Video Harmonization with Triplet Spatio-Temporal Variation Patterns

Supplementary Material

In this supplementary material, we provide details about our Triplet Transformer architecture and the user study, analyze the boundaries between short-term (ST) and long-term (LT), as well as present additional visual comparison results. Furthermore, we discuss the limitation and societal impact of our work.

8. Triplet Transformer Architecture Details

The detailed architecture of our Triplet Transformer is illustrated in Figure 8, comprising Short-Term Spatial Transformer (ST-ST), Long-Term Global Transformer (LT-GT), and Long-Term Dynamic Transformer (LT-DT) modules. Both our ST-ST and LT-GT modules perform Transformer self-attention calculations separately within the normal and shifted windows [29, 30], instead of across all frame tokens.

LT-GT with Auto-regressive. As shown in Table 3, we introduce three auto-regressive strategies to explore their ability to capture long-term temporal variation patterns in our LT-GT module. In the vanilla iGPT's auto-regressive design [3], the current token z_i is limited to conducting self-attention calculations solely with the tokens preceding it to predict its features, encompassing both semantic and appearance. The calculation process of z_i is as follows:

$$z_i = \text{TR}(z_1, z_2, ..., z_{i-1}), \tag{11}$$

where TR denotes Transformer with vanilla self-attention.

To promote self-attention with an auto-regressive focus on learning temporal appearance patterns rather than predicting semantic features, we enable the current token z_i to conduct self-attention with the tokens preceding it, named F-D. The calculation process of z_i in F-D is as follows:

$$z_i = \text{TR}(z_1, z_2, ..., z_{i-1}, z_i). \tag{12}$$

We further consider the reverse temporal trends and introduce two bidirectional temporal auto-regressive strategies, where, one utilizes a shared Transformer (Bi-D) and the other employs a separate Transformer (FB-D). The calculation process for these strategies is as follows:

$$z_i^{Bi-D} = C[\text{TR}(z_1, ..., z_i), \text{TR}(z_i, ..., z_{T-1}, z_T)], \tag{13}$$

$$z_i^{FB-D} = C[\text{TR}(z_1, ..., z_i), \text{TR}'(z_i, ..., z_{T-1}, z_T)], \tag{14}$$

where C denotes concatenate operation, and T represents the total number of input frames.

LT-GT with Masked Prediction. We also design different masking strategies for masked prediction within our LT-GT module, drawing inspiration from BERT [9] and MAE [16]. The calculation process is as follows:

$$z_i = \text{TR}(z_{[1,T]\setminus M}), \tag{15}$$

where M denotes masked tokens, randomly sampled from the sequence $[1, T]$ with a certain probability of inclusion.

As shown in Table 3, the self-window masking (MS) represents masking the current token during self-attention computation, *i.e.*, $z_i = \text{TR}(z_{[1,T]\setminus z_i})$. The self-window and random 50% masking (MS&50) represent masking the current token and 50% of all tokens. M50, M75, and M90 indicate random masking 50%, 75%, and 90% of all tokens, respectively.

9. User Study Details

As shown in Table 5, we conduct user studies on the HYouTube dataset and Real Composite Videos [31] using our custom platform. These studies analyze the visual quality of harmonized videos and the effectiveness of our temporal consistency evaluation metrics, *i.e.*, R-RTC and fR-RTC. Our custom platform page is illustrated in Figure 9.

For HYouTube dataset, we randomly select 20 groups of videos from the testing dataset. Each group consists of a composite video, a real video, and harmonized videos produced by three different methods (our VHTT, CO$_2$Net [31] and HT+ [14]). We invite users to choose the harmonized video that is most similar to the real video. Additionally, users rate the degree of flickering in each harmonized video on a 5-point Likert scale, with higher scores indicating more pronounced flickering.

For Real Composite Videos, we randomly select 20 groups of videos. Each group consists of a composite video and harmonized videos produced by three different methods. We invite users to select the visually best among the three harmonized videos and rate their degree of flickering.

Finally, we invite 50 users to participate in these studies and ask each user to complete all evaluations, and then we collect 2000 groups of comparisons for further analysis.

Following [35], we separately count the number of times each of the three harmonization methods is selected on HYouTube and Real Composite Videos, and then calculate their respective percentages (higher is better), *i.e.*, "Times" in Table 5. We further calculate each method's average flickering degree score based on the harmonized videos they correspond to, obtaining the average score for each method (lower is better), *i.e.*, "Degree" in Table 5.

10. The boundaries between ST and LT.

Based on our motivation, we set ST as 2-frame to leverage subtle temporal changes for better spatial harmonization, and set LT as 5-frame during training based on computational practicality while maximizing LT frames during inference phase for richer temporal information. Table 8

Training	**S2&L5**		S1&L5	S3&L5	S2&L3	S2&L4	S2&L8
Inference	**S2&L20**	S2&L5	S1&L20	S3&L20	S2&L20		
fMSE↓	**90.35**	100.19	98.27	97.58	100.61	92.37	90.51
fR-RTC↓	**1.26**	3.68	1.43	1.27	1.38	1.44	1.28

Table 8. Comparison of short-term (ST) and long-term (LT) in different settings. "Sx&Ly" means ST with x and LT with y frames.

demonstrates that: ST with 2-frame surpasses 1-frame and 3-frame (column: 2 vs. 4-5), because 1-frame lacks temporal variations and 3-frame overly focuses on temporal features; using more frames for LT during training is beneficial (2 vs. 6-7), but there may be an upper limit or insufficient training (2 vs. 8), indicating further study required.

11. Additional Visual Comparison Results

Additional visual comparison results are in our GitHub repository[1]. This video includes results of video harmonization on both the HYouTube dataset and Real Composite Videos, failure cases, as well as video enhancement and video demoiréing.

12. Limitation

We show failure cases in our GitHub repository[1], presenting the limitation of our method for harmonizing videos with rapid moving scenes. We will address this issue by delving deeper into modeling spatio-temporal variation patterns with foundation models and richer datasets.

13. Societal Impact

Developments in visual generative models including video harmonization models, offer new applications and creative workflows, while pose risks of misuse for producing deceptive content. It is crucial to manage and regulate the use of such models.

[1] https://github.com/zhenglab/VideoTripletTransformer

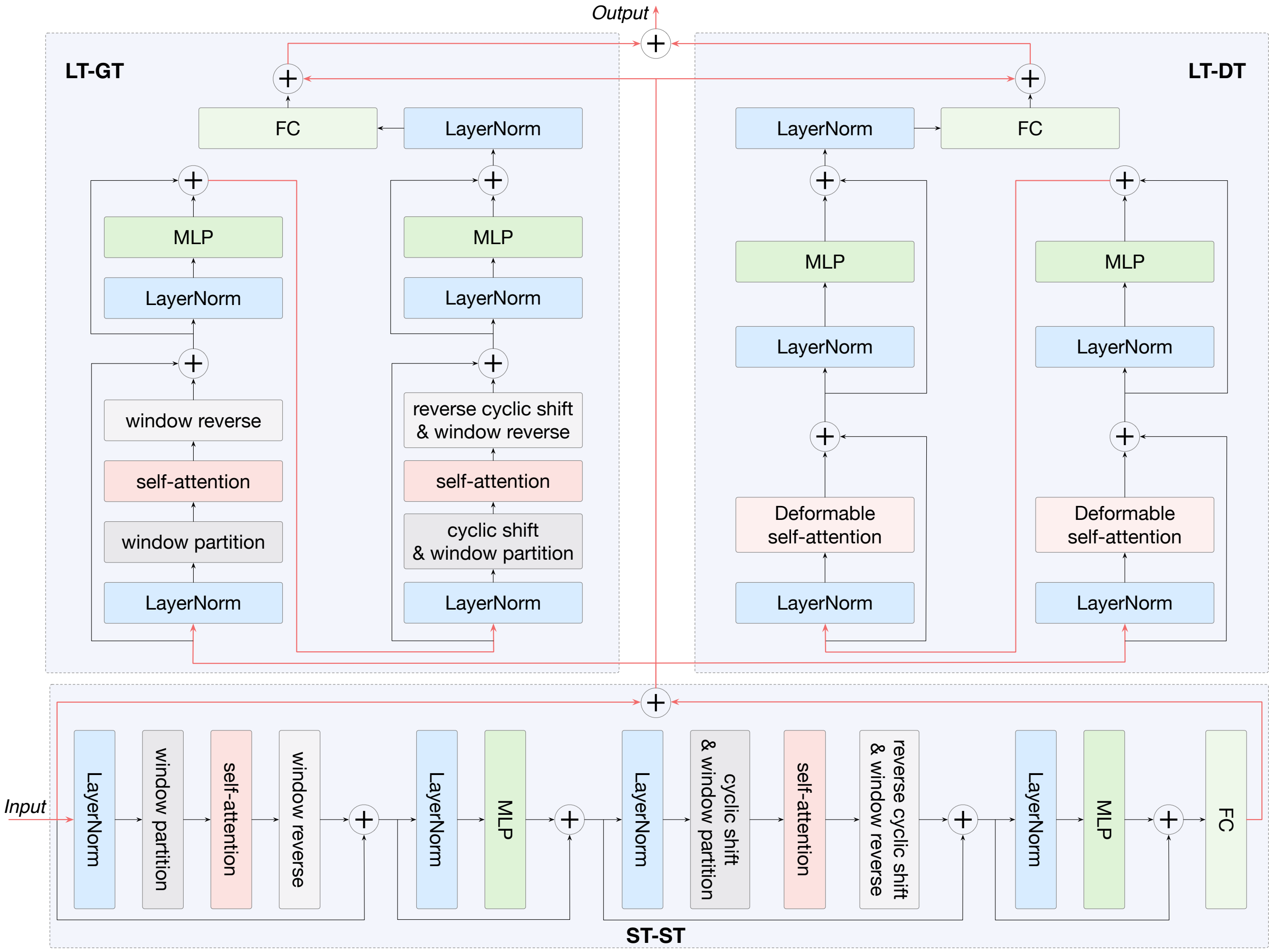
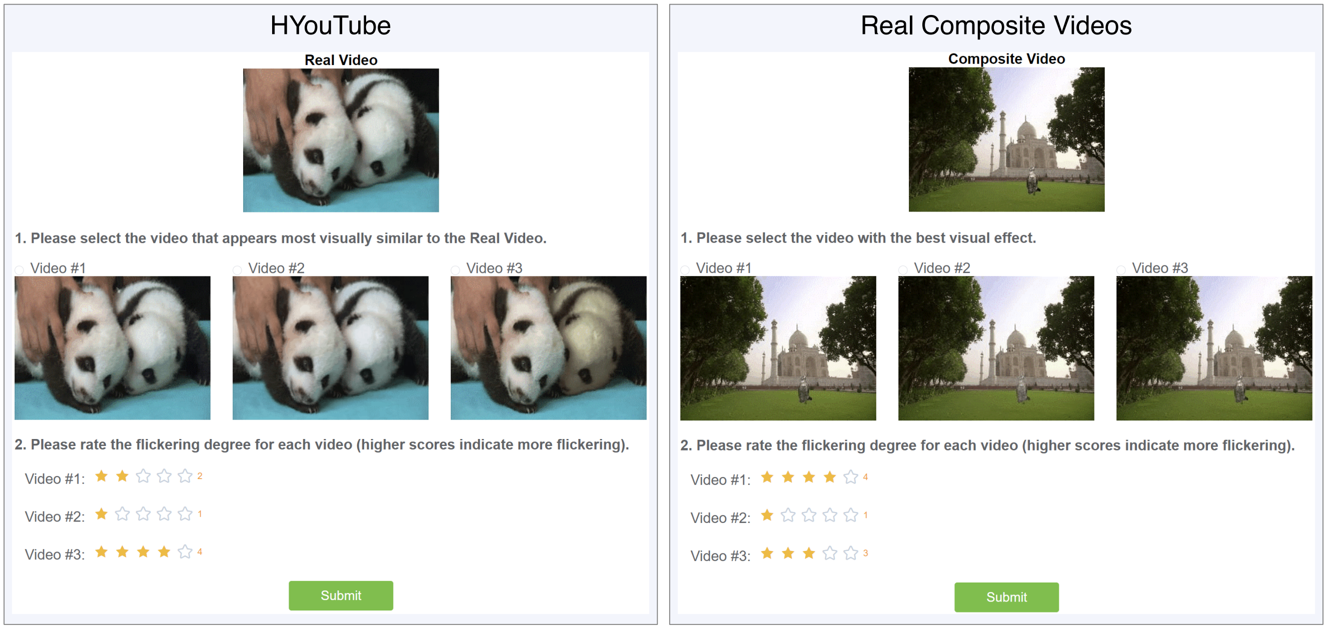


Figure 8. The architecture details of our Video Triplet Transformer, including Short-Term Spatial Transformer (ST-ST), Long-Term Global Transformer (LT-GT), and Long-Term Dynamic Transformer (LT-DT).



HYouTube

Real Video

1. Please select the video that appears most visually similar to the Real Video.

○ Video #1 ○ Video #2 ○ Video #3

2. Please rate the flickering degree for each video (higher scores indicate more flickering).

Video #1: ★★☆☆☆ 2

Video #2: ★☆☆☆☆ 1

Video #3: ★★★★☆ 4

Submit

Real Composite Videos

Composite Video

1. Please select the video with the best visual effect.

○ Video #1 ○ Video #2 ○ Video #3

2. Please rate the flickering degree for each video (higher scores indicate more flickering).

Video #1: ★★★★☆ 4

Video #2: ★☆☆☆☆ 1

Video #3: ★★★☆☆ 3

Submit

Figure 9. Our custom user study platform pages for both the HYouTube dataset (left) and Real Composite Videos (right) [31].