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

Supplemental Material

Chenyang Qi\textsuperscript{1}*, Xiaodong Cun\textsuperscript{2†} Yong Zhang\textsuperscript{2} Chenyang Lei\textsuperscript{3} Xintao Wang\textsuperscript{2} Ying Shan\textsuperscript{2} Qifeng Chen\textsuperscript{1†}

\textsuperscript{1}HKUST \hspace{1cm} \textsuperscript{2}Tencent AI Lab \hspace{1cm} \textsuperscript{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 \( \tau \) uses more self-attention map from editing to improve shape editing results.

In \textit{style} and \textit{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 \textit{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 (\textit{silver jeep} \rightarrow \textit{Porsche car}) as in the below figure. The edited background is identical to the input.

\*Intern at Tencent AI Lab

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.

\begin{tabular}{lcccccc}
Method & Frame-Null & Frame-SDEdit & NLA-Null & Tune-A-Video & Ours \\
L1-Src-Distance & 0.118 & 0.079 & 0.126 & 0.181 & 0.114 \\
Warping-error & 0.207 & 0.285 & 0.255 & 0.314 & 0.163 \\
\end{tabular}

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\textsuperscript{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.

\begin{tabular}{lcccccc}
Method & w/ recon-attn & w/o self-attn & Self-attn w/o mask & Ours \\
Tem-Con/Fram-Acc & 0.968/0.987 & 0.960/0.990 & 0.956/0.988 & 0.978/0.993 \\
L1-Distance/Warping-error & 0.1540/0.212 & 0.1430/0.219 & 0.0930/0.130 & 0.1180/0.181 \\
\end{tabular}

C. Demo Video

we provide a detailed demo video to show:

\textbf{Video Results} on style, local attribute, and shape editing to validate the effectiveness of the proposed method.
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


Algorithm 1 FateZero Algorithm

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

Hyperparameters:
- \( t_e \): Last timestep of the cross attention fusion
- \( t_s \): Last timestep of the self-attention blending
- \( \tau \): Threshold for blending mask

Output:
- \( z_f \): Final edited latent code

// DDIM for inversion latents and attention maps
for \( t = 1, 2, \ldots, T \) do
\[
\varepsilon_t, c^t_{\text{src}}, s^t_{\text{src}} = \varepsilon_0(z_t, t, p_{\text{src}})\\
z_t = \sqrt{\alpha_t} \cdot \frac{z_{t-1} - \sqrt{1 - \alpha_{t-1}} \varepsilon_t}{\sqrt{\alpha_t}} + \sqrt{1 - \alpha_t} \varepsilon_t
\]
end for

// Denoising the inverted latents with attention fusion
for \( t = T, (T - 1), \ldots, 1 \) do
Edited_index = \( p_{\text{src}} \Rightarrow p_{\text{edit}} \)
\( \text{Cross-attention mask is from the edited index}[3] \)
\( M_{\text{cross}[\text{Edited_index}]} = 1 \)
\( \text{Self-attention blending mask is from cross-attention.} \)
\( M_{\text{self}} = \text{(}\varepsilon_{t_{\text{src}}}^{\text{Edited_index}} > \tau\text{)} \)
\( \hat{\varepsilon}_{t} = \text{ATT-ATTENTION} (\varepsilon_{t_{\text{src}}}, z_{t}, t, p_{\text{edit}}, M_{\text{edit}}, M_{\text{self}}, c^t_{\text{src}}, s^t_{\text{src}}) \)
\( z_{t-1} = \sqrt{\alpha_{t-1}} \cdot \frac{z_{t-1} - \sqrt{1 - \alpha_{t-1}} \hat{\varepsilon}_{t-1}}{\sqrt{\alpha_{t-1}}} + \sqrt{1 - \alpha_{t-1}} \hat{\varepsilon}_{t-1} \)
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-ATTENTION(\( \varepsilon_{t_{\text{src}}} \), \( z_{t} \), \( t_{\text{edit}} \), \( p_{\text{edit}} \), \( M_{\text{cross}} \), \( M_{\text{self}} \), \( c^t_{\text{src}} \), \( s^t_{\text{src}} \))
for \( i = 1, \ldots, B \) do
\( s^t_{\text{edit}} = \text{Softmax}(W_{t_{\text{edit}}}^Q(z_{t}), W_{t_{\text{edit}}}^K(z_{t})/\sqrt{d_t}) \)
\( s^t_{\text{fused}} = \text{SELF-BLENDING}(s^t_{\text{edit}}, s^t_{\text{src}}, M_{\text{self}}, c^t_{\text{src}}, t) \)
\( z_{t} = W_{t}^{V} (z_{t}) \cdot s^t_{\text{fused}} \)
\( c^t_{\text{edit}} = \text{Softmax}(W_{t_{\text{edit}}}^Q(z_{t_{\text{edit}}}), W_{t_{\text{edit}}}^K(p_{\text{edit}})/\sqrt{d_t}) \)
\( c^t_{\text{fused}} = \text{CROSS-BLENDING}(c^t_{\text{edit}}, c^t_{\text{src}}, M_{\text{edit}}, t) \)
\( z_{t} = W_{t}^{V} (p_{\text{edit}}) \cdot c^t_{\text{fused}} \)
end for
return \( z_{t} \)
end function

Figure 1. limitation of our zero-shot editing.
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($c_{edit}^t, c_{src}^t, M_{cross}, t$)

if $t > t_c$ then
   return $M_{cross} \cdot c_{edit}^t + (1 - M_{cross}) \cdot c_{src}^t$
else
   return $c_{edit}^t$
end if

end function

▷ Self-attention blending with the mask from cross attention map: $M_{self}$ with shape (1, height $\times$ width)

function SLEF-BLENDING($s_{edit}^t, s_{src}^t, t$, $c_{src}^t, M_{self}, t$)

if $t > t_s$ then
   return $M_{self} \cdot s_{edit}^t + (1 - M_{self}) \cdot s_{src}^t$
else
   return $s_{edit}^t$
end if

end function