DreamTrack: Dreaming the Future for Multimodal Visual Object Tracking

Supplementary Material

The supplementary material presents additional details and analyses of our model.

- **More Challenging Benchmarks**
  We evaluate the proposed method in more challenging benchmarks, including VOT [8–10] and large-scale long-term datasets (*e.g.,* VastTrack [13]).
- **Generalization in Different Paradigms**
  We show the generalization of our method in both CNN and Transformer-based frameworks.
- **Different Prediction Modalities**
  We compare different modalities in multimodal prediction to handle the uncertainty of future forecasting.
- **Different Dreaming Steps**
  We explore the influence of dreaming steps to enjoy more bonus from the temporal learning of future dreaming.
- **Position of Future Dreaming**
  We compare different positions of future dreaming to achieve a better understanding of the environment.
- **Ground-Truth Future States**
  We explore the gap between the dreamed future and ground-truth future by inference with the extracted features of real future frames.
- **Update of Target Query**
  We evaluate the performance without the historical information by freezing the target queries in inference.
- **Reconstruction of Template and Search Region**
  We ablate the reconstruction of the template and search region to show the influence in future dreaming.
- **Scale of Training Data**
  We explore the influence of training data volume to show the effectiveness of our method with data-driven learning.
- **Attribute Results on LaSOT**
  We detail the performance of the video sequences in LaSOT which are labeled with specific attributes.
- **More Visualization Results**
  We provide more visualization results of the attention maps and dreamed future frames.

More Challenging Benchmarks. With the proposed future dreaming module to benefit from both historical and future information, our DreamTrack is capable of generalizing in new tracking scenarios. We further demonstrate this by evaluating on more challenging benchmarks, including VOT2018LT [8], VOT2020 [9], VOT2022 [10], OxUvA [14], TLP [12] and VastTrack [13]. The results are listed in Tab. 1. It shows that our method still achieves outstanding performance under more complex tracking environment, proving its effectiveness. Notably, we also apply LoRA [6] to our DreamTrack$_{256}$, which aims to improve model efficiency with trainable rank decomposition matrices. The resulted tracker DreamTrack-LoRA$_{256}$ performs better with much faster inference speed (217FPS), which achieves balanced performance and efficiency. Besides, we employ SAM [7] as the post-processing to output segmentation masks of tracking target for VOT evaluation [8–10]. The superior performance of DreamTrack$_{256}$+SAM further proves the effectiveness of our design to provide accurate localization of the tracking target.

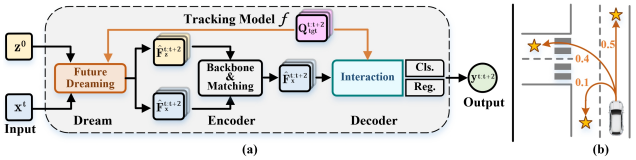

Figure 1. **(a)** Generality of our framework on both CNN- and transformer-based paradigms. **(b)** "multimodality" in the prediction module, which indicates multiple possible motion behaviors.

Generalization in Different Paradigms. Our method is general to be applied to both CNN and Transformer-based paradigms. As shown in Fig. 1(b), learnable queries $\mathbf{Q}_{\text{tgt}}^{t:t+2}$ cache temporal dynamics by interacting with image features in the **dream** and **decoder** stages, in which cross-attention (Transformer-wise) can be replaced by cross-correlation (CNN-wise). The **encoder** stage can be constructed with CNN backbone and correlation. We further demonstrate this by applying our design to a CNN tracker SiamCAR [5]. The resulted tracker DreamCAR$_{255}$ in Tab. 1 achieves superior performance in various challenging benchmarks, which proves the effectiveness and generalization of our method.

Different Prediction Modalities. The "multimodality" in our methodology refers to different possibilities of the target's future behavior instead of inputs from different sensors ("**Decoder**" of Sec. 3.2 in the main text). As shown in Fig. 1(a), the car could go straight or turn left/back. We introduce this concept into visual object tracking, and propose the Multimodal Prediction module to handle the high degree of uncertainty in predicting the target's future motion. In our design, three possible future situations are included in the prediction, *i.e.,* the modality of 3. We then explore the influence of different prediction modalities, as illustrated in Tab. 2. The results show that more modalities help to model different motion modes of the target in the future, which leads to superior performance (③ *v.s.*② *v.s.*①). It's worth noting that the prediction modality of 4 shows no obvious gains compared with the one of 3 modalities (④ *v.s.*③). This

Method	VOT2018LT			VOT2020			VOT2022			OxUvA			TLP		VastTrack		FPS
	P	R	F	A	R	EAO	A	R	EAO	TPR	TNR	MaxGM	SUC	P	SUC	P	
OSTrack$_{256}$ [16]	0.556	0.613	0.579	0.464	0.779	0.301	0.778	0.802	0.514	0.847	0.822	0.835	57.6	62.7	33.6	31.5	105
LoRAT$_{224}$ [11]	0.584	0.646	0.610	0.469	0.804	0.313	0.796	0.814	0.532	0.888	0.856	0.872	61.1	64.1	39.3	40.8	209
ARTrackV2$_{256}$ [1]	0.592	0.654	0.622	0.475	0.814	0.318	0.803	0.821	0.542	0.893	0.871	0.882	60.7	64.5	40.9	41.5	94
DreamTrack$_{256}$	0.610	0.677	0.642	0.487	**0.833**	0.335	0.825	0.837	0.561	0.920	0.887	0.903	63.3	66.7	42.7	**43.8**	139
DreamTrack-LoRA$_{256}$	**0.617**	**0.680**	**0.645**	**0.492**	0.831	**0.337**	**0.830**	**0.841**	**0.565**	**0.924**	**0.890**	**0.907**	**64.1**	**67.1**	**42.9**	43.7	**217**
ARTrackV2$_{256}$ [1]+SAM [7]	0.714	0.687	0.705	0.767	0.874	0.606	0.837	0.867	0.623	-	-	-	-	-	-	-	14
DreamTrack$_{256}$+SAM [7]	**0.740**	**0.711**	**0.728**	**0.789**	**0.908**	**0.631**	**0.850**	**0.896**	**0.649**	-	-	-	-	-	-	-	**19**
SiamCAR$_{255}$ [5]	0.442	0.481	0.461	0.414	0.711	0.235	0.691	0.673	0.374	0.494	0.521	0.507	31.3	32.7	27.0	25.3	**52**
DreamCAR$_{255}$	**0.475**	**0.503**	**0.488**	**0.437**	**0.744**	**0.266**	**0.712**	**0.733**	**0.416**	**0.519**	**0.538**	**0.528**	**33.9**	**35.5**	**31.1**	**30.6**	46

Table 1. State-of-the-art comparison on VOT2018LT [8], VOT2020 [9], VOT2022 [10], OxUvA [14], TLP [12] and VastTrack [13]. The number in the subscript denotes the search region resolution.

#	Prediction Modality	LaSOT [3]		LaSOText [4]		TNL2K [15]	
		SUC(%)	P(%)	SUC(%)	P(%)	SUC(%)	P(%)
①	1	73.1	79.0	51.9	58.4	59.5	61.1
②	2	73.6	79.8	52.8	59.5	60.2	62.5
③	3	**73.8**	**80.6**	**53.1**	59.8	**60.4**	**63.2**
④	4	73.8	80.5	53.0	**60.0**	60.3	62.9

Table 2. Ablation on the modality of the Multimodal Prediction.

#	Dreaming Step	LaSOT [3]			TNL2K [15]		FPS
		SUC(%)	P_{Norm}(%)	P(%)	SUC(%)	P(%)	
①	0	71.0	80.9	78.8	57.6	59.6	**155**
②	1	72.7	82.5	80.0	59.2	61.4	146
③	2	**73.8**	**83.4**	**80.6**	**60.4**	**63.2**	139
④	3	73.7	83.2	80.5	60.2	62.9	131
⑤	4	73.5	82.8	80.1	59.6	62.7	122

Table 3. Ablation on the dreaming steps.

indicates that three future predictions are sufficient for temporal dreaming in visual tracking, which is adopted as the default setting of our DreamTrack.

Different Dreaming Steps. In our default setting, the proposed DreamTrack predicts the future states of the next two frames, which aims to complement the temporal messages with future dynamics. Here we ablate different dreaming steps to explore the influence, as shown in Tab. 3. The version with the dreaming step of 0 only predicts the states of the current frame based on the historical observations (①), which in fact has no future information while showing comparable performance with recent SOTA trackers. When the dreaming step grows, the performance also improves and reaches the top with an SUC of 73.8% on LaSOT that dreams the next 2 frames (③). One interesting observation is that the dreaming steps of more than 2 cannot bring more performance gains but degrade the tracking accuracy. The underlying reason is that the uncertainty of predicting too far future has exceeded the capacity of the model, which distracts the tracking localization and leads to inferior performance (⑤,④ *v.s.* ③).

#	Architecture	LaSOT [3]			TNL2K [15]	
		SUC(%)	P_{Norm}(%)	P(%)	SUC(%)	P(%)
①	Dream-Encoder-Decoder	**73.8**	**83.4**	**80.6**	**60.4**	**63.2**
②	Encoder-Dream-Decoder	70.7	80.6	77.1	57.0	58.6

Table 4. Ablation on the position of the dream stage.

Position of Future Dreaming. Our DreamTrack has a dream-encoder-decoder architecture, which performs future dreaming before the interaction between the template and search region in the encoder. This order aims to preserve the information from original observations and learn the general environmental dynamics. We further explore the influence by postponing the dream stage after the encoder, *i.e.,* encoder-dream-decoder. Tab. 4 shows that late future dreaming achieves inferior performance compared with the default version (② *v.s.* ①). The underlying reason is that the encoder aims to filter target-irrelevant messages of the current frame by interacting with the template, leading to insufficient environmental dynamics for future dreaming. This also proves the effectiveness of our design.

Ground-truth Future States. As mentioned in Sec. 3.2 of the main text, we exploit the ground-truth (GT) future frames as the supervision label of the predicted future states, which empowers the tracker with the capability of future dreaming. Then what if using the GT future states instead of dreaming? We replace the dreamed future states with the features of GT future frame for inference, which align with the modeling process of the current frame. The results are presented in Tab. 5. It shows that the performance with GT future states surpasses the dreamed one for 1.8% SUC and 1.4% precision of LaSOT (② *v.s.* ①). This indicates that the discrepancies still remain between the dreamed future and the GT one. We leave it for further study to dream a more realistic future scenario.

Update of Target Query. With the input search region of each new frame, the target queries will first update the environmental dynamics with the current observation, which

#	Ground-Truth Future State	LaSOT [3]		LaSOText [4]		TNL2K [15]	
		SUC(%)	P(%)	SUC(%)	P(%)	SUC(%)	P(%)
①	✗	73.8	80.6	53.1	59.8	60.4	63.2
②	✓	**75.6**	**82.0**	**54.7**	**61.8**	**62.2**	**64.9**

Table 5. Ablation on the ground-truth future states.

#	Update of Target Query	LaSOT [3]		LaSOText [4]		TNL2K [15]	
		SUC(%)	P(%)	SUC(%)	P(%)	SUC(%)	P(%)
①	✗	71.4	79.1	51.2	58.2	59.0	60.4
②	✓	**73.8**	**80.6**	**53.1**	**59.8**	**60.4**	**63.2**

Table 6. Ablation on the update of target query.

#	Reconstruction		LaSOT [3]		TNL2K [15]		GPU Days for Training
	Template	Search Region	SUC(%)	P(%)	SUC(%)	P(%)	
①	✗	✗	72.7	79.8	59.1	61.7	**7.83**
②	✓	✗	73.1	80.3	59.8	62.4	8.74
③	✗	✓	73.8	80.6	60.4	63.2	9.57
④	✓	✓	**74.1**	**80.7**	**60.8**	**63.5**	10.66

Table 7. Ablation on the reconstruction of the template and search region in the Future Dreaming module.

are then interacted with encoded features of the search region to perform Multimodal Prediction. Here we explore the influence of updating target queries by freezing the parameters as initialization, *i.e.,* abandoning the historical information. As shown in Tab. 6, the performance degrades compared with the default DreamTrack$_{256}$ (① *v.s.* ②), proving the necessity of past experience for a better environmental understanding. Despite this, ① still demonstrates comparable tracking capability, showing the effectiveness of our design to infer the future with only current observation.

Reconstruction of Template and Search Region. As described in Sec. 3.2 of the main text, we only construct the search region based on the dreamed future states. Here we further explore the influence of reconstructing the template, and the results are presented in Tab. 7. It shows that the supervision of reconstruction helps learn the environmental dynamics to benefit visual tracking in our default DreamTrack$_{256}$ (③ *v.s.* ①). One interesting observation is that the reconstruction of the search region obtains superior performance compared with the one to reconstruct the template (③ *v.s.* ②). This indicates that learning the tracking scenario of the search region is more helpful in locating the target compared with dreaming the future template. Reconstructing both the template and search region achieves the best performance (④). Considering the increased training costs and little performance gains (④ *v.s.* ③), we take the model that only reconstructs the search region as the default version of our DreamTrack.

Scale of Training Data. The quality of temporal learning is deeply influenced by the training data volume. We then explore the influence by training our DreamTrack$_{256}$ with different scales of data and the results are presented in Tab. 8. The default setting for the scale of training data is noted as "100%". It shows that as the scale of training data reduces (*i.e.,* "75%", "50%" and "25%"), the overall performance gradually decreases (*e.g.,* $73.8\% \rightarrow 70.8\% \rightarrow 65.1\% \rightarrow 49.0\%$ of SUC on LaSOT), demonstrating that more data helps improve the model capacity of distinguishing the target in complex scenarios. Notably, even with only 50% of the data, our DreamTrack$_{256}$ has achieved comparable performance with early transformer-based trackers (*e.g.,* TransT [2]), proving the effectiveness of our design.

#	Data Volume	LaSOT [3]		LaSOText [4]		TNL2K [15]	
		SUC(%)	P(%)	SUC(%)	P(%)	SUC(%)	P(%)
①	25%	49.0	48.8	36.7	42.2	41.5	40.4
②	50%	65.1	69.8	45.3	53.4	51.3	52.5
③	75%	70.8	77.0	50.2	57.7	56.7	59.5
④	100%	**73.8**	**80.6**	**53.1**	**59.8**	**60.4**	**63.2**

Table 8. Ablation on the training data volume.

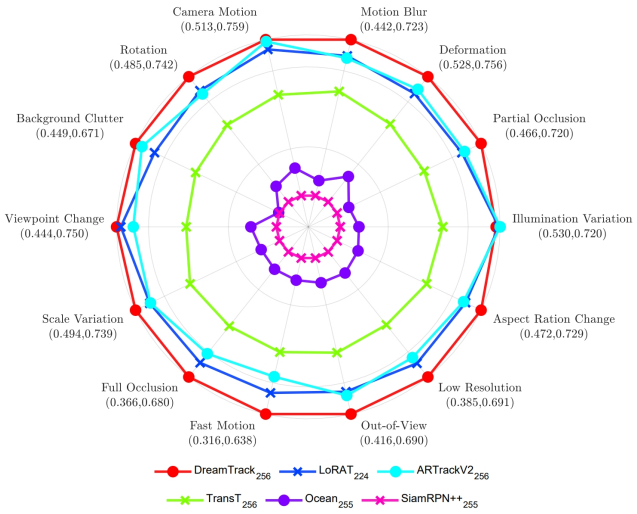

Figure 2. AUC scores of different attributes on LaSOT.

Attribute Results on LaSOT. With the offered specific attributes for each video sequence on LaSOT [3] (*e.g.,* Motion Blur, Deformation, Occlusion), we compare the performance under various tracking scenarios, as shown in Fig. 2. It shows that our DreamTrack$_{256}$ is more effective than other competing trackers on most attributes, particularly in handling scenarios involving Deformation, Full Occlusion and Partial Occlusion that raise critical challenges in long-term tracking. This proves the effectiveness of our future dreaming to improve tracking under complex scenarios with the learned environmental dynamics.

Figure 3. More attention visualization of our DreamTrack *w/.* and *w/o.* dreaming the future. The predictions and ground-truth boxes are marked in red and green, respectively.

Figure 4. More results visualization of different trackers and dreamed future frames. The comparison shows that our DreamTrack$_{256}$ could learn temporal dynamics of complex scenarios and perform robust tracking (e.g., deformation, occlusion and similar interferences), as well as future dreaming.

More Activation Analysis and Result Visualization. We introduce additional visualizations to demonstrate the enhanced generalization capability in novel situations. As depicted in Fig. 3, the activation maps without the Future Dreaming module hardly capture the target area under the circumstances of occlusion and interference from similar objects. This proves our claims that insufficient environmental understanding leads to biased target localization and error accumulation in long-term tracking. By comparison, the version with the proposed future dreaming forms focused attention on the target region even though some parts are occluded (*e.g.,* the second row). It demonstrates enhanced adaptability with our design in complex environments of new frames, which benefits the general perception of visual object tracking.

We also visualize more results of different trackers as well as the dreamed future frames under various tracking scenarios and object classes. As shown in Fig. 4, the proposed DreamTrack$_{256}$ still demonstrates superior distinguishability under appearance deformation, occlusion, background clutters and interference with similar objects. It demonstrates the effectiveness of our History-to-Future architecture to benefit tracking with the predicted future states based on the learned environmental dynamics from past experiences. The dreamed future frames correctly reflect the motion evolution of tracking scenarios, as well as the target appearance. Besides future forecasting of the subsequent frames, the predicted trajectory also successfully locates the tracking target of the dreamed frames. This further proves the effectiveness of our design to enhance the temporal learning with both history and future information and achieve generalized tracking with future dreaming.

References

[1] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *CVPR*, 2024. 2

[2] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 3

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 2, 3

[4] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Mingzhen Huang, Juehuan Liu, Yong Xu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *IJCV*, 2021. 2, 3

[5] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *CVPR*, 2020. 1, 2

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 1

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv*, 2023. 1, 2

[8] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *ECCV Workshops*, 2018. 1, 2

[9] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, et al. The eighth visual object tracking vot2020 challenge results. In *ECCV*. Springer, 2020. 1, 2

[10] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, et al. The tenth visual object tracking vot2022 challenge results. In *ECCV*. Springer, 2022. 1, 2

[11] Liting Lin, Heng Fan, Zhipeng Zhang, Yaowei Wang, Yong Xu, and Haibin Ling. Tracking meets lora: Faster training, larger model, stronger performance. *ECCV*, 2024. 2

[12] Abhinav Moudgil and Vineet Gandhi. Long-term visual object tracking benchmark. In *ACCV*. Springer, 2019. 1, 2

[13] Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. Vasttrack: Vast category visual object tracking. In *NeurIPS Datasets and Benchmarks Track*, 2024. 1, 2

[14] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *ECCV*, 2018. 1, 2

[15] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, 2021. 2, 3

[16] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 2