Language-guided Visual Consistency for Point Tracking

Abstract

Point tracking poses a formidable challenge in computer vision, aiming to estimate the corresponding positions of points across a long video sequence. Most recent advances focus on exploring temporal modeling of motion prediction while neglecting the consistency of backbone features. In this paper, we conduct an analysis to elucidate the factors contributing to the robust semantic correspondence of the diffusion feature. Our investigation reveals that text embeddings play a crucial role in enhancing the coherent representation of visual features. We integrate this insight into the point tracking task and propose an autogenic language-guided visual feature enhancement to reinforce point correspondence in long-term sequences. In contrast to other vision-language tasks, our text embedding is automatically generated from visual features through a specialized mapping network, thus it can be seamlessly adapted to any tracking task without requiring explicit text data. Furthermore, our designed consistency decoder efficiently incorporates text tokens into visual features with minimal computational overhead. Through enhancing visual consistency, our approach makes the trajectories of related points more discernible in lengthy videos with significant appearance variations. Extensive experiments on several widely used point tracking benchmarks demonstrate the superior performance of our approach, particularly the noteworthy enhancements compared to the baseline tracker, which lacks language-guided consistency enhancement.

1. Introduction

Point tracking represents a novel and formidable challenge in computer vision, holding considerable potential for unveiling insights into physical properties and 3D shape [7, 8]. The goal of point tracking is to estimate the pixel coordinates corresponding to specified points within any frame of an extensive video sequence. Although similar to the optical flow [35] which focuses on predicting pixel motion over short video frames, point tracking extends its purview to estimating point motion trajectories across extended temporal windows. Optical flow, in contrast, is constrained to predicting motion at infinitesimal distances and is incapable of capturing motion trajectories over prolonged periods.

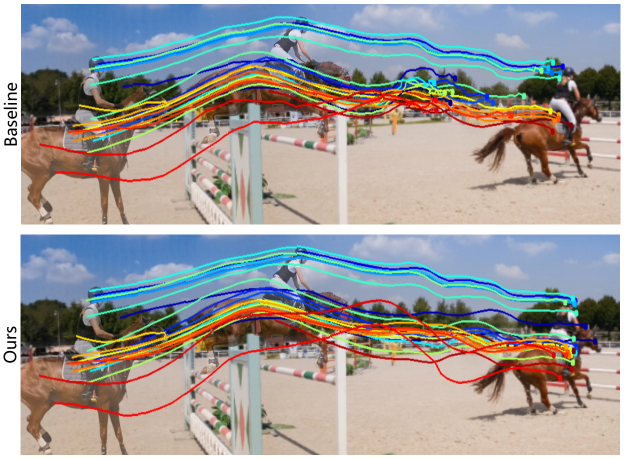


Figure 1. **Visualizing trajectories of tracked points**. We visualize the motion of the object at various times and compare the tracking trajectory between our approach and the baseline method. Our method maintains the same structural framework as the baseline, differing only in the utilization of language-guided consistency.

Recent techniques in point tracking mainly focus on enhancing temporal modeling for precise motion estimation, such as learning temporal priors for the target pixel's location [13], identifying the reliable long sequence of flows from the perspective of occlusion [23] and jointly tracking all points in a long range of frames [17]. In contrast to prior investigations, our approach prioritizes feature consistency by mapping features from diverse frames into a language-guided sharing space. This strategic emphasis ensures that, even in lengthy sequences with substantial appearance variations, as shown in Figure 1, our approach maintains robust semantic correspondence, thereby enhancing its effectiveness in long-term point tracking.

Diffusion features (DIFT) [34] exhibit notable capabilities for semantic correspondence, requiring no additional fine-tuning or supervision on task-specific data. DIFT excels in accurate keypoint matching between two distinct objects, even across different categories. Our analysis reveals that the text prompt plays a pivotal role in its semantic matching ability, under the process where different images are aligned with the same text prompt as the input of the diffusion model. Given that the text contains identical semantics, we hypothesize that embedding text tokens into visual features can diminish spatial differences across objects, consequently enhancing their overall consistency. Furthermore, we observe that the precision of textual descriptions also greatly impacts performance. For example, "*a sitting cat with gray and white color*" is more accurate than "*a cat*". Integrating diffusion features into point tracking is a compelling notion, yet it encounters two challenges: (1) most point tracking tasks lack text data, making it impractical and limiting to manually input precise descriptions for each sequence; (2) point tracking typically employs lightweight convolutional backbones to meet real-time and multi-frame computational demands, whereas diffusion's backbone is too complex for direct application in point tracking.

To address these challenges, we propose a point tracking framework augmented with autogenic **L**anguage-**G**uided visual feature enhancement (**LGTracker**). Our approach comprises three key components: (i) an automatic text prompt generation module that generates text tokens from image features through a vision-language mapping network; (ii) a text embedding enhancement module, ensuring precise text descriptions by incorporating image embeddings; and (iii) a text-image integration module designed to enrich the consistency of image features with textual information. In contrast to other vision-language tasks, our text information is automatically generated from image features, thus it can be adapted to any tracking task without requiring explicit text data. In addition, our visual consistency enhancement approach can be plugged into any point tracking method to effectively improve the tracking performance with slight computation overhead. Applying our feature enhancement to the baseline tracker enhances the Average Jaccard (AJ) score from 54.2 to 56.6 on TAP-Vid-DAVIS [7] dataset. Extensive comparison experiments on several challenging datasets including TAP-Vid [7] and PointOdyssey [42] exhibit the state-of-the-art performance, which further evidences the correctness of our analysis regarding the language impact on visual features.

In summary, our main contributions include:

1. We reveal that text prompts significantly enhance visual correspondence across images, and precise textual descriptions contribute more to semantic consistency. Subsequently, we applied this insight to track points within long video sequences.
2. We propose a language-guided visual feature enhancement method, consisting of three modules: mapping network to automatically generate text tokens, text description enhancement is designed to enrich the description of textual information, and image-text integration aimed at enhancing image features with text embeddings.

2. Related Work

Optical flow. Optical flow aims to attain pixel-level motion estimation of objects in image pairs. Traditionally, optical flow is conceptualized as an optimization problem and addressed through variational methods [2, 3, 15, 19]. Presently, convolutional network-based methods have demonstrated superior performance. FlowNet [9] employs a deep learning framework to learn end-to-end optical flow estimation models. DCFlow [40] constructs a 4D cost volume with convolutional features and refines the cost volume through Semi Global Matching. PWCNet [33] reduces computing costs by employing a feature pyramid to learn multi-scale features and incorporating wrapping techniques. RAFT [35] extracts pixel-level features, generates a 4D cost volume for each pixel, and iteratively updates the optical flow field by searching the cost volume. Recently, transformers [37] have made significant strides in optical flow research. FlowFormer [16] encodes the 4D cost volume into cost memory using an alternative group transformer layer and decodes the location cost queries through a recurrent decoder. GMFlow [41] formulates the flow estimation as a global matching problem, acquiring the matching relationship through a direct comparison of feature similarities. While optical flow methods allow for precise motion estimation between consecutive frames, they are not suited to long-range motion estimation.

Point Tracking. Several works develop the point tracker for predicting long-range pixel-level tracks in a feedforward manner. TAP-Vid [7] formulates the problem of Tracking Any Point (TAP) as continuously tracking target points through occlusion in a video sequence, which calculates a cost volume independently for each frame pair and utilizes it for coordinate regression and occlusion branches. Particle Video Revisited (PIPs) [13] revisits the classic Particle Video [29] problem, presenting a model which iteratively refines the features of multiple consecutive frames within a sliding window, enabling the prediction of the tracking point's trajectory and visibility. Recently, many concurrent works have emerged. MFT [23] identifies the most reliable sequence of flows by considering the occlusion and uncertainty map. Context-TAP [1] enhances PIPs by incorporating spatial context during the tracking of trajectories. TAPIR [8] combines two-stage approaches: a matching stage inspired by TAP-Net and a refinement stage inspired by PIPs. PointOdyssey [42] extends PIPs by removing its rigid 8-frame constraint, enabling it to consider a much broader temporal context. OmniMotion [38] represents a video using a quasi-3D canonical volume and achieves pixel-wise tracking through bijections between local and canonical space. CoTracker [17] collectively models the correlation of different points in time through spe-

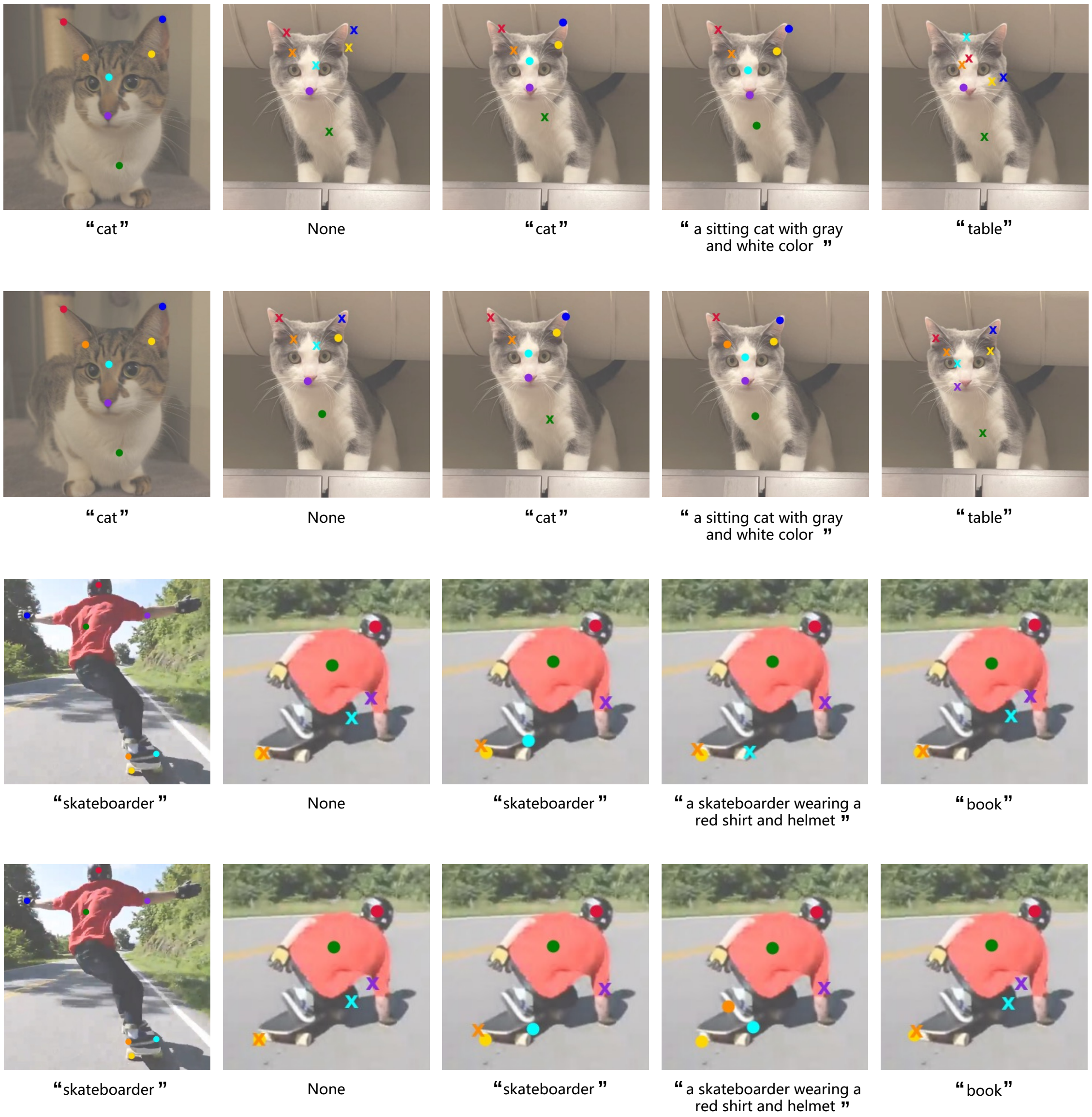

"cat"
None
"cat"
" a sitting cat with gray and white color "
"table"
"cat"
None
"cat"
" a sitting cat with gray and white color "
"table"
"skateboarder"
None
"skateboarder"
" a skateboarder wearing a red shirt and helmet "
"book"
"skateboarder"
None
"skateboarder"
" a skateboarder wearing a red shirt and helmet "
"book"

Figure 2. **Visualization of semantic correspondence with various text prompts**. The leftmost image is the source image with a set of key points; other images show correspondence under various text prompts, from left to right they represent: no text, simple description, detailed description, and semantically irrelevant text. Different colors represent different points. We use circles to denote correctly-predicted points under the threshold $\alpha_{bbox} \leq 0.1$ and crosses for incorrect matches. The results reveal that correct correspondences are established under text prompts (column 2 *vs* 3), especially the detailed description (column 4).

cialized attention layers and iteratively updates the trajectories. Our contribution is complementary to these works: the language information can be automaticlly generated and embedded into the visual feature to enhance the consistency across long-range video frames.

Vision-language models. Recently, the CLIP model [25] measures the similarity between images and text, it maps images and their corresponding text descriptions into a shared feature space that allows the model to perform various tasks, such as image segmentation [26], few-shot learning [36] and image caption [22]. Several methods explore utilizing language signals to the area of object tracking [20, 31, 32]. Some trackers [11, 39] use the language signal as an additional cue and combine it with the commonly used visual cue to compute the final tracking result. SNLT tracker [11] exploits visual and language descriptions individually to predict the target state and then

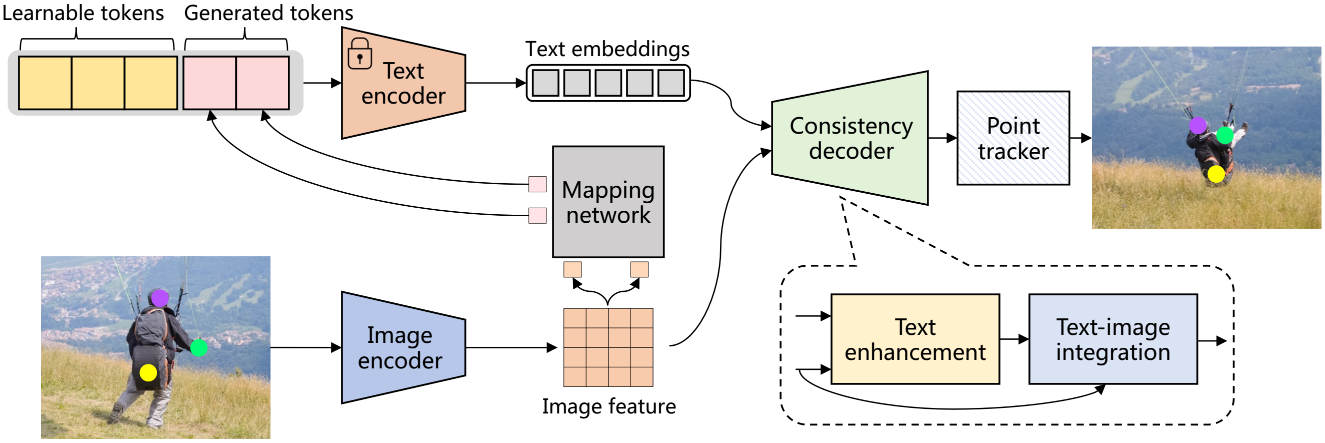


Figure 3. **The architecture of our LGTracker**. We introduce a mapping network that aligns image features with corresponding generated tokens to automaticly obtain the text information. A consistency decoder is designed to jointly process textual and visual information, the text enhancement module refines text embedding with enhanced descriptive capabilities, and an image-text integration module integrates the enhanced text embeddings seamlessly into image features. Finally, the tracking result is obtained through any point tracker.

dynamically aggregates these predictions for generating the final tracking result. Another type of method [12, 21] focuses on integrating the visual and textual signals to get an enhanced representation for visual tracking. The CapsuleTNL [21] develops a visual-textual routing module and a textual-visual routing module to promote the relationships within the feature embedding space of query-to-frame and frame-to-query for object tracking. In contrast to previous research, we leverage the semantic information of language to improve consistency in point tracking tasks over long sequences. Furthermore, our approach generates text descriptions from the image example, which eliminates the need for language annotations and expands the range of potential applications.

3. Method

In this section, we introduce our language-guided visual consistency in detail. Before proceeding, we first present an analysis on the text prompt of diffusion features.

3.1. Revisiting the correspondence in diffusion

Diffusion feature. Diffusion models[14, 30] are generative models that transform a Normal distribution to an arbitrary data distribution. In the forward process of the diffusion model, a noisy image x_t is obtained by combining a clean image x_0 with a randomly sampled noise $\epsilon \sim \mathcal{N}(0, I)$, expressed as:

$$x_t = \sqrt{\alpha_t} x_0 + (\sqrt{1-\alpha_t})\epsilon \quad (1)$$

where $t \in [0, T]$ means "step" in the diffusion process with larger time steps involving more noise. The amount of noise is determined by ${\alpha_t}_1^T$, which is a predefined noise schedule. The backward process involves obtaining a cleaner image x_{t-1} from the noisy image x_t by removing the noise ϵ, which can be iteratively estimated using a neural network f_θ. For image generation, f_θ is commonly parametrized as a U-Net[6, 28].

As outlined in DIFT [34], the diffusion feature of a given image is defined as feature maps of intermediate layers at a specific time step t during the backward process, it utilizes the Stable Diffusion [27] as the diffusion model. DIFT demonstrates that precise correspondences between two different images can be established through diffusion features using a straightforward approach: locating the maximum cosine similarity of feature maps between the target image and the search image.

Impact of text prompt. The diffusion feature exhibits superior capabilities in semantic correspondence compared to other vision models. Our analysis highlights the pivotal role played by text prompts in this process. Specifically, DIFT takes both an image and a text prompt as input, consistently using the same text prompt across different images. In pursuit of this understanding, we conduct experiments and identify key patterns: (1) Consistently using the same text prompt across images improves the correctness of correspondence, as illustrated in the 2nd and 3rd columns of Figure 2; (2) The accuracy of text description is a crucial factor influencing association ability, as depicted in the 4th column of Figure 2.

We hypothesize that the text encoder, pretrained by contrast learning [25], generates text embeddings containing semantic information similar to visual representations. This ability facilitates consistency across images when aligning these image features within the shared textual semantic space. Furthermore, finer textual descriptions yield more

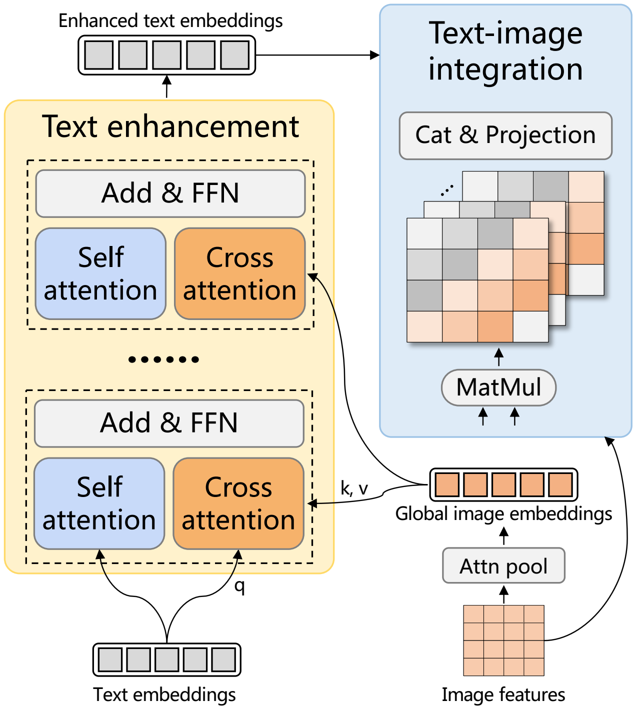


Figure 4. **The architecture of the consistency decoder**. The text enhancement module enriches text embeddings by integrating image embeddings into the attention mechanism, the text-image integration module combines the enhanced text embeddings with image features to obtain the consistency feature.

precise semantic information, a quality that distinctly impacts cross-image correspondence.

3.2. Language-guided visual consistency

According to the above findings, we propose leveraging the diffusion feature to improve the performance of point tracking algorithms. However, directly incorporating the diffusion model into the point tracking framework poses challenges due to the intricate nature of its feature extraction network. The computational load associated with extracting a single image of the diffusion model is already considerable, making it challenging to fulfill the simultaneous processing demands for long sequences in tracking. Additionally, the majority of datasets in the tracking field lack text information, it is difficult to obtain the detailed description of each sequence through manual input.

We propose the LGTracker, an autogenic language-guided visual feature enhancement for point tracking to integrate consistency of text prompts into a lightweight image encoder. As illustrated in Figure 3, we employ a text encoder and a convolutional network as image encoder adopted in the CLIP [25], we introduce a mapping network that aligns image features with corresponding text tokens. To jointly process textual and visual information, our approach incorporates a consistency decoder consisting of two key components: a text enhancement module and an image-text integration module. The former refines text embedding with precise descriptive capabilities and the latter integrates the enhanced text embeddings into image features.

Mapping network transforms image features to text tokens. Our text tokens are composed of learnable tokens and generated tokens. As the learnable tokens can be easily finetuned through the training process, we can employ a simple Multi-Layer Perception (MLP) as our mapping network. The input to the mapping network is the class token, derived from image features via a single-layer attention pooling. This class token encapsulates the global image information and facilitates a more effective alignment with text tokens.

Text description enhancement enhances the text accuracy by integrating global image embeddings. For example, "*a sitting cat with gray and white color*" is more accurate than "*a cat*" and performs more effectively in semantic correspondence. Inspired by the basic attention block in the Stable Diffusion [27], we introduce a text enhancement module that incorporates self-attention and cross-attention, as illustrated in Figure 4. Specifically, self-attention operates on the text embedding, while cross-attention operates jointly on the text embedding and the global image embedding, the text embedding serves as the query (q), and the image embedding functions as the key (k) and value (v). Throughout the multi-layer iteration, the text embedding undergoes continuous updates, while the global image embedding is consistently sourced from the original input. The global image embedding is derived by flatting the image feature through an attention pooling layer.

Image-text integration is proposed to model the interactions between vision and language. By employing matrix multiplication, we combine enhanced text embeddings $t \in \mathbb{R}^{K \times d}$ with image features $x_I \in \mathbb{R}^{H \times W \times d}$ to derive a integrated map $z = x_I t^T, z \in \mathbb{R}^{H \times W \times K}$. This integrated map can be regarded as a mapping result in a consistent space, offering consistent semantic information across a long sequence. The integrated map is concatenated with the image features to yield final features $x_f \in \mathbb{R}^{H \times W \times d}$, as defined by the equation:

$$x_f = Proj(Cat(x_I, x_I t^T)) \tag{2}$$

where the $Proj$ indicates the linear projection and the Cat means the concatenation along the last dimension. Our framework is scalable and can be applied to many video tasks to enhance feature consistency across long video sequences.

4. Experiments

In this section, we verify the individual contributions in the ablation study, and present the tracking evalua-

Method	Kinetics First			DAVIS First			DAVIS Strided		
	AJ	$<\delta^x_{avg}$	OA	AJ	$<\delta^x_{avg}$	OA	AJ	$<\delta^x_{avg}$	OA
TAP-Net [7]	38.5	54.4	80.6	33.0	48.6	78.8	38.4	53.1	82.3
PIPs [13]	31.7	53.7	72.9	42.2	64.8	77.7	52.4	70.0	83.6
MFT [23]	-	-	-	47.3	66.8	77.8	56.1	70.8	86.9
OmniMotion [38]	-	-	-	-	-	-	51.7	67.5	85.3
TAPIR [8]	**49.6**	64.2	**85.0**	56.2	70.0	**86.5**	61.3	73.6	**88.8**
Baseline	45.6	62.2	83.9	54.2	72.7	81.5	60.2	74.7	88.0
LGTracker(ours)	48.4	**64.2**	84.0	**56.6**	**74.3**	85.6	**62.9**	**77.9**	87.7

Table 1. **Evaluation on TAP-Vid-DAVIS and TAP-Vid-Kinetics datasets** [7]. The methods are evaluated under the "queried first" protocol and the "queried strided" protocol on DAVIS.

tion on several challenging benchmarks containing manually annotated trajectories in real videos, including PointOdyssey [42], TAP-Vid-DAVIS [7] and TAP-Vid-Kinetics [7].

4.1. Setting

Implementation Details. We employ a text encoder and an image encoder in the CLIP [25], the point motion estimation in the CoTracker [17] as our baseline. In the training phase, we train our approach on the PointOdyssey [42] training set for 80,000 iterations, and we randomly choose a 1/2/3-interval for consecutive frames. Points are preferentially sampled on objects and we randomly sample 256 trajectories for each batch, with points visible either in the first or in the middle frame. The size of an input image is resized to 384×512. The AdamW [18] optimizer is employed with an initial learning rate of $5e^{-4}$. We train our model on 4 Nvidia Tesla V100 GPUs. The mini-batch size is set to 4 with each GPU hosting 1 batch. Our approach is implemented in Python 3.8 with PyTorch 1.10.

Datasets. PointOdyssey [42] is a large-scale synthetic dataset for long-term point tracking of 80 videos on training set, 11 videos on validation set and 12 videos on test set, with 2035 average frames and 18,700 tracks per video. TAP-Vid-DAVIS [7] is a real-world dataset of 30 videos from the DAVIS 2017 validation set [24], which clips ranging from 34∼104 frames and an average of 21.7 point annotations per video. TAP-Vid-Kinetics [7] is a real-world dataset of 1,189 videos each with 250 frames from the Kinetics-700-2020 validation set [5] with an average of 26.3 point annotations per video.

Evaluation Metrics. We report both the position and occlusion accuracy of predicted tracks. Following the TAP-Vid and PointOdyssy benchmarks, our evaluation metrics include: Average Position Accuracy ($<\delta^x_{avg}$) measures the average position accuracy of visible points over 5 thereholds $\{1, 2, 4, 8, 16\}$; Average Jaccard (AJ) evaluates both occlusion and position accuracy on the same thresholds as $<\delta^x_{avg}$; Occlusion Accuracy (OA) evaluates the accuracy of the visibility/occlusion prediction at each frame; Median Trajectory Error (MTE) measures the distance between the estimated tracks and ground truth tracks; "Survival" rate means the average number of frames until tracking failure and is reported as a ratio of video length, failure is when L2 distance exceeds 50 pixels.

Method	MTE↓	$\delta\uparrow$	Survival↑
RAFT [35]	319.46	23.75	17.01
DINO [4]	118.38	10.07	32.61
TAP-Net [7]	63.51	28.37	18.27
PIPs [13]	63.98	27.34	42.33
PIPs++ [42]	26.95	33.64	50.47
Baseline	27.53	29.41	49.22
LGTracker(ours)	**24.44**	**33.91**	**51.37**

Table 2. Evaluation on PointOdyssey test set [42].

4.2. State-of-the-art Comparison

TAP-Vid. For the TAP-Vid benchmarks, we follow the standard protocol and downsample videos to 256 × 256 before passing them to the model, all the metrics are then computed in 256 × 256. We evaluate our models on the TAP-Vid-DAVIS and TAP-Vid-Kinetics, points are queried on objects at random frames and the goal is to predict positions and occlusion labels of queried points. In the TAP-Vid, "queried first" evaluation protocol, each point is queried only once in the video, at the first frame where it becomes visible. Hence, the model should predict positions only for future frames. In the "queried strided" protocol, points are queried every five frames and tracking should be done in both directions. We adopt the online method as our point tracker, it tracks points only forward, and we run the tracker forward and backward starting from each queried point. As "queried first" requires estimating the longest tracks, it is a more difficult setting than "strided". Moreover, "strided"

Learnable tokens	Mapping network	Text enhancement		Kinetics First			DAVIS First		
		self	cross	AJ	$<\delta^x_{avg}$	OA	AJ	$<\delta^x_{avg}$	OA
-	MLP	-	-	40.6	55.2	76.7	42.2	61.2	72.3
✓	MLP	6	-	45.6	61.2	81.1	53.9	72.7	81.1
✓	MLP	-	6	44.3	60.1	80.0	48.5	66.7	80.9
✓	MLP	4	4	46.6	60.7	84.9	55.8	74.1	85.6
✓	MLP	6	6	48.4	64.2	84.0	56.6	74.3	85.6
-	MLP	6	6	47.7	63.2	85.0	54.4	73.7	84.6
✓	Transformer	6	6	46.2	62.9	83.6	55.1	74.1	83.3
✓	MLP	10	10	48.4	64.2	84.0	56.6	74.3	85.9

Table 3. **Ablation experiments on the text generation and text enhancement module**. In the text generation module, we present the effectiveness of learning tokens and the type of mapping network. In the text enhancement module, we provide the evaluation of self-attention layers (self) and cross-attention layers (cross). Default settings are marked in gray.

demands estimating the same track from multiple starting locations and is thus much more computationally expensive. From the experiment results in Table 1, we can see that our method has achieved remarkable performance with an AJ score of 48.4 in Kinetics First and 56.6 in DAVIS First. Furthermore, our method has made significant improvements in all evaluation metrics compared to the baseline.

PointOdyssey. Inspecting results (as shown in Table 2) across rows, we can see that our LGTracker achieves the best results among all methods, achieving the highest MTE score of 24.4. Especially, compared to the baseline, our method has achieved a significant improvement by a specific gain of 3.09 of MTE. Our approach significantly outperforms the best existing tracker and demonstrates the effectiveness of the language-guided feature consistency.

Integration		DAVIS First		
Cat	Map	AJ	$<\delta^x_{avg}$	OA
✓	-	54.1	72.3	85.8
-	✓	50.9	69.7	80.9
✓	✓	56.6	74.3	85.6

Table 4. **Comparison of integration strategies**. *Cat* indicates the concatenation, *Map* means the integrated map obtained by the matrix multiplication.

4.3. Ablation Study

We ablate our approach to verify the effectiveness of our design decisions using the TAP-Vid-DAVIS and TAP-Vid-Kinetics datasets.

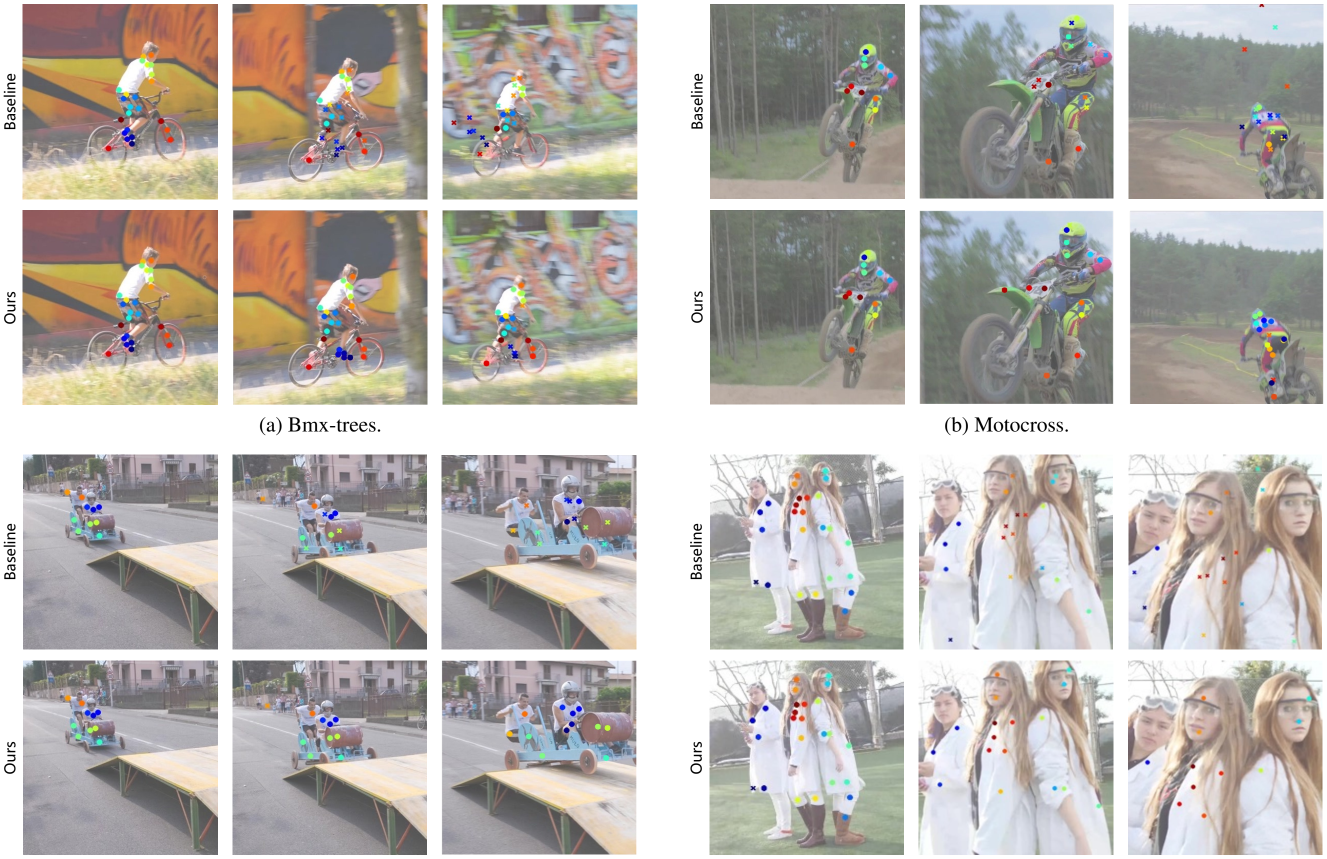
Text token generation. We compare the effectiveness of learnable tokens and different types of mapping network. Learning tokens are plugged into the generated text tokens and can be finetuned through the network training process. From Table 3 we can find that adopting the learning tokens can greatly improve the performance of point tracking (line 5 *vs* line 6 in Table 3). We also conduct the comparison between the MLP and the Transformer as our mapping network, each network adopts a three-layer basic unit. Evaluation results (line 5 *vs* line 7 in Table 3) demonstrate that MLP is a more suitable mapping network than Transformer for our purposes due to its superior performance and lower computational complexity.

Attention layers in text enhancement module. We tested several text enhancement schemes, including no text enhancement (line 1 in Table 3), self-attention enhancement only (line 2), cross-attention enhancement only (line 3), and simultaneous self-attention and cross-attention enhancement at different layers (line 4,5,8). Our experimental results demonstrate that both self-attention and cross-attention can enhance the representation ability of text embedding to varying degrees, and the impact remains constant after 6 layers. For optimal accuracy and efficiency, we have chosen 6 layers in our text enhancement module.

Text-image integration methods. Different integration strategies of text embeddings and image features have a large effect on the performance of consistency. A simple way to approach this is to concatenate these two cues (Cat) by flatting the image features and finally map them back to the original size. Our proposed integrated map (Map) through the matrix multiplication of text and image embeddings, which can be used as features alone (line 2 in Table 4) or concatenated with the original image features (line 3) for tracking purposes. The comparison results show that the concatenation of image features and integrated map can effectively improve the tracking performance.

4.4. Visualization

We offer visualizations of long-range trajectories of prototypical challenging scenarios in the tracking area to demonstrate the tracking performance of the proposed language-guided visual consistency. Figure 5 shows queried points of the shown frame between our method and baseline. As



(a) Bmx-trees.

(b) Motocross.

(c) Soapbox.

(d) Lab-coat.

Figure 5. **Visualization of our method and baseline on DAVIS** [24]. The images show tracking results over time. Different colors indicate different points. We use circles to indicate correctly-predicted points under the threshold $\alpha_{bbox} \leq 0.1$ and crosses for incorrect matches. Notably, our method yields accurate, coherent long-range motion even for fast moving (*Bmx-trees*), object deformation (*Motocross*), scale change (*Soapbox*), and similar distractor (*Lab-coat*) scenarios.

can be seen from the figure, after a long period of tracking, our LGTracker still achieves correct point tracking results. Crosses in images indicate incorrect matches. We observe that our approach has a strong discriminative ability for targets with severe scale variations and keeps a reliable associative ability in many challenging scenes.

5. Conclusion

In this study, we conduct an analysis to elucidate the factors contributing to the robust semantic correspondence of the diffusion feature. We reveal that text prompts significantly enhance visual correspondence across visual semantics, and precise textual descriptions contribute to improved semantic consistency. Incorporating this insight into the point tracking task, we propose a language-guided visual consistency to reinforce the point tracker. Our LGTracker consists of a mapping network aligning image features with corresponding text tokens and a consistency decoder to integrate textual and visual information. As our text information is automatically generated from visual features, it can be seamlessly adapted to any tracking task without requiring for explicit text input. Experiments conducted on several challenging point tracking datasets exhibited impressive performance, demonstrating that our approach renders point tracking more stable and discriminative, particularly in lengthy videos with substantial appearance variations.

Limitations. Our primary focus was on integrating autogenic language-guided consistency of convolutional networks for point tracking demands. However, we did not explore the consistent enhancement on alternative visual encoders, including vision transformers[10] commonly employed in cross-image tasks such as video object segmentation. In future, we intend to amalgamate our language-guided consistency to more image encoders, enhancing its adaptability to a broader range of scenarios.

References

[1] Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yitong Dong, Yijin Li, and Hongsheng Li. Context-tap: Tracking any point demands spatial context features. *arXiv preprint arXiv:2306.02000*, 2023. 2

[2] Michael J Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *1993 (4th) International Conference on Computer Vision*, pages 231–236. IEEE, 1993. 2

[3] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International journal of computer vision*, 61:211–231, 2005. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[7] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens Continente, Kucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. In *NeurIPS Datasets Track*, 2022. 1, 2, 6

[8] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. *ICCV*, 2023. 1, 2, 6

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 8

[11] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5851–5860, 2021. 3

[12] Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. *Advances in Neural Information Processing Systems*, 35:4446–4460, 2022. 4

[13] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *ECCV*, 2022. 1, 2, 6

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[15] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[16] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European Conference on Computer Vision*, pages 668–685. Springer, 2022. 2

[17] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv:2307.07635*, 2023. 1, 2, 6

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the ICLR*, 2018. 6

[19] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 2

[20] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. Diffusiontrack: Diffusion model for multi-object tracking. *arXiv preprint arXiv:2308.09905*, 2023. 3

[21] Ding Ma and Xiangqian Wu. Capsule-based object tracking with natural language specification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1948–1956, 2021. 4

[22] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3

[23] Michal Neoral, Jonáš Šerỳch, and Jiří Matas. Mft: Long-term tracking of every pixel. *arXiv preprint arXiv:2305.12998*, 2023. 1, 2, 6

[24] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6, 8

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4, 5, 6

[26] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 5

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[29] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *International journal of computer vision*, 80:72–91, 2008. 2

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 4

[31] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8791–8800, 2022. 3

[32] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Compact transformer tracker with correlative masked modeling. *arXiv preprint arXiv:2301.10938*, 2023. 3

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[34] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 1, 4

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the ECCV*, pages 402–419. Springer, 2020. 1, 2, 6

[36] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2

[38] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 2, 6

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021. 3

[40] Jia Xu, René Ranftl, and Vladlen Koltun. Accurate optical flow via direct cost volume processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1289–1297, 2017. 2

[41] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17592–17601, 2022. 2

[42] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 2, 6