SPARTAN: Spatiotemporal Pose-Aware Retrieval for Text-Guided Autonomous Navigation

Xiangyu Bai[1]
xiangyu.bai@northeastern.edu
Sai Anish Sreeramagiri[1]
sreeramagiri.s@northeastern.edu
Sai Siddhartha Vivek Dhir Rangoju[1]
rangoju.s@northeastern.edu
Bishoy Galoaa[1]
galoaa.b@northeastern.edu
Eric C Mortin[2]
eric.c.mortin.civ@army.mil
Sarah Ostadabbas*[1]
ostadabbas@ece.neu.edu

[1] Augmented Cognition Lab (ACLab), Department of Electrical and Computer Engineering, Northeastern University Boston, MA, USA
[2] DEVCOM Analysis Center (DAC) US Army

Abstract

Generative AI—particularly video diffusion models—offer a scalable alternative to traditional data collection for autonomous navigation. However, current models are not optimized for navigation-specific tasks, often resulting in inaccuracies in physics modeling and scene dynamics. We present SPARTAN (Spatiotemporal Pose-Aware Retrieval for Text-guided Autonomous Navigation), an open-source framework that delivers significant advancements to video diffusion components. At its core, SPARTAN features a novel spatiotemporal encoder that converts per-frame camera pose data into continuous spatiotemporal feature embeddings, enhancing representation and modeling efficiency. We also propose a camera pose-conditioned training pipeline and loss function that tightly integrate spatiotemporal features with text annotations to support more accurate retrieval and generation. In addition, we present DrivingScenePTX, a comprehensive driving video dataset that includes both frame-wise camera poses and rich textual scene descriptions. We benchmark SPARTAN against state-of-the-art contrastive language-image pretraining (CLIP) models using standard retrieval tasks, and introduce a novel evaluation method inspired by visual simultaneous localization and mapping (vSLAM) to assess performance in cross-domain trajectory retrieval. Our results show that SPARTAN retrieves driving videos with 15% higher spatial and temporal accuracy compared to the strongest fine-tuned models on SLAM-based metrics, offering a critical step forward in adapting generative AI for autonomous navigation. We release the code and dataset at https://github.com/ostadabbas/SPARTAN.

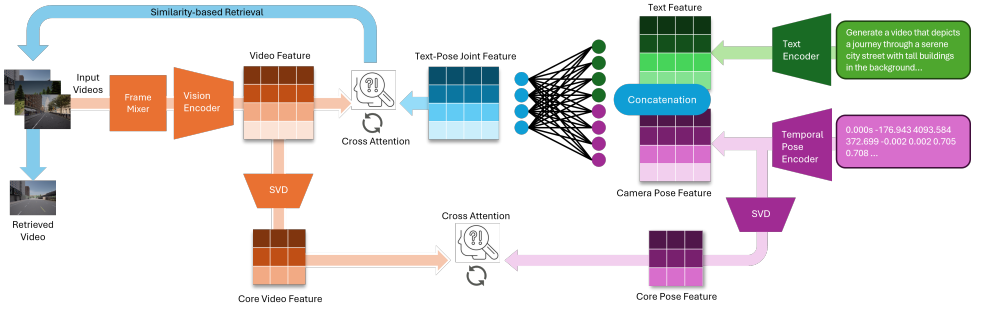

Figure 1: An overview of our SPARTAN pipeline. We start by encoding input videos, text and camera pose to obtain corresponding feature embeddings and then contrastively align them to obtain similarity scores. We leverage the score to identify and retrieve the output video that most accurately aligns with the given text and pose information.

1 Introduction

The success of autonomous vehicles relies heavily on robust navigation and control models that are trained and evaluated using accurately annotated data [20]. Recent advancements in multi-modal generative foundation models, particularly video diffusion models, have significantly accelerated research in autonomous navigation by acting as scene simulators capable of generating valuable, hard-to-obtain data—especially for edge cases and rare driving scenarios [4]. When guided by text prompts, state-of-the-art generative models can synthesize high-resolution driving videos with impressive visual realism [5]. However, the lack of accurate ground truth annotations (such as depth, segmentation, and ego-motion) along with physically incorrect or inconsistent scene dynamics, limits the reliability of generated videos for autonomy-related tasks. This ultimately reduces their effectiveness in simulation-based training and evaluation for real-world deployment [37, 40].

To address the limitations of current generative models, particularly their failure to produce videos that are both text-aligned and physically consistent, it is essential to develop architectures that not only adhere to text prompts but also respect the spatial and temporal dynamics of real-world scenes. While some approaches seek to redesign the full diffusion pipeline, this paper focuses on the contrastive language-image pretraining (CLIP) model, a critical component within video diffusion models. CLIP plays a key role in aligning prompts with visual outputs, ensuring coherence between prompts and generated images [17]. However, CLIP's reliance on text-only supervision makes it insufficient for autonomous navigation tasks, where understanding the egocentric spatial and temporal context is crucial.

To bridge this gap, we develop a powerful contrastive learning framework by incorporating ego vehicle pose information with textual annotations and RGB video frames. We introduce Spatiotemporal Pose-Aware Retrieval for Text-guided Autonomous Navigation (SPARTAN), a joint text-camera pose-video retrieval model that improves alignment accuracy and introduces missing spatial cues into the retrieval process. By grounding the generation in both textual and physical contexts, SPARTAN enhances the spatiotemporal awareness of generative models, enabling downstream autonomous evaluation and decision-making tasks.

As illustrated in Figure 1, our proposed framework builds on the limitations of current

text-only CLIP models by introducing a retrieval pipeline specifically tailored for autonomy-focused video generation and evaluation. The architecture comprises three main components: First, we introduce a temporal camera pose transformer that encodes the egocentric motion dynamics of the vehicle. This module takes per-frame timestamps along with camera translation and rotation matrices and projects them into a unified video-level pose embedding, capturing both spatial transitions and temporal evolution across frames. Second, we design a co-projection layer that tightly aligns the pose-aware video embeddings with textual scene descriptions. This fusion enriches the joint representation with contextual scene semantics and motion cues, enabling more accurate and physically consistent retrieval and generation. Lastly, to effectively balance the influence of textual and spatial information during training, we introduce a conditioned loss function alongside a dedicated training pathway. By leveraging singular value decomposition (SVD), we extract core features from both video content and camera pose trajectories, allowing the model to balance between salient dynamics and text features to avoid overfitting. Together, these components enable a retrieval model that is both text-aligned and spatiotemporal-aware, addressing a core limitation of existing video diffusion systems in autonomous applications.

To thoroughly train SPARTAN and to further advance diffusion models for autonomy, we introduce a driving scene annotation approach that utilizes vision-language foundational models (VLMs) and large language models (LLMs) with a special focus on context temporal stability. With this method, we curated DrivingScenePTX, a comprehensive driving video dataset with frame-wise camera **P**oses, **T**e**X**tual scene descriptions. Using CARLA simulator [CARLA], we collected 184K RGB frames from 11 diverse environments such as urban and rural areas with varied weather and time settings. Samples of our dataset can be found in Figure 3.

We assess SPARTAN's performance using standard retrieval tasks, reporting top-k retrieval accuracy. A demonstration is included in the supplementary materials. We further analyze the significance of the camera pose embeddings by comparing the alignment of the generated frames with provided camera poses via vSLAM metrics, which has yet been explored before. Finally, we perform ablation experiments to evaluate incorporating additional components on the accuracy of the retrieved frames.

To this end, we introduce a novel framework that advances video diffusion modeling for autonomous driving by embedding spatial and temporal scene understanding directly into the retrieval process. Our key contributions are as follows:

- We propose SPARTAN, a pose-aware, spatiotemporally grounded video retrieval model that jointly encodes textual scene descriptions, RGB videos, and frame-wise camera pose trajectories. This joint representation improves retrieval accuracy and makes the model suitable for autonomy-related tasks that require precise egocentric context.

- We introduce a camera pose-conditioned loss function that guides the training process to preserve temporal coherence and spatial consistency across video frames. This loss dynamically balances the influence of pose and text information, enabling more semantically and physically grounded retrieval.

- We present DrivingScenePTX, a large-scale, synthetic driving video dataset built in the CARLA simulator, containing 184K RGB frames with aligned frame-wise camera poses and textual scene annotations. This dataset supports downstream tasks such as video diffusion, retrieval, and evaluation in autonomous driving scenarios, particularly for rare and edge cases.

We are committed to open-source research and will make SPARTAN and DrivingScenePTX fully accessible, benefiting researchers in advancing the generative autonomy simulation.

2 Related Works

2.1 Diffusion Models for Autonomy Simulations

Using generative models for autonomy simulations has become a popular research area due to the surge in AI technologies. Researchers have built video diffusion models like SORA [30], CogVideo [42], Veo-2 [35], Hunyuan [34] and Lavie [39] to create videos and manipulate their context through text prompts. This approach has been extended to the autonomous vehicle field and further applied to various downstream tasks, such as traffic control [26], autonomous policy design [24, 31] and scene modeling [1, 12]. The majority of the existing research focuses on improving the architecture of the U-Net, which has a significant effect on the generation results in terms of quality and consistency [23]. Despite noteworthy overall progress, the inconsistency of features in spatial and temporal dimensions hinders a series of critical autonomy tasks such as visual simultaneous localization and mapping (vSLAM), object tracking and continuous depth estimation which rely on the stability of tracked features [6, 21, 43]. The proprietary nature of cutting-edge video diffusion models poses further challenges for academic researchers, limiting their ability to innovate and apply these models to downstream tasks.

2.2 Image-Text Retrieval Models

Image-text retrieval models are designed to bridge the gap between visual and textual data, allowing users to search for relevant vision data based on text queries or vice versa. CLIP-based architectures play a central role in retrieval tasks and serve as text-image encoders in video diffusion models, ensuring that generated images accurately match the context of text prompts [5, 29, 38, 44]. To extend these models from image domains to video, researchers integrate domain-specific components into the structure, allowing researchers to retain pre-trained weights and fine-tune on video datasets [19]. Despite the effectiveness of this approach, ensuring spatial and temporal stability in encoded text prompts remains an open challenge. While extensive research focuses on leveraging CLIP for semantic encoding, relatively few studies explore improvements through additional components and input data. Spatial-Temporal Auxiliary Network (STAN) by Liu et al. aims to improve temporal modeling by extending CLIP models to diverse video tasks [25]. Another approach, ST-CLIP, enhances temporal relationship modeling between video frames for dense video captioning [7]. These approaches enhance the usability of text-image encoders by integrating interactive mechanisms to generate high-quality, realistic generic videos, yet they remain insufficient for autonomy-focused applications. To address this gap, we introduce an optimized joint camera pose-text-video retrieval model designed for precise autonomous simulation. A critical aspect of ensuring retrieval accuracy in autonomy tasks is evaluating spatial and temporal stability. Conventional metrics, such as top-k retrieval rate and cosine similarity, primarily assess frame accuracy and fail to capture these essential dynamics [9, 14]. To overcome these limitations, we propose a new evaluation approach inspired by autonomous odometry tasks, ensuring a more comprehensive assessment of retrieval performance in autonomy scenarios.

2.3 Datasets for Autonomy Simulation

Synthetic data has proven effective in addressing data scarcity and imbalance challenges [4, 16, 18], offering a scalable and cost-efficient alternative to collecting real-world data in the

autonomy domain. Open-source driving simulators such as CARLA [11], GTAV-TeFS [28], and AirSim [33] have gained wide adoption for this purpose. These synthetic datasets, along with real-world datasets like KITTI [15], nuScenes [13], ApolloScape [27], and Cityscape [8], have served as benchmarks for training foundational models in autonomous driving research. However, despite their diversity and availability, current datasets used to train generative AI models for autonomy simulation still lack essential components–particularly text annotations with logically ordered temporal descriptions of scene content. This limitation hinders the ability to model event progression and causality, which are critical for reliable simulation and evaluation in autonomous navigation tasks.

3 Methodology

To overcome the limitations of the existing text-based retrieval algorithms in autonomous navigation, we present a unified framework that integrates camera pose data to capture critical spatial and temporal context. Our approach introduces a spatiotemporal encoder that transforms raw pose inputs into meaningful embeddings, which are then jointly aligned with text and video features to support more accurate and context-aware retrieval (see Figure 1).

Problem Formulation: Given a first-person driving video, each frame is associated with a six degree-of-freedom (6-DoF) camera pose and a timestamp. Let the 6-DoF camera pose for the ith frame be represented as $\mathbf{p}_6^{(i)} = [x^{(i)}, y^{(i)}, z^{(i)}, \phi^{(i)}, \chi^{(i)}, \psi^{(i)}]^T$, where $x^{(i)}, y^{(i)}, z^{(i)}$ denote the 3D position, and $\phi^{(i)}, \chi^{(i)}, \psi^{(i)}$ correspond to the Euler angles (yaw, pitch, and roll, respectively) that define the camera's orientation. Each frame is also tagged with a timestamp $t^{(i)}$ relative to the initial frame. In addition, we are given a text prompt $r^{(i)}$ that describes the scene's composition and visual appearance. The objective is to retrieve the most relevant video clip f that is consistent with the spatiotemporal context $(\mathbf{p}_6, t)$ and the narrative provided by r. To achieve this, we propose a joint feature embedding approach: (1) a pose-text encoder E_{pr} generates a joint embedding I_{pr} from the camera pose and text prompt with a pre-trained vision transformer E_v extracts visual features I_v from the video; (2) the retrieval model aligns the joint embedding I_{pr} with the video feature embeddings I_v with the supervision of (3) a joint camera pose-text CLIP loss to accurately identify the video clip f that best matches the specified pose, timestamp, and scene description. This framework leverages both the physical displacement in three-dimensional space and the orientation dynamics, as defined by a specific Euler angle convention, to enable robust spatiotemporal retrieval in dynamic driving scenarios.

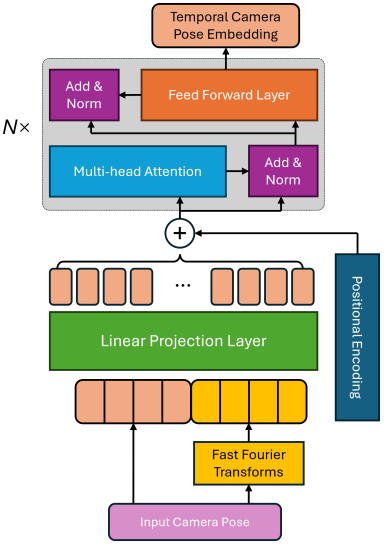

Figure 2: The implementation of spatiotemporal encoder in our model. We apply a fast Fourier transform to the input camera pose sequence and concatenate with the input sequence before applying linear projection to the data. The transformer block repeats by $N = 4$ times. Both text encoders in this illustration refer to the same encoder in the training process.

3.1 Spatiotemporal Encoder

Existing text encoders struggle to handle numerical data effectively, which is particularly relevant since camera poses consist entirely of numerical data. Additionally, each element in the $\mathbf{p}_6^{(i)}$ array has correspondence to the elements at the same position in the previous and next frames. This intrinsic relationship cannot be captured due to the tokenization steps and limited token lengths typically associated with CLIP models. To address this issue, we propose a novel sequence-aware spatiotemporal encoder to process per-frame camera pose data to continuous feature embeddings, the details of which can be found in Figure 2. To model temporal relationships, we employ the Fast Fourier transform (FFT) algorithm, converting time-dependent camera poses in the frequency domain:

$$P_{k,j} = \sum_{n=1}^{N} p6_{j,n} \cdot e^{-2\pi i k n/N}; j = 1, ..., b \tag{1}$$

N is the number of elements in each camera pose input array and b is the batch size. For our scenario, each frame contains seven input elements: a timestamp and a 6DoF camera pose, thus $N = 7$. This approach captures long-term dependencies between elements while effectively filtering out high-frequency noise caused by hardware jitter or external vibrations. We adopt a pad-and-crop approach to the transformed data and apply a normalization afterwards to avoid gradient overflowing (see Algorithm 1). A copy of the raw data is then concatenated with converted signals and passes through a linear projection layer, where positional encoding is applied to capture spatial relationships within the video frames. This data then goes through repeated attention blocks to obtain the embedding of the camera poses $\mathbf{E}_{\mathbf{p}}^{(\mathbf{i})}$.

Algorithm 1 Fast Fourier Transform

```
procedure APPLYFFT
    Poses ← a batch of raw camera pose
    F ← length of video in frames
    S_P ← FFT(Poses)
    I_P ← [Re(S_P), Imag(S_P)]
    if I_P.shape(1) < F then
        I_p ← Pad(I_p, Poses.shape, 0)      ▷ Pad I_p with zeros to match the shape of Poses
    else
        I_p ← Crop(I_p, Poses.shape)        ▷ Crop I_p to match the shape of Poses
    I_p ← (I_p − μ[I_p]) / σ[I_p]           ▷ Normalize I_p
    return [Poses, I_p]
```

3.2 Joint Pose-Text Embedding

We first tackle the issue of effective input token length. Our experiments indicate that the 77-token limitation of the pre-trained CLIP model reduces critical information, making it inadequate for our driving video retrieval tasks. To comprehensively describe the driving scenes, events, and objects in chronological order, we require lengthy text prompts. Taking design inspiration from Long-CLIP [45], we extend the effective input token length to 248.

For camera pose information, after we obtain the pose embeddings $\mathbf{E}_{\mathbf{p}}^{(\mathbf{i})}$ from the spatiotemporal encoder, we employ a fully connected layer to merge the embeddings with the text prompt embeddings $\mathbf{E}_{\mathbf{t}}^{(\mathbf{i})}$ to obtain joint text-camera pose embedding $\mathbf{E}_{\mathbf{j}}^{(\mathbf{i})}$. We initialize

the layer by constraining all the weights related to text input to sum up to 1.0 and randomly initializing all pose-related weights $\mathbf{w}_p << 1.0$. This process compresses the information during training and distills both types of input while retaining pre-trained CLIP weights, which were trained on a large corpus of data.

3.3 Joint Camera Pose-Text Retrieval Loss

To ensure that camera pose information is preserved during merging, we propose a distinct training loss with a novel paradigm that uses core video ($\mathbf{E}_{\mathbf{vc}}^{(\mathbf{i})}$) and pose features ($\mathbf{E}_{\mathbf{pc}}^{(\mathbf{i})}$). We apply the Singular Value Decomposition (SVD) algorithm to both the input videos and camera pose data to extract the essential features of the scene, including motion dynamics. These features are then contrastively aligned to train both text and spatiotemporal encoder, keeping pose information relevant. This training scheme allows us to balance the input information, ensuring that the model does not prioritize one over the other.

To better guide the training process, the loss function is designed as a weighted combination of Joint camera pose-text CLIP loss l_{joint} and core pose-video feature loss l_{core}:

$$loss = (1-\alpha)l_{\text{joint}} + \alpha l_{\text{core}} \tag{2}$$

l_{joint} is defined as:

$$l_{\text{joint}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{e^{(\text{sim}(\mathbf{E_j}^{(i)},\mathbf{E_v}^{(i)}/\tau)}}{\sum_{j=1}^{N}e^{(\text{sim}(\mathbf{E_j}^{(i)},\mathbf{E_v}^{(j)})/\tau)}} + \log\frac{e^{(\text{sim}(\mathbf{E_v}^{(i)},\mathbf{E_j}^{(i)})/\tau)}}{\sum_{j=1}^{N}e^{(\text{sim}(\mathbf{E_v}^{(i)},\mathbf{E_j}^{(j)})/\tau)}}\right) \tag{3}$$

$\mathbf{E_v}^{(i)}$ and $\mathbf{E_j}^{(i)}$ are the image and joint text-camera pose embeddings for the ith sample. $\text{sim}(\cdot,\cdot)$ is the cosine similarity between two embeddings. τ is a learnable temperature parameter that scales the logits. For l_{core}, we substitute $\mathbf{E_v}^{(i)}$ with $\mathbf{E_{vc}}^{(i)}$ and $\mathbf{E_j}^{(i)}$ with $\mathbf{E_{pc}}^{(i)}$. α is a parameter that determines the intensity of the influence of core pose and video features.

4 Experiments

In this section, we provide insights into evaluations and comparisons of SPARTAN against various state-of-the-art retrieval models on both widely recognized retrieval metrics and novel odometry-inspired tasks for autonomous evaluation. We also detail the properties and annotation procedures of DrivingScenePTX dataset.

4.1 Datasets and Settings

Our model is trained for approximately one week using four 32GB NVIDIA V100 GPUs, with a batch size of 12 and FP16 mixed precision. All of the training data is collected from eleven distinct scenes in CARLA simulator with varying time and weather conditions, featuring 184K RGB images at 720p resolution. We convert them into 2.4 seconds videos at 10 frames per second (fps) with a sliding window of random length to prevent fixed dependency between image and pose pairs. To extend image-based models to video domains, we adopt frame mixer, which is a function that reshapes video batch from $(batch, frames, -1)$ to $(batch \times frames, -1)$. This approach of adapting image-based transformers for video inputs has been proven effective in [32]. Samples can be found in the supplementary materials.

For text annotation, we adopt state-of-the-art VLM Florence2 [41] and LLM Grok-2 [10] to perform scene understanding and video summarization. By utilizing pre-trained VLMs, we use the grounding capabilities to generate detailed scene descriptions for each frame

Figure 3: Frames extracted from videos in the DrivingScenePTX dataset. Each image represents a distinct driving environment. The sample text annotation corresponds to the video from which the bottom-right images are taken.

of the dataset. We then enhance per-frame annotations into comprehensive per-video text generation instructions by summarizing the frame annotations with an emphasis on the temporal dynamics of the frame sequence in the instruction prompt to the LLM. Our approach specifically highlights the logical progression in which events are triggered by providing the (relative) timestamp of the annotation of each frame. For pose data, we extract the ground truth 6DoF camera pose along with its timestamp from the simulator. During training, we identify the corresponding poses within the video sequences and normalize them relative to the first frame. Frame and text annotation samples are provided in Figure 3.

4.2 Retrieval-based Evaluation

We evaluate SPARTAN's retrieval ability on our validation dataset. Following common practice in [22, 30, 36, 45], we examine on both text-pose-to-video and video-to-text-pose tasks and report top-k accuracy. As seen in Table 1, SPARTAN excels in both tasks, exceeding previous CLIP models by a large margin. The original, untuned image-based CLIP model performs the worst, followed by ViFi-CLIP, which is optimized for video classification with limited numbers of classes. This outcome is expected, as our dataset focuses on vehicle driving scenarios, characterized by similar patterns that cannot be effectively distinguished using a short sentence. For this very reason, tuning these models does not boost performance significantly as both original CLIP and ViFi-CLIP restrict input length to 77 tokens.

As the effective token length increases, Long-CLIP demonstrates superior out-of-the-box performance compared to tuned CLIP and ViFi-CLIP. Fine-tuning Long-CLIP further enhances accuracy, enabling the model to leverage the additional information provided by longer text. Fine-tuned Long-CLIP achieved a high top-5 accuracy compared to our model but exhibited a noticeable gap in top-1 accuracy. This is because certain videos are distinguishable only through subtle movements, such as scenes that start from the same position but feature slight variations in acceleration. The incorporation of camera pose and a temporal transformer effectively resolves this issue, thereby improving our model's top-1 accuracy.

4.3 SLAM-based Evaluation

To further demonstrate SPARTAN's camera-pose-aided retrieval ability and highlight the benefits of merging spatial-temporal information with text prompts, we evaluate our model using common odometry metrics [2, 3, 6]. Specifically, we approach the Video-to-Text-Pose retrieval task as a visual SLAM problem, and compute the Absolute Pose Error (APE) and Relative Pose Error (RPE) of the retrieved trajectory (camera pose) against the provided trajectory. To the best of our knowledge, this is the first instance of these metrics being used to evaluate CLIP-based models. We report our results in Table 2: our model achieved an

Table 1: Comparison of our model's performance on text-to-image and image-to-text tasks on DrivingScenePTX against CLIP (2021), ViFi-CLIP (2023) and Long-CLIP (2024). To ensure fairness, we adapted image-based models to the video domain using the same technique as our model. We then analyzed each model's pre-trained weights (where available) as well as weights fine-tuned on our dataset. For both tasks, we utilize top-1 and top-5 retrieval accuracy to assess the model's capability to identify the most relevant video corresponding to a given text prompt and camera pose within a batch of 100 videos. The evaluation spans 13 batches, and we report the mean and standard deviation across these batches.

CLIP Models	Text-Pose-to-Video		Video-to-Text-Pose	
	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑
CLIP []	9.33%±8.49%	32.98%±18.26%	11.46%±13.15%	30.73%±18.64%
CLIP (Fine-tuned)	16.87%±17.04%	41.33%±16.29%	17.80%±22.09%	47.07%±14.65%
ViFi-CLIP [] (Fine-tuned)	11.57%±2.40%	14.28%±3.50%	30.00%±3.86%	37.28%±2.87%
Long-CLIP []	16.87%±11.88%	42.93%±15.90%	16.87%±12.07%	40.87%±16.47%
Long-CLIP (Fine-tuned)	75.22%±7.50%	98.53%±0.96%	77.01%±9.17%	98.53%±1.25%
SPARTAN (Ours)	**81.53%±5.89%**	**99.06%±1.12%**	**80.40%±6.37%**	**99.40%±0.61%**

Table 2: Evaluation results of our model's performance using common SLAM metrics. We calculate the APE and RPE between retrieved trajectory (camera pose) and provided trajectory in each retrieval across 13 batches. We also express the percentage of errors relative to the average trajectory length to provide a more intuitive representation of accuracy.

CLIP Models	**Metrics**			
	APE(m)↓	APE(%)↓	RPE(m)↓	RPE(%)↓
CLIP	13.41±8.89	98.97%	0.98±1.60	7.23%
CLIP (Fine-tuned)	11.80±9.69	87.09%	0.61±1.27	4.50%
ViFi-CLIP (Fine-tuned)	12.92±9.33	95.35%	0.52±1.19	3.83%
Long-CLIP	12.30±8.36	90.78%	0.49±0.75	3.61%
Long-CLIP (Fine-tuned)	3.45±7.87	25.46%	0.26±0.96	1.91%
SPARTAN (Ours)	**1.34±3.82**	**9.89%**	**0.19±0.78**	**1.40%**

average APE of 1.34 meters, approximately 10% of the average trajectory length of 13.59 meters, outperforming other models by a significant margin. Furthermore, this evaluation approach uncovers insights that are overlooked in the retrieval task. Despite only a 6% difference in top-1 accuracy and a 0.5% difference in top-5 accuracy between Long-CLIP and our model, Long-CLIP's absolute pose error is nearly three times higher than that of our model, further showing the improved retrieval ability of our model.

4.4 Ablation Study

We perform an ablation study to assess the effectiveness of our proposed architecture (Table 3), demonstrating that incorporating temporal and spatial information enhances accuracy in driving video retrieval tasks. We observe significant improvement in terms of accuracy when extending input token length despite limited training, demonstrating that for driving datasets, a model's ability to process extended textual information is crucial for correctly identifying visual data. Adding spatiotemporal encoder also contributed to performance in a non-trivial way, allowing the model to better differentiate videos with highly similar text annotations by leveraging camera pose information. Our results also reveal that the model benefited from both the FFT module and the guided loss pipeline, improving its ability to extract and utilize spatial and temporal information from camera pose data. We also highlight that our model achieved a greater performance gain in the odometry (vSLAM) evaluation compared to traditional retrieval tasks (e.g., 1.8% improvement in APE versus 0.66% in top-1 accuracy), making it a more suitable candidate for potential autonomy applications.

Table 3: An ablation study of SPARTAN's components in various configurations. We trained each configuration with our dataset for one epoch for evaluation.

CLIP Models	Text-Pose-to-Video	Video-to-Text-Pose	Odometry	
	Acc@1↑	Acc@1↑	APE(m)↓	RPE(m)↓
Base CLIP Model (B)	3.13% ± 4.54%	11.60%± 23.71%	15.65±9.91	1.10±1.09
B + Extended Text Length (E)	56.13%±12.04%	53.00%±13.03%	5.34±8.88	0.59±1.15
BE + Spatiotemporal Encoder (T)	57.33%±12.33%	53.66%±13.42%	3.61±6.43	0.45±1.19
BET + FFT (F)	58.47%±12.16%	55.67%±12.29%	3.36±6.34	0.40±1.07
BETF + Guided Loss (SPARTAN)	**61.67% ±10.89%**	**60.60%±10.88%**	**3.26±6.27**	**0.43±1.15**

5 Conclusion

This paper presents SPARTAN, a novel open-source video-based CLIP architecture designed to support joint camera pose and text retrieval, specifically tailored for driving scenes. To facilitate effective training, we introduce our curated video driving dataset, DrivingScenePTX, which includes text annotations and camera pose data. Through innovative evaluation methods comparing several other CLIP-based approaches, we validate SPARTAN's superior retrieval performance. Furthermore, SPARTAN has the potential to assist generative vision models in producing accurate and cohesive driving videos for autonomy simulations, aligning with our future vision.

6 Acknowledgement

Research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Alexander Amini, Tsun-Hsuan Wang, Igor Gilitschenski, Wilko Schwarting, Zhijian Liu, Song Han, Sertac Karaman, and Daniela Rus. Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

[2] Xiangyu Bai, Yedi Luo, Le Jiang, Aniket Gupta, Pushyami Kaveti, Hanumant Singh, and Sarah Ostadabbas. An Evaluation Platform to Scope Performance of Synthetic Environments in Autonomous Ground Vehicles Simulation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[3] Xiangyu Bai, Yedi Luo, Le Jiang, Aniket Gupta, Pushyami Kaveti, Hanumant Singh, and Sarah Ostadabbas. Bridging the domain gap between synthetic and real-world data for autonomous driving. *ACM Journal on Autonomous Transportation Systems*, 2023.

[4] Xiangyu Bai, Le Jiang, Yedi Luo, and Sarah Ostadabbas. Dual-conditioned temporal diffusion modeling for driving scene generation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Carlos Campos, Richard Elvira, Juan J. Gomez, Jose M. M. Montiel, and Juan D. Tardos. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[7] Huimin Chen, Pengfei Duan, Mingru Huang, Jingyi Guo, and Shengwu Xiong. St-clip: Spatio-temporal enhanced clip towards dense video captioning. In *International Conference on Intelligent Computing*, pages 396–407. Springer, 2024.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European conference on computer vision*, pages 88–105. Springer, 2022.

[10] Murillo Edson de Carvalho Souza and Li Weigang. Grok, gemini, chatgpt and deepseek: Comparison and applications in conversational artificial intelligence. *INTELIGENCIA ARTIFICIAL*, 2:1, 2025.

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[12] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

[13] W. Fong, R. Mohan, J. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. In *ICRA*, 2022.

[14] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 35:5207–5218, 2022.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria DaneshvarKakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based Infant Action Recognition in the Wild with Very Limited Data. In *CVPR Workshop on Learning with Limited Labelled Data (CVPRW)*, 6 2023.

[17] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

[18] Le Jiang, Shuangjun Liu, Xiangyu Bai, and Sarah Ostadabbas. Prior-aware synthetic data to the rescue: Animal pose estimation with very limited real data. In *The British Machine Vision Conference (BMVC)*, 11 2022.

[19] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022.

[20] Dhanoop Karunakaran, Julie Stephany Berrio Perez, and Stewart Worrall. Generating edge cases for testing autonomous vehicles using real-world data. *Sensors*, 24(1):108, 2023.

[21] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

[22] Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. Dpc: Dual-prompt collaboration for tuning vision-language models. *arXiv preprint arXiv:2503.13443*, 2025.

[23] Senmao Li, Taihang Hu, Fahad Shahbaz Khan, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of unet encoder in diffusion models. *CoRR*, 2023.

[24] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024.

[25] Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H Li. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6555–6564, 2023.

[26] Jinxiong Lu, Shoaib Azam, Gokhan Alcan, and Ville Kyrki. Data-driven diffusion models for enhancing safety in autonomous vehicle traffic simulations. *arXiv preprint arXiv:2410.04809*, 2024.

[27] Weixin Lu, Yao Zhou, Guowei Wan, Shenhua Hou, and Shiyu Song. L3-net: Towards learning based lidar localization for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6389–6398, 2019.

[28] Yedi Luo, Xiangyu Bai, Le Jiang, Aniket Gupta, Eric Mortin, Hanumant Singh, and Sarah Ostadabbas. Temporal-controlled frame swap for generating high-fidelity stereo driving data for autonomy analysis. *British Machine Vision Conference*, 2023.

[29] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[30] OpenAI. Video generation models as world simulators, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: 2024-08-28.

[31] Ali Rabiee, Sima Ghafoori, Xiangyu Bai, Sarah Ostadabbas, and Reza Abiri. Streams: An assistive multimodal ai framework for empowering biosignal based robotic controls. In *International Conference on Artificial Intelligence, Robotics, and Control (AIRC)*, 2025.

[32] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6545–6554, 2023.

[33] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, pages 621–635. Springer, 2018.

[34] Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, et al. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*, 2024.

[35] Mohamad-Hani Temsah, Rakan Nazer, Ibraheem Altamimi, Raniah Aldekhyyel, Amr Jamal, Mohammad Almansour, Fadi Aljamaan, Khalid Alhasan, Abdulkarim A Temsah, Ayman Al-Eyadhy, et al. Openai's sora and google's veo 2 in action: A narrative review of artificial intelligence-driven video generation models transforming healthcare. *Cureus*, 17(1), 2025.

[36] Gustavo A. Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Clipartt: Adaptation of clip to new domains at test time. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7092–7101, 2025. doi: 10.1109/WACV61041.2025.00689.

[37] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[38] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023.

[39] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024.

[40] Wayve. Introducing gaia-1: A generative ai model for autonomy, 2023. URL https://wayve.ai/thinking/introducing-gaia1/.

[41] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[43] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. Video object segmentation and tracking: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–47, 2020.

[44] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. *arXiv preprint arXiv:2403.14148*, 2024.

[45] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024.