Supplementary Material of "Exploring Pre-trained Text-to-Video Diffusion Models for Referring Video Object Segmentation"

In the following sections, we first introduce the additional implementation details, including various losses and the matching process. Then we provide more experimental results and visualization results.

A Additional Implementation Details

Additional Training Details. In the pre-training phase on Ref-COCO [10], we initiate with distinct learning rates for different components: 2.5e-6 for the text encoder, and 2.5e-5 for the remaining parts of the model. This stage spans 12 epochs, focusing on single-frame inputs. The learning rates undergo a reduction by a factor of 10 at the 8-th and 10-th epochs to refine the training process. In our main training phase, we adjust the training length to 6 or 9 epochs, depending on whether we use pre-training or not. The text encoder is kept frozen during training. Initial learning rates are set at 5e-5 for the whole model. For a 9-epoch training setting, learning rates are decreased by a factor of 10 at the 6-th and 8-th epochs. In contrast, for a 6-epoch schedule, reductions occur at the 3-rd and 5-th epochs.

Additional Training Loss Details. To ensure a fair comparison, we utilize the identical training losses and weights that are mentioned in [1, 5, 8]. Therefore, our segmentation consists of three sub-heads: bounding box head, classification head and mask head. These correspond to three types of outputs: bounding boxes $\mathcal{B}$, classification scores $\mathcal{S}$, and segmentation masks, which include both a low-resolution $\mathcal{M}_o$ and a refined high-resolution $\mathcal{M}$. The overall training loss equation used to optimize these outputs is

$$\mathcal{L}_{\text{train}} = \lambda_{\mathcal{M}_o}\mathcal{L}_{\mathcal{M}_o} + \lambda_{\mathcal{M}}\mathcal{L}_{\mathcal{M}} + \lambda_{\mathcal{B}}\mathcal{L}_{\mathcal{B}} + \lambda_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}, \tag{1}$$

where $\mathcal{L}$ represents the different loss components and λ denotes their respective weights. Specifically, $\mathcal{L}_{\mathcal{M}}$ and $\mathcal{L}_{\mathcal{M}_o}$ are mask losses. $\mathcal{L}_{\mathcal{B}}$ is bounding box loss and $\mathcal{L}_{\mathcal{S}}$ is classification loss. Specifically, we use Dice loss [3] and Focal loss [4] for masks $\{\mathcal{M}_o, \mathcal{M}\}$, Focal loss [4] for confidence scores $\mathcal{S}$, and L1 and GIoU [6] loss for bounding boxes $\mathcal{B}$.

Additional Matching Loss Details. Our approach to instance matching is in line with the recent transformer-based methods [1, 5, 7–9]. In particular, we adopt the Hungarian algorithm [2] to select the most suitable match between the $Q = 5$ instance queries and the ground truth. For this purpose, final masks $\mathcal{M}$, bounding boxes $\mathcal{B}$ and classification scores $\mathcal{S}$ are used to compute the matching loss $\mathcal{L}_{\text{match}}$ for each query and Hungarian algorithm is used to find the best match that has the minimum loss. $\mathcal{L}_{\text{match}}$ lies in three parts, which can be formulated as

$$\mathcal{L}_{\text{match}} = \lambda_{\mathcal{M}}\mathcal{L}_{\mathcal{M}} + \lambda_{\mathcal{B}}\mathcal{L}_{\mathcal{B}} + \lambda_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}. \tag{2}$$

Table A: Ablation study on the effectiveness of attention mechanism in Text-Guided Image Projection. "Attn," denotes the attention mechanism. "Concat." denotes concatenation. All results are reported on Ref-YouTube-VOS.

Method	Fusion	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
VD-I	-	59.7	57.9	61.6
VD-T	-	61.9	60.1	63.7
VD-IT	Concat.	62.4	60.8	64.0
VD-IT	Attn.	**64.8**	**63.1**	**66.6**

Table B: Ablation study on the multi-step diffusion process. All results are reported on Ref-YouTube-VOS.

Step	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	**64.8**	**63.1**	**66.6**
5	63.4	61.6	65.1
10	63.5	61.7	65.1
50	63.1	61.3	64.9
100	62.7	60.8	64.5

B Additional Experiment Results

Text-Guided Image Projection. In our Text-Guided Image Projection, we propose leveraging both referring text and visual tokens from each frame to guide the T2V model in producing the latent feature, rather than solely relying on visual tokens. This approach not only ensures visual feature consistency across time aiding temporal instance matching, but also enriches feature detail for better spatial differentiation. In Equation 1 presented in the paper, the referring text tokens employ the attention mechanism to guide the visual tokens to get final prompt. To demonstrate the effectiveness of our design, we contrast our approach with a method that concatenates referring text tokens and visual tokens as the prompt. It is noteworthy that after the concatenation operation, we follow up with an MLP to ensure the number of trainable parameters is consistent with those in the attention mechanism for a fair comparison. Table A reveals that mere concatenation results in the inability of U-Net to effectively utilize the details of visual tokens, hindering the extraction of fine-grained visual features.

Noise Level. As discussed in Section 4.1 of the paper, to extract visual features using the denoising U-Net, it is essential to add noise to the video signal. However, the optimal level of noise strength is unknown. Specifically, in the forward process of the diffusion model, the level of noise is represented by steps; the larger the step, the greater the intensity of the noise. Table B reports that when the step is growing, the performance is dropping. We believe that higher noise levels might obscure the original signal, resulting in less detailed visual features, which in turn leads to poorer performance.

C Additional Visualization Results

To further highlight the benefits of using the T2V diffusion model as the visual encoder over V-Swin, we showcase a comparison between our VD-IT and several V-Swin based methods in Figure A. VD-IT demonstrates superior temporal consistency in tracking individuals, particularly when the person is extensively obscured. This enhancement becomes especially evident in challenging scenarios where maintaining continuity and accuracy in tracking is critical, highlighting VD-IT's robustness in handling complex visual occlusions.

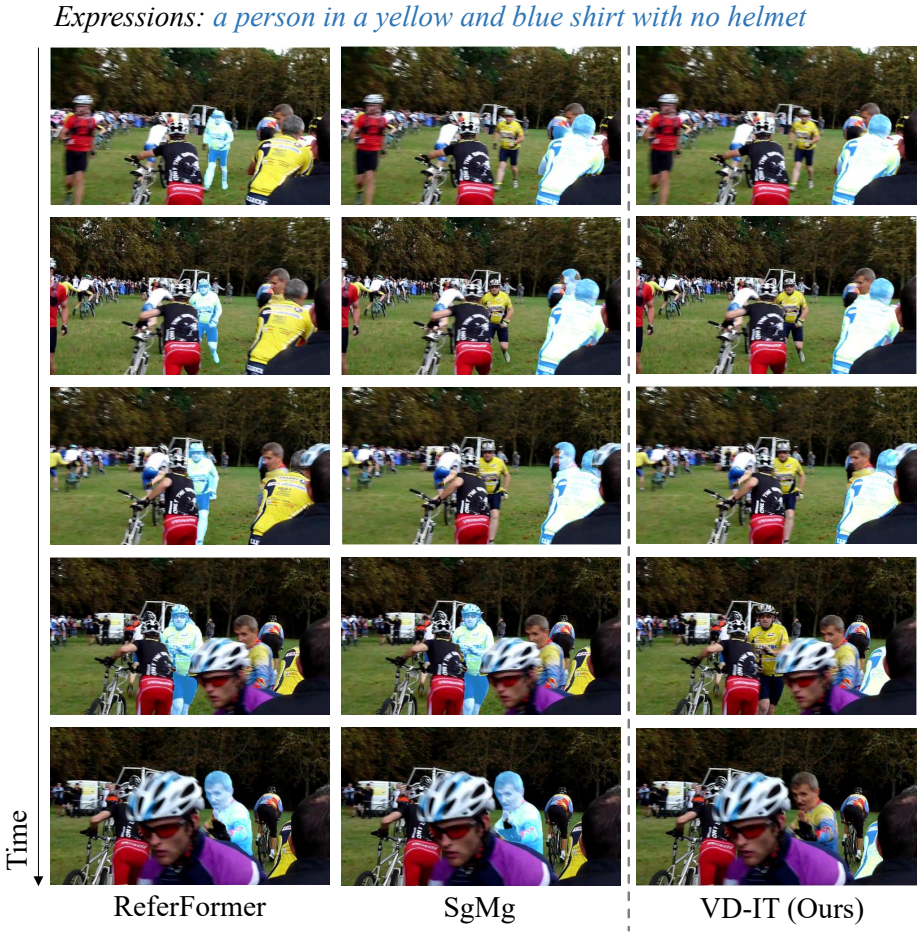

Fig. A: Qualitative comparison of our method with others. Both ReferFormer [8] and SgMg [5] use V-Swin as the visual encoder.

References

1. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4985–4995 (2022) 1
2. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics **52** (1955) 1
3. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. arXiv preprint arXiv:1911.02855 (2019) 1
4. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017) 1
5. Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Spectrum-guided multi-granularity referring video object segmentation. In: ICCV. pp. 920–930 (2023) 1, 3

6. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–666 (2019) 1
7. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR. pp. 8741–8750 (2021) 1
8. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022) 1, 3
9. Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. In: ECCV. pp. 588–605. Springer (2022) 1
10. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85. Springer (2016) 1