Diff-Tracker: Text-to-Image Diffusion Models are Unsupervised Trackers (Supplementary Material)

Zhengbo Zhang[1], Li Xu[1], Duo Peng[1], Hossein Rahmani[2], and Jun Liu[1] *

[1] Singapore University of Technology and Design
{zhengbo_zhang, li_xu, duo_peng}@mymail.sutd.edu.sg, jun_liu@sutd.edu.sg
[2] Lancaster University
h.rahmani@lancaster.ac.uk

1 More Analysis and Discussion of Results

1.1 Main results

As shown in Tab. 1&2 of the main paper, we conduct abundant experiments on 5 datasets and our method achieves SOTA performance on all datasets. This success can be attributed to that our method leverages the rich knowledge stored in the pre-trained diffusion model, including the understanding of semantic and structural information of video frames. As shown in Tab. 2 of the main paper, our method shows a significant accuracy improvement on the long video dataset LaSOT. A possible reason is that we use target's motion information for online updating, facilitating effective and continuous tracking, which is crucial for the long-term tracking task.

1.2 Ablation study.

As shown in Tab. 3 of the main paper, when using attention harmonization, the performance of our method improves. This is because attention harmonization allows our method to leverage the relationship between target and its background for tracking. Moreover, as shown in Tab. 4 of the main paper, utilizing short-term motion information for online updating improves our method's performance. This is because the short-term motion information effectively captures target's immediate motion, providing timely data for prompt online updating. Besides, incorporating long-term motion information brings greater performance improvements to our method. This enhancement is attributed to the stronger spatio-temporal continuity of long-term motion data compared to short-term motion data.

* corresponding author

2 Additional Ablation Studies

2.1 Number of frames input to the motion encoder.

We explore the number of video frames in the sequence input to the motion encoder for extracting long-term motion information (see Fig. 1). We believe that the reason why a greater number of input frames leads to better tracking performance might be that the long-term motion information obtained from a longer video sequence is more accurate. However, once the number of input frames exceeds six, the performance improvement is trivial. Therefore, we set the number of frames input into the motion encoder to six.

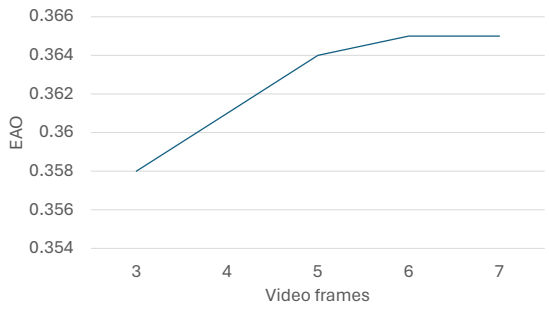

Fig. 1: Ablation study for impact of numbers of input frames. We report the EAO scores from the VOT2018 benchmark dataset as the evaluation metric.

2.2 Hyper parameters α and β.

We also validate the sensitivity our hyper parameters α and β. As shown in Fig. 2 and Fig. 3, the optimal values for α and β are 0.5 and 0.7, respectively.

3 Visualization Results

As discussed in our main paper, the cross-attention maps obtained from the pre-trained text-to-image diffusion models can showcase the underlying text-vision correspondences. We present examples of cross-attention maps derived from the text-to-image diffusion models (see Fig. 4), including the Stable Diffusion [2] and DALL-E mini [1]. Fig. 4 illustrates the process where both the image and corresponding text are fed into the text-to-image diffusion models. We can observe that the activated areas on the cross-attention maps are related to the textual meanings, affirming the capability of text-to-image diffusion models to capture and represent the associations between textual prompts and visual elements effectively.

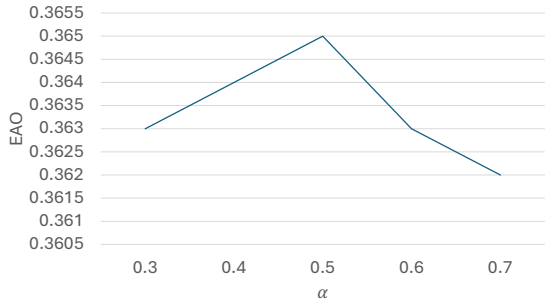

Fig. 2: Ablation study for impact of the hyper parameter α. We report the EAO scores from the VOT2018 benchmark dataset as the evaluation metric.

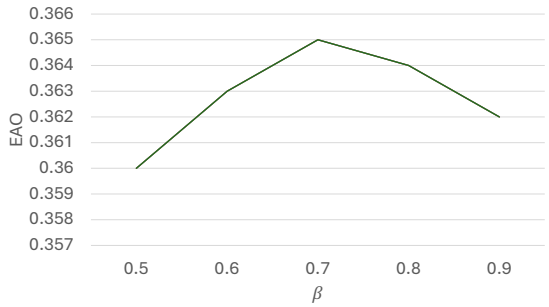

Fig. 3: Ablation study for impact of the hyper parameter β. We report the EAO scores from the VOT2018 benchmark dataset as the evaluation metric.

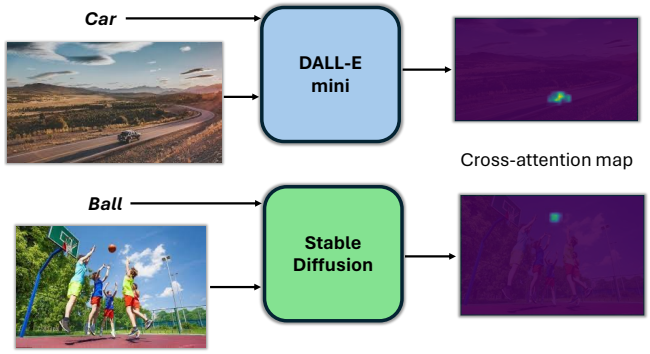

Fig. 4: Visualization of the text-vision correspondences across different text-to-image diffusion models, specifically the Stable Diffusion [2] and DALL-E mini [1].

References

1. Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., et al: Dall·e mini (7 2021). https://doi.org/10.5281/zenodo.5146400, https://github.com/

borisdayma/dalle-mini 2, 3
2. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 3