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

Xiangyu Bai, Sai Anish Sreeramagiri, Sai Siddhartha Vivek Dhir Rangoju, Bishoy Galoaa, Eric C Mortin, Sarah Ostadabbas

Augmented Cognition Lab at Northeastern University

BMVC 2025

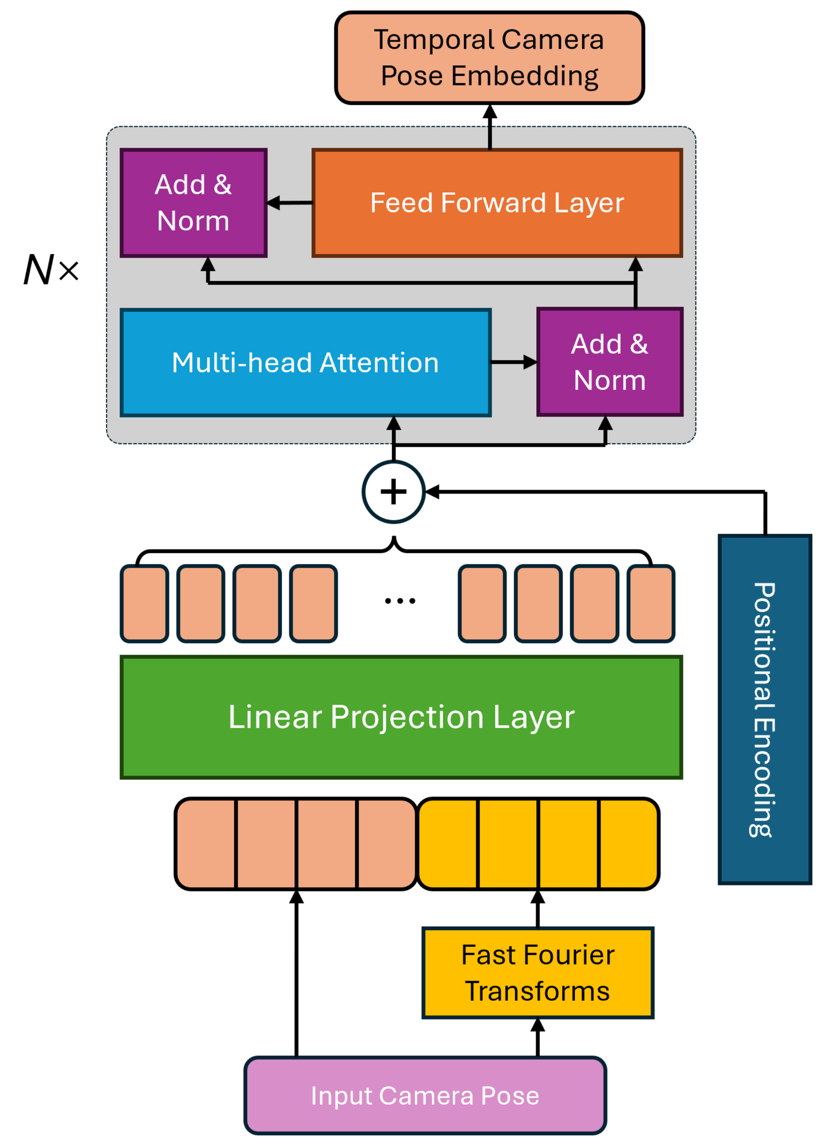
Background

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. 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. Our work 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. 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.

Our Solution: SPARTAN

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.

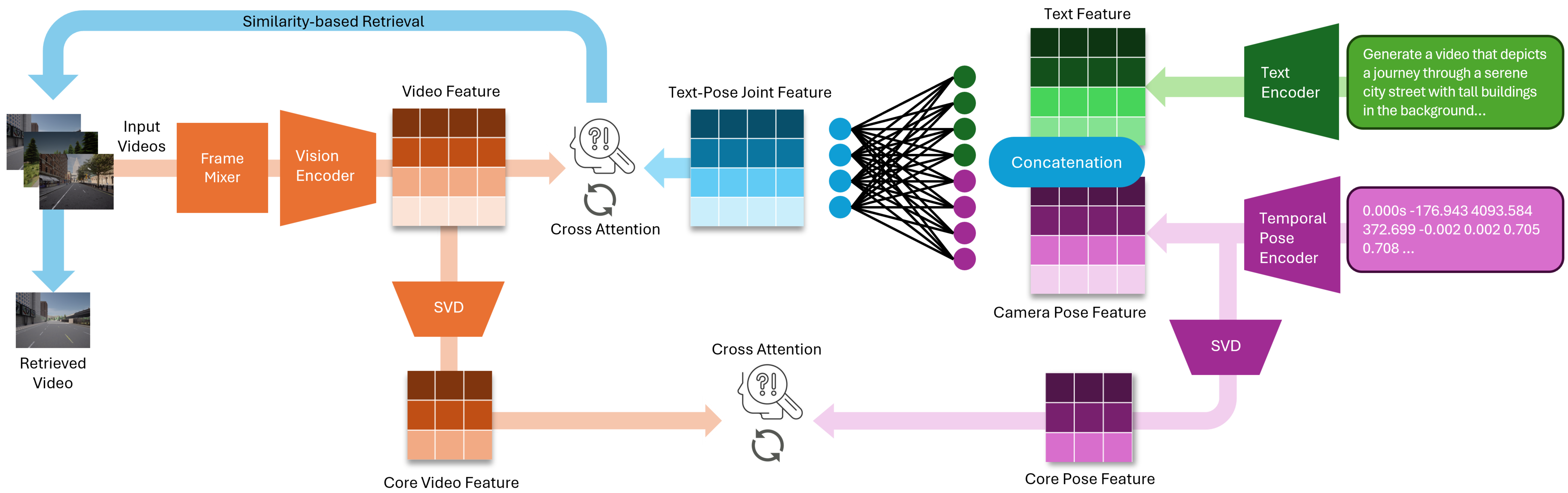
Temporal Camera Pose Encoder



Existing text encoders struggle to handle numerical data effectively, which is particularly relevant since camera poses consist entirely of spatiotemporally-related numerical data. This intrinsic relationship cannot be captured due to the tokenization steps and limited token lengths 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. To model temporal relationships, we employ the Fast Fourier transform (FFT) algorithm, converting time-dependent camera poses in the frequency domain. This approach captures long-term dependencies between elements while effectively filtering out high-frequency noise caused by hardware jitter or external vibrations. Additionally, 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.

Architecture

- 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.
- 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.
- We propose a conditioned loss function alongside a dedicated training pathway to effectively balance the influence of textual and spatial information during training. 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.



Dataset

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 Poses, TeXtual scene descriptions. Using CARLA simulator, we collected 184K RGB frames from 11 diverse environments such as urban and rural areas with varied weather and time settings.

Generate a video that depicts driving on a wet, empty road with trees on both sides and a hill on the right. The sky is grey and overcast, and the road appears wet from recent rain. Initially, there's a red stop sign and a street lamp on the left. As the car moves, the environment remains consistent with the stop sign still visible, but a small building appears on the left. Continuing along, the road remains wet and empty, with the same hill and trees, but now a "No Parking" sign replaces the stop sign. Raindrops start falling, blurring the windshield slightly. Approaching a speed limit sign, the road transitions to a more rural setting with houses and buildings in the background. The speed limit sign changes from a general warning to a specific "30 km/h" indication as the car continues on this gloomy, wet road.

Evaluations

Following common practice, we assess SPARTAN's performance using standard retrieval tasks, reporting top-k retrieval accuracy. SPARTAN excels in both tasks, exceeding previous CLIP models by a large margin on the evaluation set with vehicle driving scenarios characterized by similar patterns that cannot be effectively distinguished using a short sentence.

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. We show that SPARTAN achieves 3× lower Absolute Pose Error than Long-CLIP despite similar retrieval accuracy, highlighting its superior spatial grounding ability.

Finally, we perform ablation experiments to evaluate incorporating additional components on the accuracy of the retrieved frames, demonstrating that incorporating temporal and spatial information enhances accuracy in driving video retrieval tasks.

Acknowledgment 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.

Access Our I/O Page

CLIP Models	Text-Pose-to-Video		Video-to-Text-Pose		SLAM-based Evaluation			
	Acc@1↑	Acc@5↑	Acc@1↑	Acc@5↑	APE(m)↓	APE(%)↓	RPE(m)↓	RPE(%)↓
CLIP	9.33%±8.49%	32.98%±18.26%	11.46%±13.15%	30.73%±18.64%	13.41±8.89	98.97%	0.98±1.60	7.23%
CLIP (Fine-tuned)	16.87%±17.04%	41.33%±16.29%	17.80%±22.09%	47.07%±14.65%	11.80±9.69	87.09%	0.61±1.27	4.50%
ViFi-CLIP (Fine-tuned)	11.57%±2.40%	14.28%±3.50%	30.00%±3.86%	37.28%±2.87%	12.92±9.33	95.35%	0.52±1.19	3.83%
Long-CLIP	16.87%±11.88%	42.93%±15.90%	16.87%±12.07%	40.87%±16.47%	12.30±8.36	90.78%	0.49±0.75	3.61%
Long-CLIP (Fine-tuned)	75.22%±7.50%	98.53%±0.96%	77.01%±9.17%	98.53%±1.25%	3.45±7.87	25.46%	0.26±0.96	1.91%
SPARTAN (Ours)	**81.53%±5.89%**	**99.06%±1.12%**	**80.40%±6.37%**	**99.40%±0.61%**	**1.34±3.82**	**9.89%**	**0.19±0.78**	**1.40%**