FateZero: Fusing Attentions for Zero-shot Text-based Video Editing
Supplemental Material

Chenyang Qi[1*] Xiaodong Cun[2†] Yong Zhang[2] Chenyang Lei[3]
Xintao Wang[2] Ying Shan[2] Qifeng Chen[1†]

[1]HKUST [2]Tencent AI Lab [3]CAIR, HKISI-CAS

https://fate-zero-edit.github.io

A. Implementation Details

Pseudo algorithm code Our full algorithm is shown in Algorithm 1 and Algorithm 2. Algorithm 1 presents the overall framework of our inversion and editing, as visualized in the left of Fig. 1 in the main paper. Algorithm 2 shows that the cross-attention is fused based on a mask of the edited words, and the self-attention is blended using a binary mask from thresholding the cross-attention (the right of Fig. 1 in the main paper).

Hyperparameters Tuning. There are mainly three hyperparameters in our proposed designs:
- $t_s \in [1, T]$: Last timestep of the self-attention blending. Smaller t_s fuses more self-attention from inversion to preserve structure and motion.
- $t_c \in [1, T]$: Last timestep of the cross attention fusion. Smaller t_c fuses more cross attention from inversion to preserve the spatial semantic layout.
- $\tau \in [0, 1]$: Threshold for the blending mask used in shape editing. Smaller τ uses more self-attention map from editing to improve shape editing results.

In **style** and **attribute** editing, we set $t_s = 0.2T$, $t_c = 0.3T$, $\tau = 1.0$ to preserve most structure and motion in the source video. In **shape** editing, we set $t_s = 0.5T$, $t_c = 0.5T$, $\tau = 0.3$ to give more freedom in new motion and 3D shape generation.

B. More results

B.1. Inpainting

We implement the latent inpainting for shape editing (*silver jeep → Porsche car*) as in the below figure. The edited background is identical to the input.

Input Video Our Edited Results

B.2. Source Similarity and Warping-error.

Note that the video editing quality can not be well evaluated by the similarity with the original video using the "frame-wise L1 distance" (as it reaches its minimum in the unedited video) or "warping error" (unsuitable for our general editing task that produces new shape and attribute). Thus, we follow previous video editing works (NLA [4], Text2live [1], Shape-aware [5], Vtoonify [6], GEN-1 [2]) to design the metrics. Here, the table below is provided as a reference, where our method achieves the best temporal consistency and the second similarity with the source video.

Method	Frame-Null	Frame-SDEdit	NLA-Null	Tune-A-Video	Ours
L1-Src-Distance↓	0.118	**0.079**	0.126	0.181	<u>0.114</u>
Warping-error↓	<u>0.207</u>	0.285	0.255	0.314	**0.163**

B.3. Quantitative Ablation on Self-attention Fusion.

We provide the results on shape editing videos in the table below. We use both CLIP metrics (Tem-Con. / Fram-Acc. as in the paper), and pixel-wise similarity (L1 and warping errors). Our full method achieves the best clip metrics and 2^{nd} similarity with the source video. Note that, fusing self-attention at each pixel without a spatial mask severely reduces the editing quality, which produces the worst Tem-Con/Fram-Acc.

Method	w/ recon-attn	w/o self-attn	Self-attn w/o mask	Ours
Tem-Con↑/Fram-Acc↑	0.968/0.987	0.969/0.990	0.956/0.908	**0.970/0.993**
L1-Distance↓/Warping-error↓	0.154/0.212	0.143/0.219	**0.093/0.130**	<u>0.118</u>/<u>0.181</u>

C. Demo Video

we provide a detailed demo video to show:
Video Results on style, local attribute, and shape editing to validate the effectiveness of the proposed method.

*Intern at tencent AI Lab

black swan → yellow pterosaur.

Figure 1. limitation of our zero-shot editing.

Method Animation to provide a better understanding of the proposed method.
Baseline Comparisons with previous methods in video.
More Promising Applications We have shown the effectiveness of the proposed method in the main paper for style, attribution, and shape editing. In the demo video, we also show some potential applications of the proposed method, including (1) object removal by removing the word of the target object in the source prompt and mask the self-attention of the corresponding area using its cross attention, (2) video enhancement by adding the specific prompt (*e.g.*, 'high-quality', '8K') in the target editing prompt.

D. Limitation and Future Work

Our zero-shot editing is not good at new concept composition or generation of very different shapes. For example, the result of editing 'black swan' to 'yellow pterosaur' in Fig 1 is unsatisfactory. This problem may be alleviated using a stronger video diffusion model, which we leave to future work.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 1

[2] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 1

[3] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[4] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 1

[5] Yao-Chih Lee, Ji-Ze Genevieve Jang Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing demo. *arXiv preprint arXiv:2301.13173*, 2023. 1

Algorithm 1 `FateZero` Algorithm

Input:
- z_0: Latent code from source video
- p_{src}: Source text prompt for input video
- p_{edit}: Target text prompt for edition

Hyperparameters:
- t_c: Last timestep of the cross attention fusion
- t_s: Last timestep of the self attention blending
- τ: Threshold for blending mask

Output:
- $\hat{z}_0$: Final edited latent code

▷ DDIM for inversion latents and attention maps
for $t = 1, 2, ..., T$ **do**
  $\epsilon_t, c_t^{\text{src}}, s_t^{\text{src}} \leftarrow \epsilon_\theta(z_t, t, p_{src})$
  $z_t = \sqrt{\alpha_t}\, \frac{z_{t-1} - \sqrt{1-\alpha_{t-1}}\epsilon_t}{\sqrt{\alpha_{t-1}}} + \sqrt{1-\alpha_t}\epsilon_t$
end for

▷ Denoising the inverted latents with attention fusion
for $t = T, (T-1), ..., 1$ **do**
  Edited_index $= (p_{src}$!= $p_{edit})$
  ▷ Cross-attention mask is from the edited index [3]
  M_{cross}[Edited_index] $= 1$
  ▷ Self-attention blending mask is from cross-attention.
  $M_{\text{self}} = (c_t^{\text{src}}$[Edited_index] $> \tau)$
  $\hat{\epsilon_t} \leftarrow$ ATT-FUSION$(\varepsilon_\theta, z_t, t, p_{\text{edit}}, M_{\text{edit}}, M_{\text{self}}, c_t^{\text{src}}, s_t^{\text{src}})$
  $z_{t-1} = \sqrt{\alpha_{t-1}}\, \frac{z_t - \sqrt{1-\alpha_t}\hat{\epsilon_t}}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\hat{\epsilon_t}$
end for
▷ Fuse the inversion and editing attention of all B blocks.
▷ We only show the operation of attention and omit the feed-forward, residual convolution layer for simplicity.
function ATT-FUSION$(\varepsilon_\theta, z_t, t, p_{\text{edit}}, M_{\text{cross}}, M_{\text{self}}, c_t^{\text{src}}, s_t^{\text{src}})$
  for $i = 1...B$ **do**
    $s_t^{\text{edit}} = \text{Softmax}(W_i^Q(z_t)W_i^K(z_t)/\sqrt{d_i})$
    $s_t^{\text{fused}} =$ SELF-BLENDING$(s_t^{\text{edit}}, s_t^{\text{src}}, M_{\text{self}}, c_t^{\text{src}}, t)$
    $z_t \quad = W_i^V(z_t) \cdot s_t^{\text{fused}}$
    $c_t^{\text{edit}} = \text{Softmax}(W_i^Q(z_t)W_i^K(p_{edit})/\sqrt{d_i})$
    $c_t^{\text{fused}} =$ CROSS-FUSION$(c_t^{\text{edit}}, c_t^{\text{src}}, M_{\text{edit}}, t)$
    $z_t \quad = W_i^V(p_{\text{edit}}) \cdot c_t^{\text{fused}}$
  end for
  return z_t
end function

[6] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 1

Algorithm 2 Attention Fusion and Blending Algorithm

▷ Following prompt-to-prompt, Cross-attention fusion using the difference mask between source and editing prompt: M_{cross} with shape (text_embedding, words_count)
function CROSS-FUSION($c_t^{\text{edit}}, c_t^{\text{src}}, M_{\text{cross}}, t$)
    if $t > t_c$ **then**
        return $M_{\text{cross}} \cdot c_t^{\text{edit}} + (1 - M_{\text{cross}}) \cdot c_t^{\text{src}}$
    else
        return c_t^{edit}
    end if
end function

▷ Self-attention blending with the mask from cross attention map: M_{self} with shape $(1, \text{height} \times \text{width})$
function SLEF-BLENDING($s_t^{\text{edit}}, s_t^{\text{src}}, c_t^{\text{src}}, M_{\text{self}}, t$)
    if $t > t_s$ **then**
        return $M_{\text{self}} \cdot s_t^{\text{edit}} + (1 - M_{\text{self}}) \cdot s_t^{\text{src}}$
    else
        return s_t^{edit}
    end if
end function