

Supplementary Materials

A. Prompt Engineering

In this work, we present a prompt engineering strategy designed to refine initial video caption outputs into a more structured and concise representation of the scene. This approach plays a critical role by guiding the language model to extract only the essential elements of the scene, thereby standardizing the output for subsequent processing. We used LLaMA3 in our prompt engineering pipeline, and our evaluations confirmed that LLaMA3 delivered the best performance for our specific objectives.

A.1. Video Caption Generation

The foundation of our approach is illustrated by the following example prompt:

```
question = "List main subjects(  
    pedestrian, car, bicyclist, etc.)  
    and their actions, background scenes  
    (weather, time, road structure(  
    intersection or narrow lane or multi-  
    lane)), and other object or  
    specification in the scene. Answer  
    should be list of phrases containing  
    up to 8 items."
```

Raw captions generated from video content are post-processed using the above prompt. The objective of this step is to eliminate unnecessary details while retaining only key subjects, actions, and contextual information. This standardized scene representation is critical for subsequent processing stages.

A.2. Partial Element Generation

Partial object generation is a post-processing step designed to refine scene descriptions extracted from video captions. The aim is to filter out extraneous details, retaining only the essential elements required for further analysis. The process can be outlined as follows:

1. Initial Input Provision: The model receives an input list that contains descriptive phrases about a driving scene. For example: "Cars driving on street, Narrow lane with parked cars on both sides, Buildings in background, Intersection with crosswalk, Time of day appears to be mid-afternoon, Multiple lanes on street, Trees along sidewalk."
2. System Instructions: The model is then guided by a system prompt that instructs it to extract only the main subjects and objects from each phrase. The task is to respond with a set of keywords that concisely capture the

core elements. For instance, the expected output might be: "Cars, Narrow lane, Parked car, Building, Intersection, Mid-afternoon, Crosswalk, Multiple lanes, Tree, Sidewalk."

3. Application to Actual Inputs: By calibrating the model with this predefined example, the system ensures that when new lists are provided, the model consistently extracts the key components. This approach minimizes noise and standardizes the input for subsequent processing steps.

This calibration method not only enhances the accuracy of the extracted scene elements, but also contributes to a more robust overall framework for video retrieval.

A.3. Negative Object Set Generation

Negative object set generation is a post-processing step designed to enhance scene representation by removing elements that do not accurately reflect the primary context of the scene. This process leverages a corpus-based filtering approach to ensure that only objects irrelevant to the main scene are selected. The methodology is outlined in the following steps:

1. Corpus Provision and Input Setup: A predefined corpus containing a broad spectrum of object phrases is provided to the system. In parallel, an input list is supplied, detailing descriptive elements of a driving scene. For instance: "Cars driving on street, Pedestrians crossing street, Buildings in background, Intersection with crosswalk, Cars stopped at crosswalk, Multiple lanes on street, Trees along sidewalk."
2. System Instructions: The model is instructed to choose phrases from the corpus that do not overlap with any element present in the given list. The prompt explicitly directs the model to avoid selecting phrases that belong to the same category as any items in the input, ensuring that negative examples remain distinct. For example, if person-related terms are included in the input, the model is guided to exclude any phrases pertaining to humans.
3. Example Output Generation: An illustrative calibration example is provided to the model. In this example, the expected output might include phrases such as: "Orange cones on sidewalk, Bicyclist riding on the road, Narrow street with no cars, Yellow taxi, Sunny day, Traffic light." This example serves to establish a reference for the type of output required and the structure of the negative examples.
4. Application to Actual Inputs: Once calibrated, the model is applied to new input lists, consistently extracting a set of negative objects. This ensures that the generated negative object set is both coherent and distinct from the primary scene descriptors, thereby supporting a more robust and refined overall scene analysis.

By systematically filtering out irrelevant or misaligned el-

ements, the negative object set generation process contributes significantly to the accuracy and completeness of the final scene representation.

B. Effect of Attention Module

To demonstrate the effectiveness of our attention-based approach, we compared it with a multi-layer perceptron (MLP) baseline. The MLP consists of four fully connected layers followed by a sigmoid activation. It concatenates the caption and text embeddings and outputs a single similarity score, while keeping all other details the same. As shown in Tab. 1, the attention-based model demonstrates superior performance, confirming that it effectively catches object-level relations.

	F1-Score	<i>N-Acc</i>
4-layer MLP	8.0	43.9
Attention module	70.8	99.3

Table 1. Ablation on model architecture

C. Performance Stability with Changing Object Counts

The results in Tables 3 and 4 show the performance for configurations with 8 and 10 objects, respectively, in comparison to the baseline experiment using 6 objects. As shown in both tables, our approach maintains stable performance as the number of objects increases, which can be attributed to its object-level matching mechanism that averages scores over individual objects. This design effectively reduces the negative impact of adding more objects. In contrast, the SBERT-RB method, which uses a holistic similarity measure, exhibits a more pronounced decline in performance when additional objects are introduced. In Tab. 4, the ‘‘Similar’’ column for the 9-object configuration is marked with ‘‘–’’ because no result was available for that specific measure.

D. More Comparison

We additionally compare our method with TC-MGC [27], a recent state-of-the-art T2VR model. As shown in Table 4, while TC-MGC outperforms DRL, it still shows a significant performance gap compared to our model, CARIM.

This result highlights the importance of domain-specialized modeling. Unlike general T2VR benchmarks, where each video is associated with distinct and diverse descriptions, driving scene datasets such as DRAMA consist of clips that share highly similar visual and semantic patterns—e.g., repeating objects like vehicles, roads, and inter-

	R@1	R@5	R@10	<i>MdR</i> ↓	<i>MnR</i> ↓
DRL	3.3	10.3	15.8	78.2	114.4
TC-MGC	5.6	16.9	23.1	38.5	70.5
BEV-CLIP	41.0	62.5	75.4	13.0	17.6
CARIM(Ours)	74.3	86.8	90.4	2.1	5.4

Table 2. The additional performance comparison on DRAMA dataset.

sections. This makes it particularly challenging for generic retrieval models to discriminate between scenes effectively.

CARIM addresses this challenge by leveraging fine-grained object-level representations and contrastive training strategies tailored for the structured nature of driving environments, leading to significantly improved retrieval accuracy.

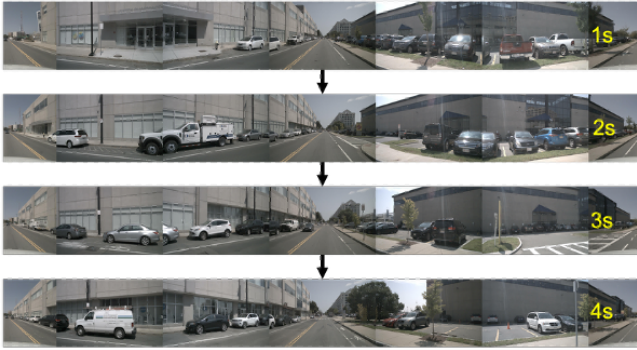
E. Extension to Panorama-based Video Captioning

To evaluate the generalizability of our method, we additionally conducted video captioning experiments on the nuScenes and Waymo Open datasets. Both datasets provide multi-view camera images suitable for panoramic stitching. Specifically, nuScenes dataset includes six surrounding views (front, front-left, front-right, rear, rear-left, and rear-right), from which we generate 360-degree panoramic frames. Similarly, the Waymo Open Dataset provides five directional views (front, front-left, front-right, side-left, and side-right), enabling the construction of wide-field panoramic video inputs.

As shown in Figure 1, our caption generation framework successfully produces informative descriptions of the driving scenes, even for these stitched panoramas. The generated captions capture not only static scene elements (e.g., urban structures, number of lanes) but also dynamic information such as agent actions (e.g., pedestrian crossing, vehicles turning left or moving straight). This rich semantic information enables fine-grained retrieval.

Since our CARIM model is designed to perform retrieval based on these generated captions, it naturally supports diverse video input types. Furthermore, inspired by recent advances such as WiseAD [31], which fine-tunes general vision-language models for autonomous driving tasks, we expect that task-specific fine-tuning of the captioner will further improve retrieval accuracy.

Original Video



Query to generate video caption

"List the main subjects (e.g., pedestrian, car, bicyclist, etc.) and their actions, background scenes (e.g., weather, time, road structure such as intersection, narrow lane, multi-lane, etc.), ego vehicle's action (e.g., turning left or going straight, etc.), and other relevant objects or notable specifications in the scene. Answer should be a list of phrases containing up to 8 items."



Video caption generated

1. Cars parked on side of road
2. Vehicles driving on road
3. Buildings in background
4. Overcast sky
5. Ego vehicle driving straight
6. Yellow lines on road
7. Trees along sidewalk
8. No pedestrians visible

(a) Panorama-based video captioning result on **nuScenes**.

Original Video



Query to generate video caption

"List the main subjects (e.g., pedestrian, car, bicyclist, etc.) and their actions, background scenes (e.g., weather, time, road structure such as intersection, narrow lane, multi-lane, etc.), ego vehicle's action (e.g., turning left or going straight, etc.), and other relevant objects or notable specifications in the scene. Answer should be a list of phrases containing up to 8 items."



Video caption generated

1. Car driving on road.
2. People standing on sidewalk.
3. Street lights illuminating area.
4. Road with multiple lanes.
5. Ego vehicle turning left.
6. Signs on sidewalk.
7. Overhead street lights.
8. Traffic cones on road.

(b) Panorama-based video captioning result on **Waymo**.

Figure 1. Panorama-based video captioning results on (a) nuScenes and (b) Waymo datasets. Each row corresponds to one video clip, stitched from multiple surround-view camera images. Captions are generated based on temporal dynamics and panoramic context.

# of Object	Method	F1-Score	Precision	Recall	<i>N</i> -Acc	Similar	R@1	R@5	R@10	<i>MdR</i> ↓	<i>MnR</i> ↓
1	SBERT-RB	0.4	5.3	0.2	60.8	0.2	7.2	17.1	24.1	48.0	89.1
	CARIM	63.7	72.6	71.6	100.0	74.6	14.4	34.6	46.8	13.0	31.4
2	SBERT-RB	3.3	9.2	2.9	49.4	1.4	16.6	32.8	42.7	18.0	49.2
	CARIM	49.1	54.8	64.0	100.0	71.3	50.3	80.3	88.0	1.0	4.8
3	SBERT-RB	15.6	23.1	16.0	37.1	5.7	33.9	58.0	68.5	3.0	24.2
	CARIM	56.2	57.3	74.9	100.0	74.0	78.8	95.6	98.7	1.0	2.0
4	SBERT-RB	42.0	41.6	57.8	21.1	36.6	57.5	74.0	80.5	1.0	13.7
	CARIM	68.0	68.2	83.0	100.0	63.5	93.2	99.6	99.8	1.0	1.2
5	SBERT-RB	48.1	41.9	84.7	12.7	58.8	81.0	89.9	91.9	1.0	7.0
	CARIM	81.7	80.9	90.7	100.0	64.7	98.0	100.0	100.0	1.0	1.0
6	SBERT-RB	37.9	31.0	95.9	6.2	66.7	91.0	96.5	96.9	1.0	3.4
	CARIM	88.0	87.1	93.6	100.0	66.7	99.3	100.0	100.0	1.0	1.0
7	SBERT-RB	33.3	26.5	98.9	1.5	100.0	95.4	97.6	98.2	1.0	2.1
	CARIM	90.5	90.2	93.2	100.0	100.0	99.3	100.0	100.0	1.0	1.0
Avg.	SBERT-RB	25.8	25.5	50.9	27.0	38.5	54.7	66.6	71.8	10.4	26.9
	CARIM	71.0	73.0	81.6	100.0	73.5	76.2	87.2	90.5	2.7	6.1

Table 3. Experimental Results for 8 objects

# of Object	Method	F1-Score	Precision	Recall	<i>N</i> -Acc	Similar	R@1	R@5	R@10	<i>MdR</i> ↓	<i>MnR</i> ↓
1	SBERT-RB	0.0	0.0	0.0	93.3	0.0	4.8	11.6	19.0	64.0	102.0
	CARIM	68.7	84.5	67.7	100.0	64.2	15.3	33.7	44.9	14.0	33.7
2	SBERT-RB	0.2	0.4	0.1	85.6	0.0	11.8	26.5	33.7	28.0	62.2
	CARIM	53.9	69.2	58.4	100.0	64.8	47.3	74.4	84.9	2.0	6.4
3	SBERT-RB	1.7	5.7	1.5	75.7	0.1	23.9	42.2	53.4	9.0	38.1
	CARIM	52.9	65.5	59.1	100.0	65.4	76.6	96.7	98.7	1.0	1.7
4	SBERT-RB	14.0	21.8	12.7	62.0	3.2	42.2	64.3	71.8	2.0	20.9
	CARIM	60.2	69.0	66.8	100.0	61.2	91.7	99.3	99.8	1.0	1.2
5	SBERT-RB	37.7	45.0	39.1	48.4	9.5	61.7	78.8	83.4	1.0	10.7
	CARIM	67.0	71.4	75.6	100.0	74.8	98.0	99.8	100.0	1.0	1.0
6	SBERT-RB	52.6	50.6	67.6	30.3	18.9	75.5	88.2	92.1	1.0	5.9
	CARIM	77.2	79.6	85.0	100.0	79.2	99.8	100.0	100.0	1.0	1.0
7	SBERT-RB	53.8	48.8	81.6	18.2	25.4	84.9	91.5	94.5	1.0	4.0
	CARIM	83.7	85.7	89.1	100.0	71.4	99.8	100.0	100.0	1.0	1.0
8	SBERT-RB	49.0	42.1	92.1	9.5	38.8	93.0	96.3	97.4	1.0	2.1
	CARIM	88.6	89.3	92.5	100.0	33.3	100.0	100.0	100.0	1.0	1.0
9	SBERT-RB	46.3	39.2	98.0	2.7	49.7	97.6	99.1	99.3	1.0	1.1
	CARIM	91.9	92.3	94.9	100.0	-	100.0	100.0	100.0	1.0	1.0
Avg.	SBERT-RB	28.4	28.2	43.6	47.3	16.2	55.0	66.5	71.6	12.0	27.4
	CARIM	71.6	78.5	76.6	100.0	64.3	80.9	89.3	92.0	2.6	5.3

Table 4. Experimental Results for 10 objects