Towards Consistent Video Editing with Text-to-Image Diffusion Models ——NeurIPS 2023 Supplementary Material

Zicheng Zhang[1]* **Bonan Li**[1]* **Xuecheng Nie**[2] **Congying Han**[1]† **Tiande Guo**[1] **Luoqi Liu**[2]

[1]University of Chinese Academy of Sciences [2]MT Lab, Meitu Inc.

Contents

A Details of Theoretic Analysis and Proof **1**

A.1 Definitions and notations 1

A.2 Distribution transition in Transformer Module 2

A.3 Covariate shift for tuning W_q 3

A.4 Covariate shift for appending Temporal Attention Module 4

A.5 Covariate shift for appending Shift-restricted Temporal Attention Module 4

B More Results **4**

B.1 Position for inserting Temporal Attention Module 5

B.2 More qualitative results 5

A Details of Theoretic Analysis and Proof

In this section, we illustrate the details of analysis and proof. For clarity, Some symbols may differ slightly from the paper.

A.1 Definitions and notations

Attention mechanism. Given a spatial-temporal feature $z \in \mathbb{R}^{b\times n\times l_1\times d_1}$, and textual embedding $c \in \mathbb{R}^{b\times n\times l_2\times d_2}$, where b, n, l, d denote the lengths of batch, frame, sequence and feature dimensions, the attention mechanism obtains three elements Query Q, Key K and Value V by $Q = zW_q,\ K = cW_k,\ V = cW_v$, where $W_q \in \mathbb{R}^{d_1\times d}$, W_k and $W_v \in \mathbb{R}^{d_2\times d}$. After that, they will interact to generate the transferred feature via

$$\text{Attention}(z, c) = M_{Q,K}V, \text{ where } M_{Q,K} = \text{softmax}(QK^T/\sqrt{d}). \tag{1}$$

Layer Normalization. Considering data $x \in \mathbb{R}^{l\times d}$, the *Layer Norm* [1] is defined as

$$Norm(x) = \alpha\frac{x-\mu_x}{\sigma_x} + \beta, \text{ where } \mu_x = \frac{1}{ld}\Sigma_{i,j=1}^{l,d}x^{i,j},\ \sigma_x = \sqrt{\frac{1}{ld}\Sigma_{i,j=1}^{l,d}(x^{i,j}-\mu_x)^2}, \tag{2}$$

where α and β are learnable scalar parameters.

*Equal contribution

†Corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Remark 1. *In practice, there is an alternative definition of Layer Norm, where μ_x and σ_x are computed only along feature dimension, e.g., $\mu_x = \frac{1}{d}\Sigma_{j=1}^{d} x^{\cdot,j}$. Both the two definitions are appropriate for the subsequent analysis.*

Transformer module. A classical Transformer module in LDM [4] has the residual structure:

$$\text{Transformer}(z, c) = z + Linear(\text{Attention}(norm(z), norm(c))), \tag{3}$$

where $Linear(z) = zW_L + b_L$, W_L and b_L are learnable parameters.

Remark 2. *In LDM [4], the Self Attention (SA) Module is defined by*

$$\text{Transformer}(z) = z + Linear(\text{Attention}(norm(z), norm(z)), \tag{4}$$

and the Cross Attention (CA) Module is defined by

$$\text{Transformer}(z, c) = z + Linear(\text{Attention}(norm(z), c), \tag{5}$$

We generalize and unify them into a common form in Eq. (3) for the sake of analysis.

A.2 Distribution transition in Transformer Module

Proposition 1. *We assume that for Transformers in Eq. (3), any input feature $z \in \mathbb{R}^{l\times d}$ (or $\mathbb{R}^{n\times d}$) is a sample of i.i.d d-dimensional Gaussian random variables, denoted as $rv(z) \sim N(\mu_{rv(z)}, \Sigma_{rv(z)})$, and $\hat{z} = Norm(z)$, same notation for c. **(i)** For the output from Eq. (1), the i-th row vector variable, $\text{Attention}(\hat{z}, \hat{c})^i$, follows the Gaussian distribution $N(\mu_{rv(V)}, \omega_{Q,K}\Sigma_{rv(V)})$, where $\mu_{rv(V)} = W_v^T\mu_{rv(\hat{c})}$, $\Sigma_{rv(V)} = W_v^T\Sigma_{rv(\hat{c})}W_v$, and $\omega_{Q,K} = \|M_{Q,K}^i\|_2^2$. **(ii)** For the output from Eq. (3), each row vector variable in $\text{Transformer}(z, c)$ follows the Gaussian distribution $N(\mu'_{rv(z)}, \Sigma'_{rv(z)})$, where $\mu'_{rv(z)} = \mu_{rv(z)} + W_L^T\mu_{rv(V)} + b_L$, and $\Sigma'_{rv(z)} = \Sigma_{rv(z)} + \omega_{Q,K}W_L^T\Sigma_{rv(V)}W_L$.*

Proof. The proof can be obtained directly by using the property of Gaussian distribution:

Under the assumption, z can viewed as the sample from $[rv(z)_1^T; \ldots; rv(z)_l^T]$, where $rv(z)_i, i = 1, \ldots, l$ are *i.i.d* and follow $N(\mu_{rv(z)}, \Sigma_{rv(z)})$. Given the affine mapping in Eq. (2), it is intuitive that the random variable $rv(\hat{z}) = \alpha\frac{rv(z)-\mu_z}{\sigma_z} + \beta$ follows Gaussian distribution, hence $\hat{z}$ can be viewed as sample from $[rv(\hat{z})_1^T; \ldots; rv(\hat{z})_l^T]$. Similar conclusion can be obtained for c, thus each row vector in V of Eq. (1) follows the Gaussian distribution, where the random variable

$$rv(V) = W_v^T rv(\hat{c}) \sim N(W_v^T\mu_{rv(\hat{c})}, W_v^T\Sigma_{rv(\hat{c})}W_v). \tag{6}$$

(i) Considering that the i-th row vector from Eq. (1),

$$A^i := \text{Attention}(\hat{z}, \hat{c})^i = \Sigma_{j=1}^{l} M_{Q,K}^{i,j}(\hat{c}W_v)^j. \tag{7}$$

According to the additive property of Gaussian distribution, the random variable

$$\begin{aligned} &rv(A^i) = \Sigma_{j=1}^{l} M_{Q,K}^{i,j} rv(V)_j \sim N(\mu_{rv(V)}, \omega_{Q,K}\Sigma_{rv(V)}), \\ &\text{where } \omega_{Q,K} = \Sigma_{j=1}^{l}(M_{Q,K}^{i,j})^2 = \|M_{Q,K}^i\|_2^2. \end{aligned} \tag{8}$$

(ii) For each $i = 1, \ldots, l$, we have

$$\text{Transformer}(z, c)^i = z^i + \text{Attention}(\hat{z}, \hat{c})^i W_L + b_L, \tag{9}$$

Therefore, we can derive the random variable

$$\begin{aligned} &rv(z) + W_L^T rv(A^i) + b_L \sim \\ &N(\mu_{rv(z)} + W_L^T\mu_{rv(V)} + b_L,\ \Sigma_{rv(z)} + \omega_{Q,K}W_L^T\Sigma_{rv(V)}W_L). \end{aligned} \tag{10}$$

□

A.3 Covariate shift for tuning W_q

Since *Tune-A-Video* solely tunes the parameters of Transformers including W_q in *SCA* and *CA*, which inherited from pre-trained LDM, to prevent covariate shift, $\mu'_{rv(z)}$ and $\Sigma'_{rv(z)}$ should not be largely changed before and after tuning. At first, we can observe that $\mu'_{rv(z)}$ is unrelated to W_q, hence tuning W_q does not affect the mean value. For $\Sigma'_{rv(z)}$, we can conclude that tuning W_q has little impact on it through the following theoretic analysis.

Lemma 1. *Given that $x \in \mathbb{R}^n$, the softmax function is defined by*

$$\mathrm{softmax}(x) = \frac{1}{\sum_{j=1}^{n} \exp(\lambda x_j)} \begin{bmatrix} \exp(\lambda x_1) \\ \vdots \\ \exp(\lambda x_n) \end{bmatrix}, \lambda > 0. \tag{11}$$

The softmax function is L-Lipschitz with respect to $\|\cdot\|_2$ with $L = \lambda$, that is, for all $z, z' \in \mathbb{R}^n$,

$$\|\mathrm{softmax}(x) - \mathrm{softmax}(x')\|_2 \leq \lambda \|x - x'\|_2 .$$

Please refer to Proposition 4 in [2] for the proof.

Proposition 2. *Along with notations in Proposition 1, denoting the tuned results of W_q as W_q^{tuning}, and the tuned Covariance of $\Sigma'_{rv(z)}$ as $\Sigma'^{tuning}_{rv(z)}$, we have*

$$\|\Sigma'^{tuning}_{rv(z)} - \Sigma'_{rv(z)}\| \leq \alpha \|W_q^{tuning} - W_q\|_2, \tag{12}$$

where α is a constant related to the definition of matrix norm but unrelated to the tuning process.

Proof. First, we expand the formula

$$\begin{aligned} \|\Sigma'^{tuning}_{rv(z)} - \Sigma'_{rv(z)}\| &= \|\omega_{Q_{tuning},K} W_L^T \Sigma_{rv(V)} W_L - \omega_{Q,K} W_L^T \Sigma_{rv(V)} W_L\| \\ &= |\omega_{Q_{tuning},K} - \omega_{Q,K}| \|W_L^T \Sigma_{rv(V)} W_L\|. \end{aligned} \tag{13}$$

In the following, we consider the first term and have

$$\begin{aligned} |\omega_{Q_{tuning},K} - \omega_{Q,K}| &= |\|M^i_{Q_{tuning},K}\|_2^2 - \|M^i_{Q,K}\|_2^2| \\ &= |(M^i_{Q_{tuning},K} - M^i_{Q,K})^T (M^i_{Q_{tuning},K} + M^i_{Q,K})| \\ &\leq \|M^i_{Q_{tuning},K} - M^i_{Q,K}\|_2 \|M^i_{Q_{tuning},K} + M^i_{Q,K}\|_2 \\ &\leq 2\|M^i_{Q_{tuning},K} - M^i_{Q,K}\|_2 \\ &= 2\|\mathrm{softmax}(Q^i_{tuning} K^T/\sqrt{d}) - \mathrm{softmax}(Q^i K^T/\sqrt{d})\|_2. \end{aligned} \tag{14}$$

By Lemma 1, softmax is a Lipschitz function under $2-$norm, thus we have

$$\begin{aligned} |\omega_{Q_{tuning},K} - \omega_{Q,K}| &\leq \frac{2}{\sqrt{d}} \|\frac{Q^i_{tuning} K^T}{\sqrt{d}} - \frac{Q^i K^T}{\sqrt{d}}\|_2 \\ &\leq \frac{2}{d} \|Q^i_{tuning} - Q^i\|_2 \|K\|_2 \\ &= \frac{2}{d} \|\hat{z}^i W_q^{tuning} - \hat{z}^i W_q\|_2 \|K\|_2 \\ &= \frac{2}{d} \|W_q^{tuning} - W_q\|_2 \|\hat{z}^i\|_2 \|K\|_2. \end{aligned} \tag{15}$$

In summary, we can derive

$$\|\Sigma'^{tuning}_{rv(z)} - \Sigma'_{rv(z)}\| \leq \frac{2}{d} \|W_q^{tuning} - W_q\|_2 \|\hat{z}^i\|_2 \|K\|_2 \|W_L^T \Sigma_{rv(V)} W_L\|. \tag{16}$$

□

Remark 3. *Because W_q^{tuning} is optimized iteratively from W_q, which is well initialized from pre-trained LDM to have a small gradient value under the diffusion loss, and*

$$\|W_q^{tuning} - W_q\|_2^2 < \mathrm{Trace}((W_q^{tuning} - W_q)(W_q^{tuning} - W_q)^T) = \|W_q^{tuning} - W_q\|_F^2, \tag{17}$$

it is intuitive that under a small learning rate, this term will have a small value.

A.4 Covariate shift for appending Temporal Attention Module

For a batch of temporal data $z \in \mathbb{R}^{(bl)\times n\times d}$, Temporal Attention (*TA*) Module has the following form

$$\text{Transformer}(z) = z + Linear(\text{Attention}(norm(z), norm(z)). \tag{18}$$

Considering a temporal feature $z \in \mathbb{R}^{n\times d}$, following Proposition 1 the impact on mean value $\mu'_{rv(z)} - \mu_{rv(z)} = W_L^T W_v^T \mu_{rv(\hat{c})} + b_L$, and the impact on covariance $\Sigma'_{rv(z)} - \Sigma_{rv(z)} = \omega_{Q,K} W_L^T \Sigma_{rv(V)} W_L$.

A.5 Covariate shift for appending Shift-restricted Temporal Attention Module

For a batch of temporal data $z \in \mathbb{R}^{(bl)\times n\times d}$, STAM has the following form

$$\text{Transformer}(z) = z + \bar{Linear}(\bar{\text{Attention}}(IC(z), IC(z)). \tag{19}$$

Considering a sample $z \in \mathbb{R}^{n\times d}$, the impact on mean value $\mu'_{rv(z)} - \mu_{rv(z)} = W_L^T W_v^T \mu_{rv(\hat{c})} + b_L$, and the impact on covariance $\Sigma'_{rv(z)} - \Sigma_{rv(z)} = \omega_{Q,K} W_L^T \Sigma_{rv(V)} W_L$. However, due to $\mu_{rv(\hat{c})} = 0$, we have $\mu'_{rv(z)} - \mu_{rv(z)} = b_L$. The covariance shift is illustrated by the following proposition.

Proposition 3. *Along with notations in Proposition 1, given* $\Sigma'_{rv(z)} = \Sigma_{rv(z)} + \omega_{Q,K} W_L^T \Sigma_{rv(V)} W_L$, *we have*

$$\|\Sigma'_{rv(z)} - \Sigma_{rv(z)}\| \le \|W_L\|^2 \|W_v\|^2 \|\Sigma_{rv(\hat{z})}\|. \tag{20}$$

Proof. Since $\sum_{j=1} M_{Q,K}^{i,j} = 1$, and $M_{Q,K}^{i,j} \ge 0$, we can derive $0 \le \omega_{Q,K} \le 1$. Thus

$$\begin{aligned} \|\Sigma'_{rv(z)} - \Sigma_{rv(z)}\| &= \|\omega_{Q,K} W_L^T \Sigma_{rv(V)} W_L\| \\ &\le \|W_L^T \Sigma_{rv(V)} W_L\| \\ &\le \|W_L\|^2 \|\Sigma_{rv(V)}\| \\ &\le \|W_L\|^2 \|W_v\|^2 \|\Sigma_{rv(\hat{z})}\|. \end{aligned} \tag{21}$$

□

In STAM, since $\Sigma_{rv(\hat{z})} = \Sigma_{rv(z)}$, $\|W_L\| = 1$ and $\|W_v\| = 1$, thus the proposition shows that $\|\Sigma'_{rv(z)} - \Sigma_{rv(z)}\| \le \|\Sigma_{rv(z)}\|$. By experiments, the most variances are small values around $1e-2$, hence STAM indeed alleviates the covariate shift compared with *TA* Module.

B More Results

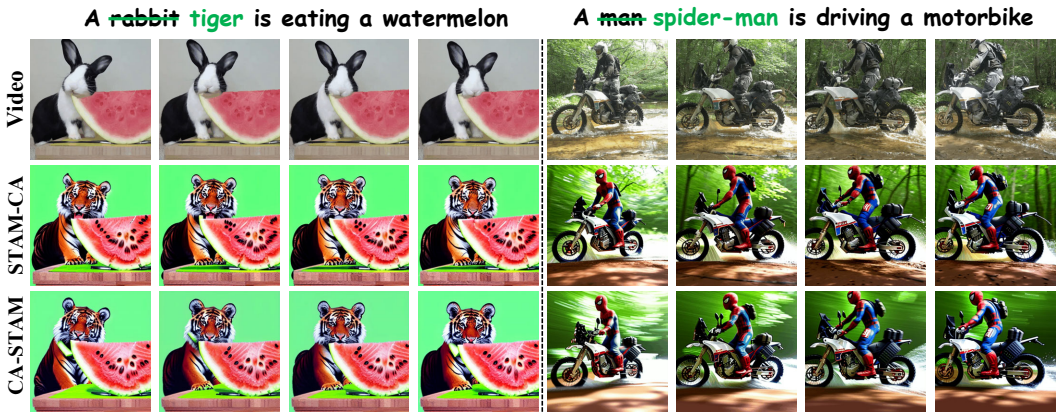

Figure 1: Inserting STAM before (STAM-*CA*) or after (*CA*-STAM) *CA* Module. The examples show that different orders have subtle visual effects on the results.

B.1 Position for inserting Temporal Attention Module

We note that *Tune-A-Video* appends *TA* Module after CA Module, where the module sequence is *SCA* Module—*CA* Module—*TA* Module. Our EI^2 inserts STAM before CA Module, where the module sequence is FFAM—STAM—*CA* Module. This sequence is more suitable for the theoretical analysis of STAM. In practice, the sequence FFAM—*CA* Module—STAM for inflating LDM also performs well in experiments, and we do not observe an obvious visual advantage from reordering the module sequence. Figure 1 depicts a few examples.

B.2 More qualitative results

For comprehensive analysis and evaluation, we have supplied a video in the supplementary materials, which contains the ***Qualitative Comparison***, ***Ablation Study*** and ***Combination of EI^2 and P2P*** [3]. It can be observed that STAM and FFAM facilitate EI^2 to surpass previous state-of-the-art methods in terms of textual alignment and temporal consistency. Incorporating P2P [3] with EI^2 can further improve the stability and structure preservation of results.

References

[1] Ba, J., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv: 1607.06450 (2016)

[2] Gao, B., Pavel, L.: On the properties of the softmax function with application in game theory and reinforcement learning. ArXiv **abs/1704.00805** (2017)

[3] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv: 2208.01626 (2022)

[4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)