128$, although our STEM inversion still derives com-

petitive reconstruction, the editing ability is weakened compared with $K = 256$. Thus, we use $K = 256$ as default.

5.3. Inversion Comparison

Qualitative comparison. We apply an 8-frame video to evaluate performance, where we consider two reconstruction fashions: (i) the typical DDIM reconstruction, used via TokenFlow [9]; (ii) additionally leveraging attention fusion during DDIM reconstruction, used via FateZero [30]. Fig. 4 shows STEM inversion consistently boosts the performance of DDIM reconstruction and the one with attention fusion. Especially, the distant mountains can be accurately reconstructed via ours which DDIM inversion failed. See more reconstruction comparisons in the SM.

Quantitative comparison. We give a quantitative evaluation of reconstruction quality with different inversions. In Fig. 4, we report *PSNR* and *SSIM* of each reconstructed frame, where the typical DDIM reconstruction (TokenFlow and STEM-TokenFlow) and the one with attention fusion [30] (FateZero and STEM-FateZero) are both listed.

Settings	DDIM Inversion			STEM Inversion		
	1-frame context	2-frame context	all-frame context	$K = 128$	$K = 256$	$K = 512$
Time (min)	0.80	1.12	5.04	1.02	1.50	2.26

Table 1. Time comparation between different inversion, where the inversion steps are 50. The frame number N is 48 and the spatial resolution of each frame is 640×360 here. We measure them in minute (min).

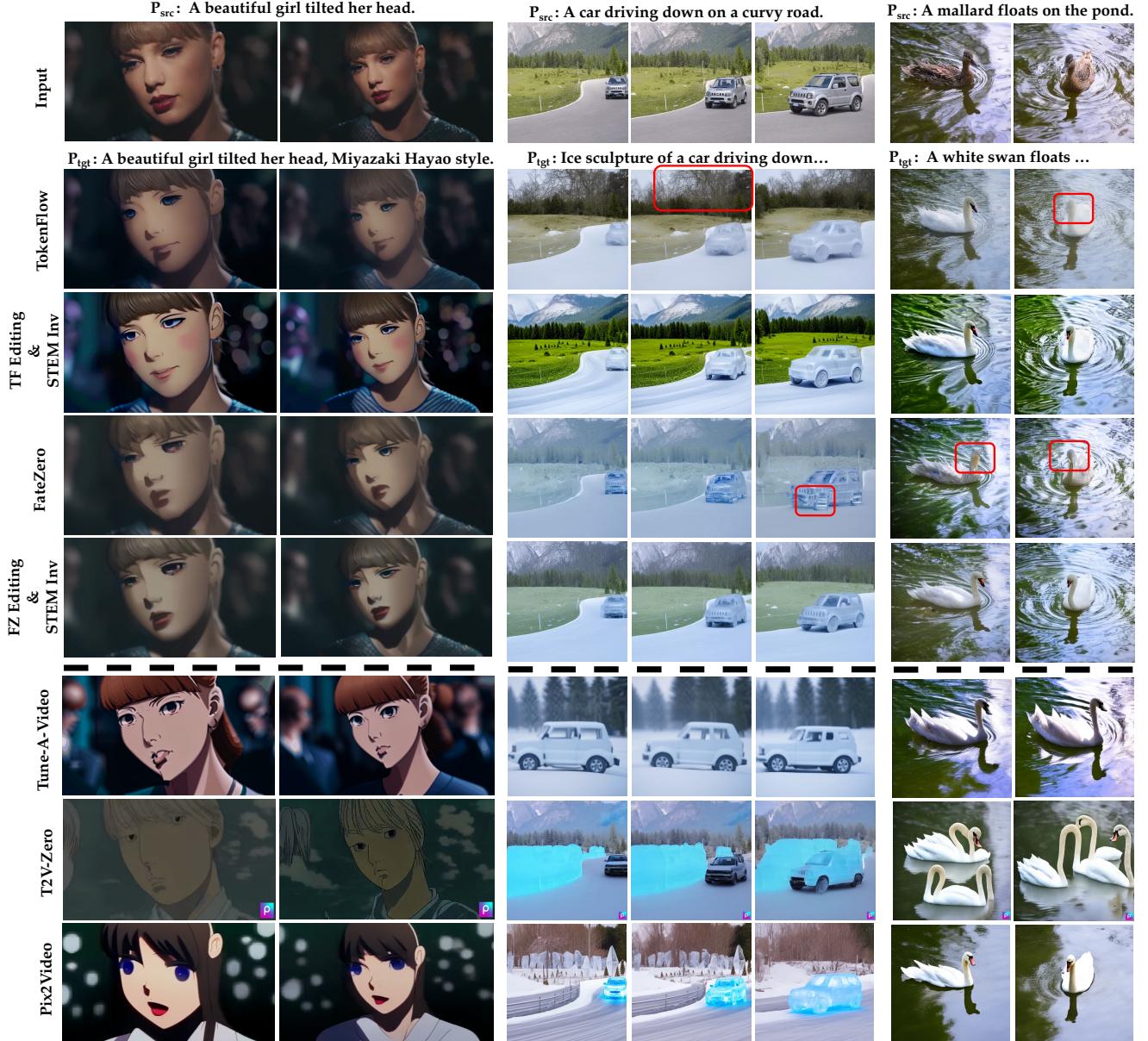


Figure 5. Qualitative comparison between different video editing methods. The editing scenarios here include style transfer, attribute editing, and shape editing. Best viewed with zoom-in.

We find our STEM inversion always achieves better performance than DDIM inversion for video reconstruction.

Feature visualization. In Fig. 6, we first estimate the optical flow [41] of the input video. Then, we apply PCA on the output features of the last SA layer from the UNet decoder. The 4-th column shows the feature visualization when we

use optical flow to warp the former-frame features. Last, we give the cosine similarity of the warped features and the target ones. Here, the brighter, the better temporal consistency can be achieved from the perspective of optical flow.

Time Cost. We measure time cost for various inversions, where a 48-frame video with 640×360 size is applied. As

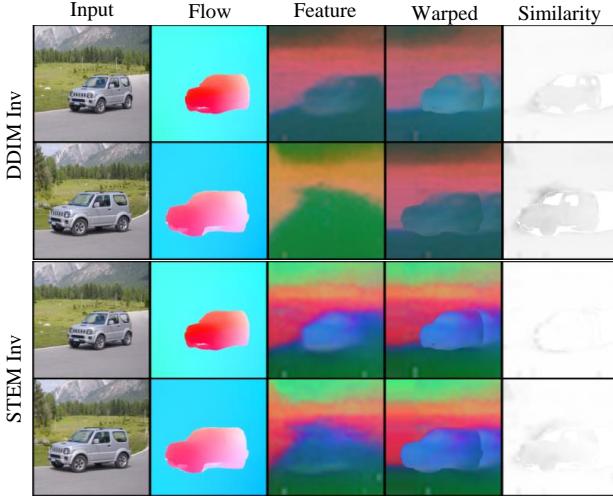


Figure 6. Visualization of different inversions. We first estimate optical flow [41] of the input video. Then, we apply PCA to the features of last SA layer from the decoder with different inversion. Next, we use the optical flow to warp the features of former-frame to obtain the warped features in the 4-th column. Last, we show the cosine similarity of features from 3-rd and 4-th column.

seen in Table.1, the typical DDIM inversion using 1-frame context for each frame costs the smallest time 0.8 min. The naive spatial-temporal DDIM inversion exploring a 2-frame context would increase the time cost by 40%, yielding 1.12 min. We also provide a more radical inflated DDIM version for reference, which uses all-frame context and needs 5.04 min. In contrast, our STEM ($K = 128$ and $K = 256$) not only harvests the spatial context from the entire video but also achieves a similar time cost with naive spatial-temporal DDIM inversion using a 2-frame context.

One may wonder why STEM inversion can reduce computational complexity, but it still takes more time than the typical DDIM inversion. The reason is the iterations of EM algorithm are serial, which is unfavourable for time cost to some extent. Both the radical inflated DDIM inversion and ours can explore the global context from the entire video, but our time cost is very advantageous compared with it.

5.4. Comparison of Video Editing

Qualitative Comparison. We first give the qualitative comparison of editing results when using DDIM and our STEM inversion respectively. From Fig. 5, we find that STEM inversion can improve the editing performance of both TokenFlow and FateZero. Moreover, FateZero struggles to perform shape editing with Stable Diffusion (see Mallard example). Instead, it usually uses the pre-trained model [46] for shape editing. Fortunately, by inserting our STEM inversion, we can empower FateZero with shape editing capabilities. Besides, we compare our method with current state-of-the-art zero-shot video editing: Tune-A-Video [46], Pix2Video [3], and Text2Video [18], where ours

Method	CLIP score \uparrow	Warp error ($1e-3$) (\downarrow)
Tune-A-Video [46]	0.28	32.3
Pix2Video [3]	0.32	6.2
Text2Video-Zero [18]	0.31	22.7
FateZero [30]	0.29	7.2
TokenFlow [9]	0.31	4.9
Ours (STEM-FateZero)	0.30	4.3
Ours (STEM-TokenFlow)	0.31	3.5

Table 2. Qualitative comparison with other video editing methods.

Method	edit fidelity	temporal consistency
FateZero [30]	17%	33%
Ours (STEM-FateZero)	83%	67%
TokenFlow [9]	25%	33%
Ours (STEM-TokenFlow)	75%	67%

Table 3. User study results (%), where we show the averaged selection percentages of each method.

exceed them on both edit fidelity and temporal consistency. We give more examples in the SM.

Quantitative Comparison. The quantitative comparison is in Table. 2. We achieve competitive CLIP score and the lowest warp error. Moreover, compared with FateZero (or TokenFlow), our STEM-Fatezero (or STEM-TokenFlow) can always yield a better performance, demonstrating the superiority of our STEM inversion.

User Study. To prove the effectiveness of our STEM inversion, we additionally provide a user study to make people choose FateZero *vs* STEM-FateZero and TokenFlow *vs* STEM-TokenFlow. Corresponding to two key objectives of video editing, we ask users to choose: *a) the one with the higher edit fidelity, and b) the one with the better temporal consistency*. For each user, we randomly sample 12 videos. We report the selected ratios in Table. 3. It can be seen that those editings that cooperated with our STEM inversion always earned the higher user preference.

6. Conclusion

In this paper, we present a STEM video inversion method. Concretely, like EM algorithm, we evaluate a more compact basis set by performing E step and M step iteratively. Considering our STEM inversion does not require any finetuning, it can be seamlessly integrated into other diffusion-based video editing methods that utilize DDIM inversion as initial noise. With its clear technical advantages, We hope that our STEM inversion can serve as a new inversion method for video editing for future contributions.

While our work shows some promising results, there are still some limitations. For example, our STEM inversion chose 256 bases as the low-rank representation of the whole video. However, for long videos or videos with high resolution, 256 bases would be not enough to obtain sufficient editing ability. An adaptive number of bases is more reasonable. We will leave it as future work.

References

- [1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2
- [2] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 2
- [3] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 1, 2, 3, 6, 8, 11
- [4] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. 2, 3
- [5] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 1
- [6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2
- [7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2
- [8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1
- [9] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1, 2, 3, 5, 6, 8, 11, 12, 13
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(1):139–144, 2020. 2
- [11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1
- [12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 3
- [15] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 2
- [16] Quan Huynh-Thu and Mohammed Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. *Telecommunication Systems*, 49:35–48, 2012. 6
- [17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2
- [18] Levon Khachatryan, Andranik Mojsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *ICCV*, 2023. 8, 11
- [19] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2
- [20] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 2
- [21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- [22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 1
- [23] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2
- [24] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019. 2, 4
- [25] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 1
- [26] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 2
- [27] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 5
- [28] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 3

- [29] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5
- [30] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3, 5, 6, 8, 11, 12, 13
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6
- [32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2
- [33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1, 2
- [34] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*, 2019. 2
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 5
- [36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 1
- [37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2
- [38] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 2
- [39] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2
- [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 3
- [41] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II* 16, pages 402–419. Springer, 2020. 6, 7, 8
- [42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2
- [43] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 2
- [44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [45] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nūwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022. 2
- [46] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3, 5, 8, 11
- [47] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédéric Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 1
- [48] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 2
- [49] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunnar Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2
- [50] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023. 1
- [51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2
- [52] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 2
- [53] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 2

Appendix

A. Inversion Comparison

A.1. Reconstruction Comparison

We provide the reconstruction comparison of DDIM inversion and STEM inversion in Fig. 8 and the supplement video. Since Fatezero [30] stores the intermediate self-attention maps and cross-attention maps at each timestep t , it is memory-consuming and cannot perform video editing over 20 frames on a single A100 GPU. In each example of the supplement video, FateZero results are sampled at a proper rate.

As seen in Fig. 8, two reconstruction fashions are applied for DDIM and STEM inversion separately: (i) the typical DDIM reconstruction (used by TokenFlow [9]), (ii) DDIM reconstruction with extra attention fusion (used by FateZero [30]). The proposed STEM inversion always delivers better reconstruction than DDIM inversion, especially under the typical DDIM reconstruction fashion. Such a benefit is derived from our STEM inversion modelling global and fixed context for each frame while DDIM inversion explores a time-varying and limited spatial-temporal context.

A.2. Feature Similarity in Different Forward Steps

Recall that in Fig. 6 of our main paper, we first use optical flow to warp the former-frame features and obtain the warped feature. Then, we calculate the cosine similarity between the wrapped feature from the former frame and the current frame feature. The more similar, the reconstructed video is more coherent in time dimension.

In this supplement, we provide the mean cosine similarity across different time steps t in Fig. 7. The higher similarity indicates that our STEM inversion can achieve better temporal consistency from the perspective of optical flow.

B. More Comparison of Various Text-driven Zero-shot Video Editing

To prove the efficiency of our STEM inversion, we compare our STEM-TokenFlow and STEM-FateZero with the current state-of-the-art video editing methods in Fig. 9 and our supplement video. Specifically, although T2V-Zero [18] can perceive the style and subject to be edited, it always deviates greatly from the original video and cannot maintain a satisfying temporal consistency. Besides, Tune-A-Video [46] needs to perform training on the video before editing while its performance is inferior to ours. Moreover, it is difficult for Pix2Video [3] to maintain the background. Please see the second from last row of Fig. 9.

Note that FateZero [30] struggles to conduct shape editing (see "cow" → "boar"). Our STEM inversion is able to

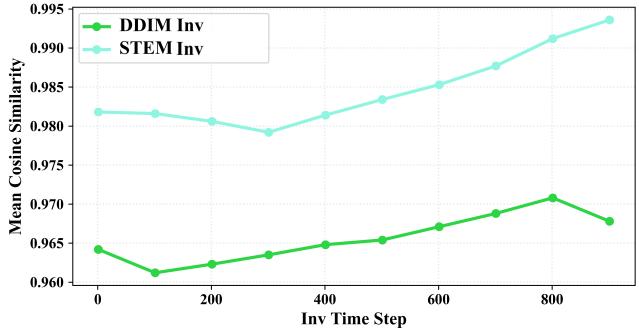


Figure 7. The mean cosine similarity between the warpped features from the former frame and the target features during various inversion steps. The more similar, the better.

endow shape-editing ability to FateZero, which also demonstrates the superiority of our method. Besides, by replacing DDIM inversion with the proposed STEM one, TokenFlow also yields more high-quality video editing results. As seen in Fig. 9, our STEM-TokenFlow has better editing fidelity when transferring the video to *Johannes Vermeer style*.

C. More Implementation Details

The attention fusion ratio in FateZero [30] is a hyper-parameter that controls the editing effect. Specifically, it fuses both cross and self-attention at DDIM time step $t \in [0.5T, T]$. However, we experimentally discovered a small fusion ratio for cross attention time step is better when using our STEM inversion for Fatezero editing. Concretely, we adopt the cross attention time step ratio as $t \in [0.2T, T]$, while the same ratio $t \in [0.5T, T]$ for self-attention. The possible underlying reason is that our STEM inversion is more sensitive to capturing the semantics from the target prompt than the DDIM one. Thus, a smaller cross-attention fusion ratio is sufficient under the FateZero editing scenario. Besides, in terms of TokenFlow editing, we use identical hyper-parameters when replacing DDIM inversion with our STEM inversion.

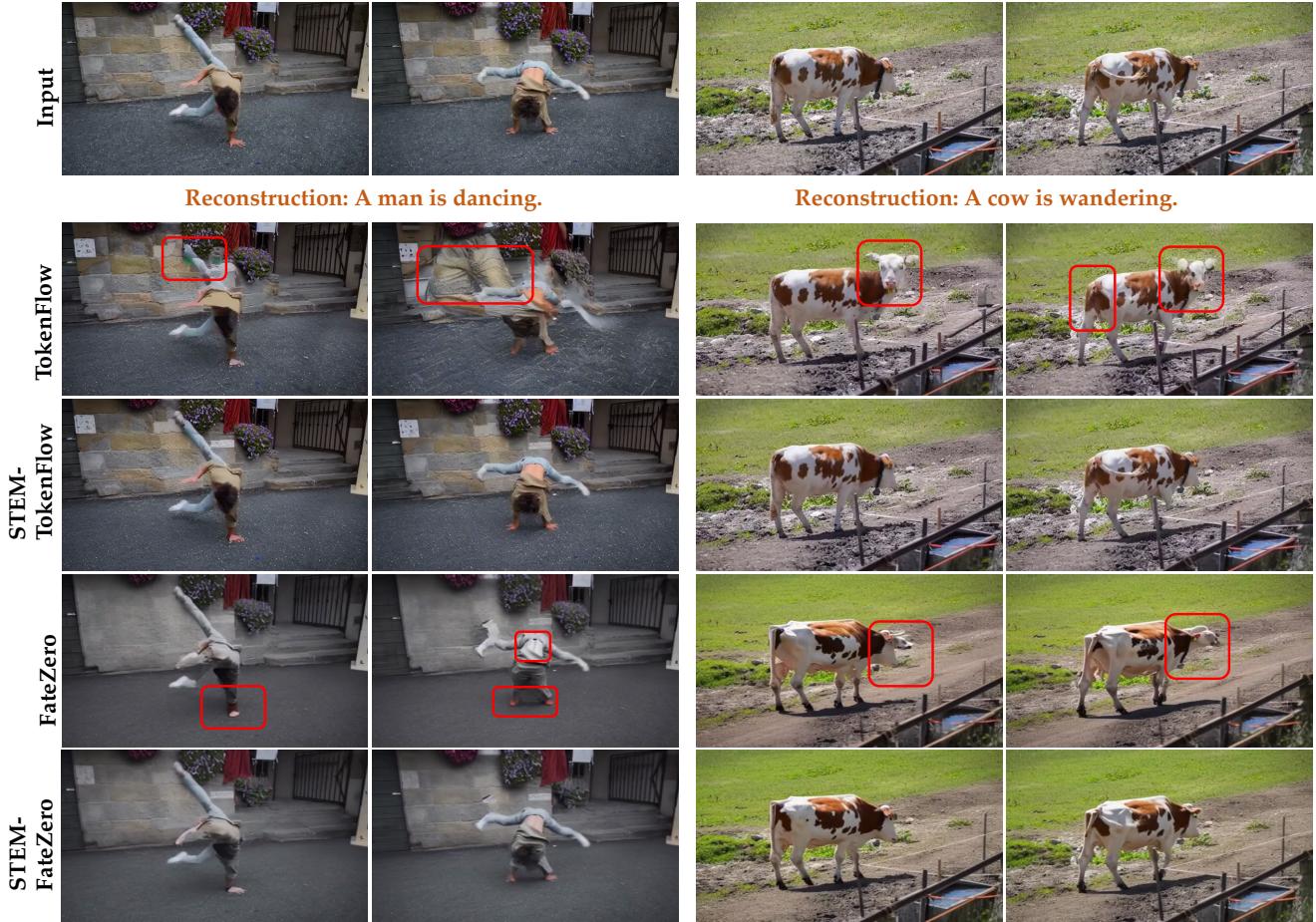


Figure 8. Qualitative comparison of the reconstruction with DDIM and STEM inversion, where two reconstruction fashions are applied: (i) DDIM reconstruction (i.e., TokenFlow [9] reconstruction)), (ii) DDIM reconstruction with additional attention fusion (i.e., Fatezero [30] reconstruction).



Figure 9. Qualitative comparison between different text-driven video editing methods. Our STEM-inversion can consistently improve the editing performance of TokenFlow [9] and FateZero [30]. Best viewed with zoom-in.