Click&Describe: Multimodal Aerial Grounding and Tracking

Rupanjali Kukal **Jay Patravali** **Fuxun Yu** **Simranjit Singh**
Nikolaos Karianakis **Rishi Madhok**
{rkukal, jaypatravali, fuxunyu, simsingh, nikolaos.karianakis, rishi.madhok}@microsoft.com
Microsoft

Abstract

The fusion of vision and language has driven progress in grounding and tracking tasks. However, aerial single-object tracking (SOT) has lagged in this domain due to a lack of text annotations in existing datasets. To address this, we provide text annotations for five aerial datasets, promoting multi-modal research in aerial tracking. Additionally, we introduce a third modality: **click** (or point prompt), offering a user-friendly alternative to bounding box annotations, enabling approximate target specification with less effort. We propose **CLaVi**, a multimodal framework that integrates click and language inputs, improving target localization and tracking efficiency. Our experiments on these datasets form the **AerTrack-460 benchmark**, which outperforms prior language-based methods, setting a strong baseline for future research.

1 Introduction

Single Object Tracking (SOT) in aerial imagery presents unique challenges, such as motion blur caused by mechanical vibrations of aerial vehicles and rapid changes in camera angles. These factors, combined with harsh lighting and weather conditions, significantly affect the appearance of the target. Traditional tracking methods rely on bounding box (BBox) annotations [1, 2, 9, 11, 13, 14], which are time-consuming [24] and difficult to create in aerial environments with small, occluded, or blurry objects.

Recent advances in natural language (NL) initialized tracking have shown promise by capturing global semantic information [8, 19, 27, 28], maintaining consistency amid frequent appearance changes. However, NL alone struggles in scenarios with clusters of similar-looking objects or tiny, unclear targets. Integrating an additional input modality, such as a click input, can address these challenges. Recent works [4, 5, 12] have trained neural networks to track points with fast inference but have not integrated point/click inputs with other modalities for single-object tracking. The click modality pinpoints object location despite appearance changes, while natural language provides global context for accurate predictions. Clicking once is more user-friendly and faster than drawing bounding boxes, allowing users to specify the target with minimal effort.

In this paper: (1) We provide text descriptions for five aerial datasets, collectively named AerTrack-460, to facilitate multimodal aerial tracking research; (2) We introduce CLaVi, a grounding and tracking framework that integrates click input with language-vision guidance, serving as a baseline for AerTrack-460; (3) We leverage spatio-temporal click data through the Point Trajectory Memory Module in CLaVi for path localization. Finally, we provide the AerTrack-460 benchmark through experiments, providing a foundation for future research.

2 AerTrack-460 Dataset

To extend vision-language trackers to the aerial domain, we develop AerTrack-460, which is a collection of official videos and language annotations from five diverse aerial datasets, including

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

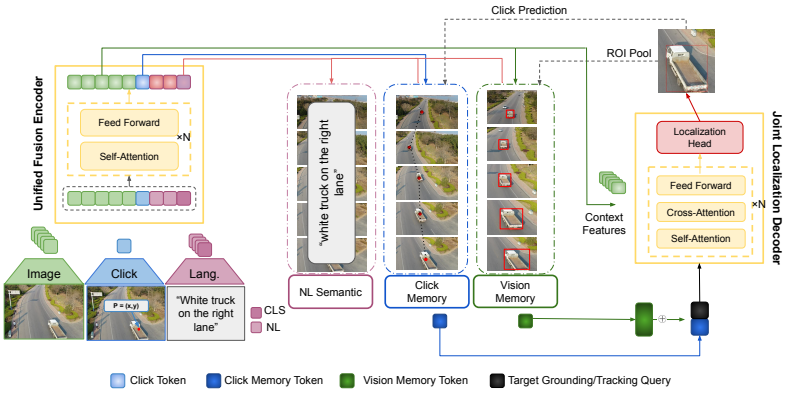

Figure 1: **CLaVi Overview:** (a) Inputs (image, language prompt, and click) are encoded and fused; (b) Memory modules process each modality; (c) A transformer decoder predicts bounding boxes using fused information.

UAVDT [6], UAV123 [21], UAVDark135 [17],DTB70 [18], and UAVTrack112 [10]. This diverse dataset spans various conditions and environments, offering a broad range of object classes with captions, making it a valuable resource for visual grounding. We annotate this collection of 5 dataset with the target's color, action, and surroundings in the first frame of each video. In total, AerTrack-460 has 461 sentences, reviewed and refined by our team.

3 CLaVi Framework

Building on the JointNLT framework for grounding and tracking, we extend it to the aerial domain by introducing an additional click modality and utilizing point tracking for path localization. Fig. 1 shows the CLaVi framework and its three key components.

Multi-Modal Feature Encoding and Fusion A pre-trained Swin Transformer [20] encodes both the input image and template into flattened embeddings. A BERT model [3] encodes the tokenized language prompt with added CLS and SEP tokens, producing a text embedding. Clicks are encoded using Gaussian Random Fourier features [23] and a learnable embedding vector [16] for positional information. During training, the click is the ground-truth bounding box center with random jitter; during evaluation, it's user-provided for the first frame and predicted in subsequent frames. The encodings from the image, template, click, and language are concatenated and passed to a Feature Fusion Encoder, yielding enhanced representations.

Modal-Tailored Memory Modules To leverage spatio-temporal information, we design memory modules for each modality:

NL Semantic Memory Module. Natural language descriptions provide consistent global information about the target object, we store the learned NL token in the NL memory module, keeping it fixed throughout video frames.

Click Memory Module. An object's center point remains a consistent spatial reference in video sequences despite changes in shape, appearance, and visibility [12], providing valuable trajectory information for tracking. To leverage this, we design a click memory module with a transformer architecture that stacks the enhanced click outputs from k previous frames and combines them with enhanced language encoding. A transformer decoder with a learnable target query then cross-attends to the encoder output to yield a click temporal clue.

Vision Memory Module. Similar to [28], this module stores ROI-pooled features from previous bounding boxes, which, along with language embeddings, provide visual cues for tracking in future frames.

Joint Localization Decoding The embeddings from the fusion encoder and memory modules are passed to a cross-attention transformer-based decoder. The target query is formed by combining a learnable query embedding with click and visual temporal clues. A unified localization head predicts bounding boxes for both grounding and tracking tasks.

Table 1: **Quantitative Comparison** of CLaVi with traditional aerial BBox tracking, recent BBox, and NL tracking on the 5 aerial datasets.

Dataset	Method	Initialize	AUC	PRE	N-PRE	Dataset	Method	Initialize	AUC	PRE	N-PRE
UAVDT [6]	HiFT	BBox	-	53.58	**72.46**	**UAVTrack112** [10]	HiFT	BBox	-	59.34	72.36
	SiamRPN	BBox	-	48.85	64.25		SiamRPN	BBox	-	62.60	77.34
	TCTrack	BBox	-	51.38	67.02		TCTrack	BBox	-	62.55	75.15
	JointNLT	NL	51.22	41.53	52.41		JointNLT	NL	64.82	71.68	84.66
	CLaVi (Ours)	Click+NL	**69.43**	**61.05**	72.32		**CLaVi (Ours)**	Click+NL	**71.16**	**72.59**	**86.14**
DTB70 [18]	HiFT	BBox	-	65.57	84.65	**UAVDark135** [17]	HiFT	BBox	-	35.89	45.86
	SiamRPN	BBox	-	69.16	87.92		SiamRPN	BBox	-	49.80	61.65
	TCTrack	BBox	-	**70.51**	**90.46**		TCTrack	BBox	-	44.94	46.55
	JointNLT	NL	69.82	61.80	80.01		JointNLT	NL	62.09	52.65	62.03
	CLaVi (Ours)	Click+NL	71.50	63.27	81.98		**CLaVi (Ours)**	Click+NL	**66.01**	**56.97**	**67.99**
UAV123 [21]	HiFT	BBox	-	50.51	63.56						
	SiamRPN	BBox	-	56.08	73.56						
	TCTrack	BBox	-	51.89	63.67						
	JointNLT	NL	67.08	62.24	75.46						
	CLaVi (Ours)	Click+NL	**69.01**	**65.28**	**79.72**						

Figure 2: **Qualitative Results.** For one sequence, we compare bounding box predictions from JointNLT (NL) (yellow) and CLaVi (red). Ground truth bounding boxes are marked as green. Since click-only predictions are substantially worse, we omit them for clarity.

4 Experiments

Following JointNLT, we first train CLaVi on LaSOT [7], TNL2K [25], and OTB [26], then fine-tune and benchmark five models on AerTrack-460 sub-datasets. CLaVi accepts test images of size 320×320, click coordinates (x, y), and language prompts up to 40 tokens; the grounding template image is 120×120, and the memory modules' history length is set to $k = 8$. We train for 300 epochs on a single 80GB Nvidia A100 GPU using the Adam optimizer [15], optimizing for GIoU [22] and L1 loss. For inference, our method achieves approximately 33 FPS on a 16GB Nvidia 3090 GPU. We evaluate our method by comparing it to others based on three criteria: (a) BBox-initialized trackers, (b) language-initialized trackers, and (c) BBox and language-initialized trackers, and also demonstrate how CLaVi performs relative to single-modality counterparts.

Results and Analysis As shown in Table 1 and Figure 2, our method achieves superior tracking and grounding results across all aerial datasets compared to single-modality counterparts. We significantly outperform the language-initialized grounding and tracking method [28], demonstrating the effectiveness of incorporating the "click" modality to enhance both tasks. The statistical significance of our results confirms the robustness of our approach. We also present a comparison of CLaVi with current state-of-the-art aerial tracking methods. As one of the first attempts at click-guided language-initialized tracking, our method surpasses existing conventional aerial trackers on UAV123 [21], UAVTrack112 [10], UAVDark135 [17] and UAVDT [6].

5 Discussion and Conclusion

We addressed the gap in aerial single-object tracking by introducing AerTrack-460, a collection of five aerial datasets annotated with text descriptions to support multimodal research. We introduce the click modality—a user-provided point on the screen—as a distinct input alongside traditional visual and textual data. We present CLaVi, a framework that integrates click, language, and vision inputs to disambiguate predictions in scenarios with small or visually similar objects. Experiments show that CLaVi outperforms traditional tracking methods and achieves state-of-the-art grounding and tracking performance. Our AerTrack-460 benchmark serves as a robust baseline for future advancements in this domain.

References

[1] Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., Fu, C.: Tctrack: Temporal contexts for aerial tracking. 2022 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14778–14788 (2022) 1

[2] Dai, K., Wang, D., Lu, H., Sun, C., Li, J.: Visual tracking via adaptive spatially-regularized correlation filters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4670–4679 (2019) 1

[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 2

[4] Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: Tap-vid: A benchmark for tracking any point in a video. Advances in Neural Information Processing Systems **35**, 13610–13626 (2022) 1

[5] Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10061–10072 (2023) 1

[6] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. CoRR **abs/1804.00518** (2018), http://arxiv.org/abs/1804.00518 2, 3

[7] Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019) 3

[8] Feng, Q., Ablavsky, V., Bai, Q., Li, G., Sclaroff, S.: Real-time visual object tracking with natural language description. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 700–709 (2020) 1

[9] Fu, C., Cao, Z., Li, Y., Ye, J., Feng, C.: Onboard real-time aerial tracking with efficient siamese anchor proposal network. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021) 1

[10] Fu, C., Cao, Z., Li, Y., Ye, J., Feng, C.: Onboard real-time aerial tracking with efficient siamese anchor proposal network. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2021) 2, 3

[11] Fu, C., Li, B., Ding, F., Lin, F., Lu, G.: Correlation filters for unmanned aerial vehicle-based aerial tracking: A review and experimental evaluation. IEEE Geoscience and Remote Sensing Magazine **10**(1), 125–160 (2022). https://doi.org/10.1109/MGRS.2021.3072992 1

[12] Harley, A.W., Fang, Z., Fragkiadaki, K.: Particle video revisited: Tracking through occlusions using point trajectories. In: European Conference on Computer Vision. pp. 59–75. Springer (2022) 1, 2

[13] Held, D., Thrun, S., Savarese, S.: Learning to track at 100 fps with deep regression networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 749–765. Springer (2016) 1

[14] Kiani Galoogahi, H., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: Proceedings of the IEEE international conference on computer vision. pp. 1135–1143 (2017) 1

[15] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 3

[16] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 2

[17] Li, B., Fu, C., Ding, F., Ye, J., Lin, F.: All-day object tracking for unmanned aerial vehicle. CoRR **abs/2101.08446** (2021), https://arxiv.org/abs/2101.08446 2, 3

[18] Li, S., Yeung, D.Y.: Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (Feb 2017). https://doi.org/10.1609/aaai.v31i1.11205, https://ojs.aaai.org/index.php/AAAI/article/view/11205 2, 3

[19] Li, Z., Tao, R., Gavves, E., Snoek, C.G., Smeulders, A.W.: Tracking by natural language specification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6495–6503 (2017) 1

[20] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 2

[21] Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 445–461. Springer International Publishing, Cham (2016) 2, 3

[22] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union (June 2019) 3

[23] Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. NeurIPS (2020) 2

[24] Wang, B., Wu, V., Wu, B., Keutzer, K.: Latte: accelerating lidar point cloud annotation via sensor fusion, one-click annotation, and tracking. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 265–272. IEEE (2019) 1

[25] Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13763–13773 (2021) 3

[26] Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2411–2418 (2013) 3

[27] Yang, Z., Kumar, T., Chen, T., Su, J., Luo, J.: Grounding-tracking-integration. IEEE Transactions on Circuits and Systems for Video Technology **31**(9), 3433–3443 (2020) 1

[28] Zhou, L., Zhou, Z., Mao, K., He, Z.: Joint visual grounding and tracking with natural language specification (2023) 1, 2, 3