Sim-to-Real Traffic Scene Understanding by Decoupling Semantics from Caption Generation with V-JEPA

Thuong Bui Nguyen Hoai[1*], Nguyen Vo Thanh[1], Giang Nguyen Trinh Tra[1], and Duc Bui Ha[1**]

Ho Chi Minh City University of Technology and Engineering (HCMUTE), Ho Chi Minh City, Vietnam
ducbh@hcmute.edu.vn

Abstract. Track 2 of the AI City Challenge 2026 requires both visual question answering (VQA) and traffic event description generation under a challenging synthetic-to-real domain shift. Existing vision-language approaches often entangle semantic understanding with language generation, making them susceptible to hallucination and inconsistent reasoning across event phases. In this work, we propose a decoupled semantic understanding framework that first resolves predefined traffic questions into structured semantic facts and subsequently uses these facts to guide caption generation. A frozen V-JEPA encoder extracts predictive scene representations, while a lightweight Llama-based predictor produces answers for VQA queries. To improve reliability, we introduce a training-free structured refinement mechanism that exploits statistical priors, inter-question relationships, and temporal event consistency to correct prediction errors. The refined semantic facts are then provided to Qwen3-VL-8B to generate pedestrian and vehicle descriptions for each traffic event. Experimental results on the official 2026 AI City Challenge Track 2 benchmark show that the proposed method achieves 87.09% VQA accuracy and an overall S2 score of 60.0853, ranking first among all participating teams. These results demonstrate that predictive world representations combined with structured semantic refinement enable more accurate and reliable traffic understanding, leading to higher-quality language generation.

Keywords: Traffic scene understanding · Sim-to-real transfer · Visual question answering · Caption generation · V-JEPA

1 Introduction

Autonomous driving requires a vehicle to continuously interpret complex traffic situations in which pedestrians, vehicles, and road infrastructure interact and evolve over time [20,21]. Reliable driving decisions therefore depend not only on

* The first three authors contributed equally.
** Corresponding author.

detecting the elements present in a scene, but also on understanding how their states, behaviors, and interactions jointly define the ongoing event. Such understanding provides the semantic context needed to recognize what is happening, assess potential hazards, and inform downstream prediction and planning.

Recent advances in vision-language models (VLMs) have significantly expanded the capabilities of traffic scene understanding by enabling open-ended question answering, semantic captioning, and language-based reasoning directly from visual observations [17,25]. Compared with conventional perception pipelines that predict predefined labels, VLMs provide richer descriptions of complex traffic situations and explain the interactions among road users in natural language. Reflecting this trend, Track 2 of the 2026 AI City Challenge evaluates whether models can understand simulated driving scenes through visual question answering and comprehensive scene captioning, emphasizing both semantic accuracy and reasoning ability [1,11].

Despite these advances, directly relying on VLMs to infer traffic semantics from raw video remains challenging. Driving scenes are inherently dynamic, containing multiple agents whose behaviors continuously evolve and influence one another. Understanding such scenes requires jointly reasoning about object identities, temporal evolution, spatial interactions, and causal relationships before forming a coherent semantic description. Asking a language model to perform this entire reasoning process directly from visual tokens can easily produce hallucinated events or logically inconsistent interpretations [16]. Moreover, Track 2 introduces an additional sim-to-real challenge [7,22] in which discrepancies in visual appearance, environmental conditions, and traffic characteristics reduce the robustness and generalizability of end-to-end language reasoning.

To address these challenges, we propose a two-stage approach in which traffic scene understanding begins with structured semantic analysis and is followed by language generation. Rather than prompting a language model to directly describe the entire traffic scene from visual observations, our approach first answers the structured visual questions provided in Track 2. Each question focuses on a specific semantic aspect of the driving scenario, such as participating agents, their attributes, behaviors, spatial relationships, or traffic events. The resulting pairs of questions and answers form a structured semantic representation that grounds the subsequent caption generation process. By reasoning over explicit semantic facts instead of raw visual tokens alone, the language model can produce more coherent and faithful descriptions while reducing hallucinations caused by complex traffic dynamics.

Specifically, we first leverage V-JEPA [18] to extract high-level scene representations that are less sensitive to appearance variations. A Llama-based predictor is then trained to infer structured semantic facts by answering the visual questions directly from these features. To eliminate contradictions, these predicted semantics are refined by a training-free mechanism that enforces temporal consistency and logical relationships among the questions. Finally, this refined semantic representation guides a Qwen model to generate the final description. This design allows language generation to serve as a means of ex-

pressing grounded scene understanding, rather than compensating for uncertain visual evidence with linguistic priors.

The main contributions of this paper are summarized as follows:

1. We propose a structured traffic scene understanding framework that grounds caption generation in explicit scene semantics, which are first extracted through reasoning guided by the benchmark questions.
2. We introduce a training-free semantic refinement mechanism that exploits relationships among questions and the temporal structure of events to improve semantic consistency and produce a coherent interpretation of the traffic event.
3. We achieve the top-ranked performance in Track 2 on the 2026 AI City Challenge leaderboard, demonstrating the effectiveness and domain transferability of our framework under challenging sim-to-real conditions.

2 Related Work

2.1 Traffic Scene Understanding and Captioning

Recent driving VLMs have moved beyond isolated recognition outputs toward structured language reasoning. DriveLM represents this shift through Graph VQA, which links perception, prediction, and planning questions so that driving decisions build on earlier scene evidence [20]. DriveVLM follows the same principle at the system level by organizing reasoning into scene description, scene analysis, and hierarchical planning [21]. This structured view is also important for WTS, where captions must describe fine-grained pedestrian–vehicle interactions across multiple views and event phases [11]. Accordingly, Kachhadiya et al. convert question–answer outputs into explicit facts before caption generation [10]. However, facts predicted independently can still contradict one another across related questions or event phases. Our method addresses this limitation by refining their relational and temporal consistency before using them, together with video frames, to guide caption generation.

2.2 Latent Video Representation Learning

Self-supervised video learning seeks to capture temporal structure without requiring task-specific annotations. While reconstruction objectives devote capacity to recovering visual details, V-JEPA predicts the latent representations of masked spatiotemporal regions from their visible context [4,13]. V-JEPA 2 scales this principle to large video collections, producing representations that support motion understanding and action anticipation [2]. VL-JEPA then extends latent prediction to vision-language learning by predicting continuous text embeddings, enabling retrieval and discriminative VQA without autoregressive decoding [5]. This progression motivates our use of frozen V-JEPA features and our formulation of answer prediction as retrieval in a semantic embedding space.

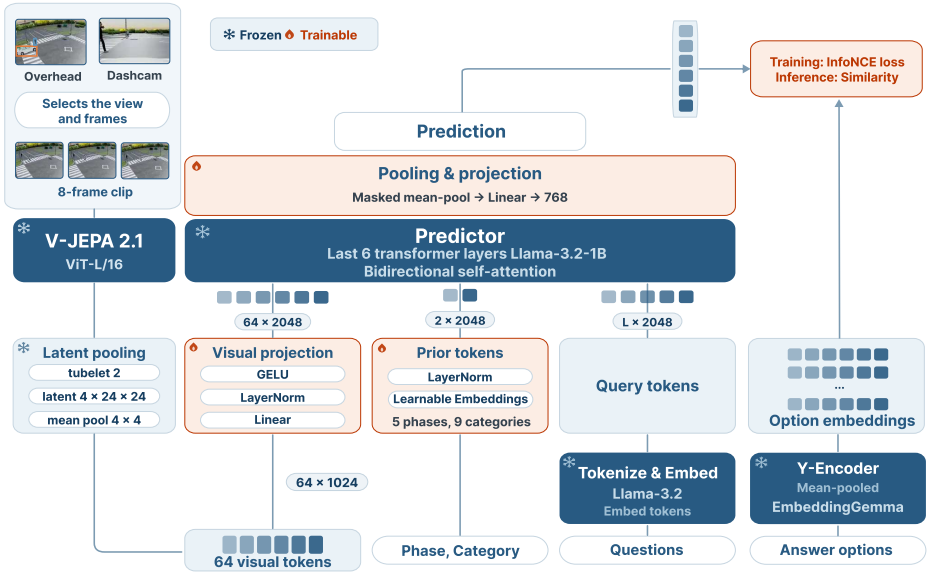

Fig. 1: Answer prediction. The selected camera video is encoded by a frozen V-JEPA 2.1. The video representation is combined with the question and mapped into the answer embedding space. Cosine scores provide the InfoNCE objective during training and the candidate scores for structured decoding during inference. Only the modules shown in orange are optimized on SynWTS.

2.3 Parameter-Efficient Adaptation and Structured Prediction

Adapting large pretrained models to limited task data commonly relies on learning a small set of parameters while keeping the backbone frozen. Prompt tuning [14, 15] learns continuous input tokens that condition a frozen language model, whereas LoRA introduces trainable low-rank updates to selected model weights [9]. Although these methods reduce training cost, they do not enforce agreement among predictions made independently. Structured prediction provides the complementary ability to model such dependencies through compatibility scores and sequence decoding [12,24]. Our framework combines both ideas: compact conditioning and projection modules adapt the frozen visual and language backbones, while a training-free decoder uses answer priors, question relations, and phase transitions to refine the predicted answers jointly.

3 Method

3.1 Overview

Track 2 requires both fine-grained visual question answering and the generation of pedestrian and vehicle descriptions for each traffic event. Rather than treating

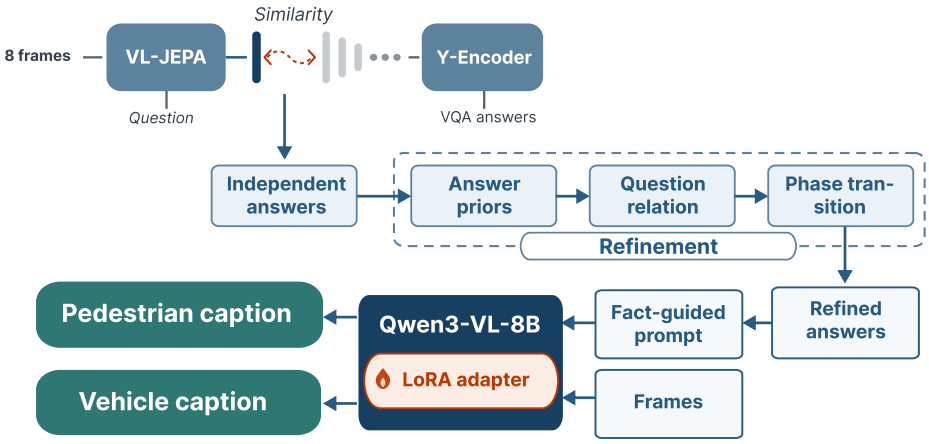

Fig. 2: Refinement and caption generation. The independent VQA answers pass through the training-free refinement (answer prior, question relation, and phase transition), and the refined answers become facts that, together with the frames, guide Qwen3-VL-8B to write the pedestrian and vehicle captions.

these as independent tasks, we exploit their complementary nature: the benchmark questions explicitly specify the semantic attributes that should appear in the final descriptions. Our framework therefore first resolves these queries into structured semantic facts and subsequently uses these facts to drive caption generation. This design is well suited to the benchmark because the questions explicitly enumerate the safety-critical attributes that should be reported, including actor awareness, gaze direction, relative position, and behavioral changes across event phases. In contrast, free-form caption generation leaves these attributes unspecified and often emphasizes visually salient but less informative content.

Figure 1 illustrates the overall pipeline, which consists of five stages. For each question, we first select the camera view that best captures the relevant actors and encode the corresponding event clip using a frozen V-JEPA 2.1 backbone. Inspired by VL-JEPA, the resulting latent representation is conditioned on the question, event phase, and category, allowing the answer to be retrieved from embedded candidate responses rather than generated autoregressively. Since questions are answered independently, the predictions may violate logical relationships or temporal consistency across event phases. To address this issue, we introduce a training-free semantic refinement module that combines answer priors, directed message passing, and temporal Viterbi decoding to enforce relational and temporal consistency. The refined answers are then normalized into structured facts about the environment, pedestrians, and vehicles, which serve as grounded semantic guidance for Qwen3-VL to generate the final captions (Fig. 2).

Given the limited size of the Track 2 dataset, directly fine-tuning large foundation models (V-JEPA 2.1, Llama-3.2-1B [8], and EmbeddingGemma) is im-

practical and may lead to overfitting rather than improved traffic understanding. We therefore freeze all pretrained backbones and train only lightweight projection, normalization, and conditioning modules to adapt the information flow between components. This parameter-efficient strategy reduces training costs while retaining the generalizable knowledge acquired during large-scale pretraining.

3.2 View Selection and Temporal Windowing

For each question, the corresponding event phase is captured from multiple camera viewpoints, but not all views provide equally informative evidence for semantic understanding. Depending on the viewpoint, the relevant pedestrian or vehicle may appear distant, partially occluded, or provide insufficient visual cues for identifying its state and behavior. To identify the most relevant perspective, we apply a straightforward, annotation-based selection rule. Questions associated with the driver's viewpoint are answered using the ego camera view. For other questions, we examine the overhead cameras during the queried event phase and select the camera containing the largest bounding box of the relevant pedestrian or vehicle.

After selecting the camera view, we further extract an actor-centric temporal window to remove unrelated frames before encoding. During training, the temporal window is defined by the first and last frames in which the queried actor is annotated with a bounding box within the selected event phase. If this interval is too short, we extend it by four frames at both ends to provide additional temporal context, while remaining within the phase boundaries. During inference, bounding boxes are available only for a small set of sparse key frames. We therefore define the temporal window as spanning from the earliest to the latest annotated frame of the queried actor, with an additional 0.5-second margin on either side. If the resulting window is shorter than one second, we extend it to one second. The window is also constrained to the queried event phase. In both settings, scene-level environment questions are the sole exception, since they are not associated with any specific event phase, and their temporal window therefore spans the entire clip.

Finally, the resulting sequence is then uniformly sampled to eight frames before being processed by V-JEPA, providing a fixed-length input while reducing the computational. Bounding box annotations are strictly used for view selection and temporal cropping and are not provided to the model.

3.3 Encoding with V-JEPA

Each selected frame is resized to 384×384 without cropping and normalized using ImageNet statistics. We employ a frozen V-JEPA 2.1 ViT-L/16 encoder [18] to extract spatiotemporal visual representations. For an eight-frame input clip, the encoder produces four temporal feature maps with a 24×24 spatial resolution, where each spatial location is represented by a 1,024-dimensional feature vector.

The resulting visual representation contains dense spatial tokens that are rich in information but excessive for the subsequent semantic reasoning module.

To efficiently adapt the visual features to the downstream task, we apply adaptive average pooling independently to each temporal feature map, reducing the spatial resolution from 24×24 to 4×4 while maintaining the temporal structure. The resulting four temporal grids are flattened into 64 visual tokens, which are then projected from 1,024 to 2,048 dimensions through a trainable linear layer to match the hidden dimension of Llama-3.2-1B. Layer normalization, GELU activation, and dropout are subsequently applied to improve feature adaptation and stability.

3.4 Question-guided Semantic Reasoning

The benchmark questions are designed to evaluate different aspects of traffic scene understanding through a structured reasoning process. Each question is associated with one of five predefined phases: Pre-recognition, Recognition, Judgment, Action, Avoidance, or related to Environment. We also group the questions into nine semantic categories according to their queried attributes, as listed in Table 1. The phase and category annotations are used as task-aware priors during question encoding. Lester et al. demonstrated that a similar use of learned conditioning tokens can effectively adapt frozen language models to downstream tasks [14, 15].

Table 1: Semantic question categories. The benchmark questions are assigned to nine categories according to the attributes they query.

No.	Category	Queried attributes
1	Environment	Pedestrian appearance and scene, road, traffic, and obstacle attributes
2	Pedestrian orientation	Pedestrian body orientation and direction of travel
3	Pedestrian position	Relative position between the pedestrian and vehicle
4	Pedestrian to vehicle distance	Relative distance between the pedestrian and vehicle
5	Pedestrian gaze	Line of sight, vehicle awareness, and visual status
6	Pedestrian behavior	Coarse and fine-grained pedestrian actions
7	Pedestrian speed	Pedestrian speed
8	Vehicle orientation	Vehicle field of view
9	Vehicle behavior	Vehicle action

For each question, linguistic tokens are extracted using the frozen Llama-3.2-1B embedding layer and truncated to a maximum length of 64 tokens. Meanwhile, the question phase and category indices into learnable token embeddings, to encode task-specific priors. The visual representation, phase tokens, category tokens and question tokens are subsequently concatenated into a unified sequence and fed into the predictor.

Following the bidirectional predictor design of VL-JEPA, the input sequence is processed by the final six frozen transformer layers of Llama-3.2-1B with the causal mask removed. Since the predictor performs latent representation inference rather than autoregressive generation, causal attention is unnecessary. Removing the mask enables bidirectional interactions among visual tokens, question tokens, phase tokens and category tokens. After the final normalization, all valid

output tokens are mean-pooled and mapped through a trainable 2,048-to-768 projection. The resulting vector $\hat{\mathbf{y}}_i \in \mathbb{R}^{768}$ represents the predicted semantic answer embedding conditioned on the video input and question i.

By embedding both predictions and candidate answers into a shared semantic space, answer prediction is formulated as an embedding retrieval problem rather than an autoregressive text generation task. For question i, let K_i be the number of valid candidate answers and let $o_{i,k}$ denote the text of its k-th candidate, where $1 \le k \le K_i$. Each candidate is independently encoded using the frozen 300-million-parameter EmbeddingGemma encoder $e(\cdot)$ [23]. The non-padding token representations are mean-pooled to produce $e(o_{i,k}) \in \mathbb{R}^{768}$, with each candidate text truncated to a maximum of 64 tokens. The predicted embedding is compared with every valid candidate embedding using cosine similarity, and the candidate with the highest score is selected:

$$s_{i,k} = \cos\big(\hat{\mathbf{y}}_i,\, e(o_{i,k})\big), \qquad k_i^\star = \underset{1 \le k \le K_i}{\arg\max}\, s_{i,k}, \qquad a_i = o_{i,k_i^\star}, \tag{1}$$

where $s_{i,k}$ is the cosine similarity score, $k_i^\star$ is the index of the highest-scoring candidate (the superscript $\star$ denotes the selected optimum), and a_i is the corresponding textual answer. The comparison depends on the content of each answer rather than its option label or position in the list. Changing the candidate order therefore does not change the semantic comparison. No textual answer is generated at this stage; instead, the similarity scores of all candidates are retained for the subsequent consistency refinement process.

The predictor is trained with an InfoNCE objective [19] to align the predicted representation with the semantic embedding space of candidate answers. Let k_i^+ denote the index of the ground-truth candidate for question i. The loss is defined as

$$\mathcal{L}_{\mathrm{VQA},i} = -\log \frac{\exp\big(s_{i,k_i^+}/\tau\big)}{\displaystyle\sum_{k=1}^{K_i} \exp\big(s_{i,k}/\tau\big)}, \tag{2}$$

where $\tau = 0.07$ is the temperature coefficient. The candidate indexed by k_i^+ is the positive target, while all other candidates for question i serve as negatives. This objective encourages the predicted representation to approach the semantic meaning of the correct answer while separating it from competing interpretations of the same question. By using alternative candidates for the same question as negative samples, the predictor is required to distinguish fine-grained attribute differences rather than only separating unrelated answer concepts.

3.5 Semantic Consistency Refinement

Although the predictor estimates each question independently, traffic semantics are inherently structured. Valid answers should not only be supported by visual evidence, but also be statistically plausible, semantically consistent with related questions, and temporally coherent across event phases. We therefore apply a training-free refinement during inference, with all statistics estimated from

SynWTS. It re-ranks only the candidates the model is unsure about, applying a statistical prior, relational message passing, and temporal Viterbi decoding in sequence.

Statistical prior. For a given question and event phase, some answers are far more frequent than others. We adjust each candidate's cosine score by its empirical log probability under SynWTS, smoothed with Laplace smoothing so that unseen answers retain a small nonzero mass. The refined score is

$$s_i^{(1)}(a) = s_i(a) + \gamma \log P(a \mid q_i) + \gamma_\varphi \log P(a \mid q_i, \varphi_i), \tag{3}$$

where $s_i(a)$ is the cosine similarity score of candidate answer a, q_i and φ_i denote the corresponding question and event phase, and $P(\cdot)$ is the empirical probability estimated from SynWTS. The weights $\gamma = 0.02$ and $\gamma_\varphi = 0.12$ are intentionally small, allowing the priors to support uncertain predictions without overriding the visual evidence.

Relational consistency. Independent predictions may violate semantic relationships between related questions, such as distance agreement, inverse relative positions, body orientation, and movement direction. For example, if the pedestrian is predicted to be *close* to the vehicle, the vehicle must be equally *close* to the pedestrian, since both questions measure the same physical gap. As another example, if the pedestrian is predicted to be *in front of the vehicle*, the vehicle must be the spatial inverse, *behind the pedestrian*. To address this issue, we perform one directed message-passing step over a predefined graph of question relations. For each relation $u \to v$, the compatibility between candidate answers is estimated from SynWTS with Lidstone smoothing [6] as the empirical conditional probability $C_{uv}(a \mid o_k)$, where o_k is an answer candidate of source question u and a is a candidate answer of target question v. When an observed source answer has no target occurrences, compatibility is assigned a small probability of 10^{-4}. Unseen source answers are assigned a uniform compatibility distribution.

The message from question u to question v is computed in the log domain as

$$m_{u \to v}(a) = \log \sum_{o_k \in \mathcal{A}_u} \exp\left(\frac{s_u^{(1)}(o_k)}{\tau_m} + \log C_{uv}(a \mid o_k) \right), \tag{4}$$

where $\mathcal{A}_u$ denotes the candidate set of question u, $s_u^{(1)}$ is the score after refinement using answer priors, and $\tau_m = 0.32$ controls the message sharpness. The target score is then updated as

$$s_i^{(2)}(a) = s_i^{(1)}(a) + \alpha w_{uv} m_{u \to v}(a), \tag{5}$$

where $\alpha = 0.8$ controls the overall contribution of relational information and w_{uv} assigns a separate weight to each relation type. A single shared weight would assign too much importance to loose relations and too little importance to near-deterministic ones, so each relation is weighted independently, tuned on the validation set to match how reliably it holds in the data. This update increases

the confidence of answers supported by related questions while reducing scores of semantically inconsistent predictions.

Temporal consistency. Questions describing the same attribute across multiple event phases should evolve smoothly over time. We therefore jointly decode repeated questions using Viterbi decoding [24]. Initialized as $\delta_1(a) = s_1^{(2)}(a)$, the optimal cumulative score for candidate a at phase t is

$$\delta_t(a) = s_t^{(2)}(a) + \max_b \Big[\delta_{t-1}(b) + \beta\, \ell_q\big(a \mid b, \varphi_{t-1}, \varphi_t\big) + \lambda\, \mathbf{1}[a = b] \Big], \qquad (6)$$

where ℓ_q is the transition log probability estimated from SynWTS, φ_{t-1} and φ_t denote the previous and current event phases, $\mathbf{1}[\cdot]$ is the indicator function, $\beta = 0.8$ controls the contribution of transition probabilities, and $\lambda = 0.2$ encourages answer persistence. After the loop completes, the temporal score of question i is its cumulative score at that phase,

$$s_i^{(3)}(a) = \delta_t(a). \qquad (7)$$

Temporal decoding is applied only to pedestrian behavior, body orientation, relative position, and gaze, where meaningful temporal evolution exists. Intuitively, ℓ_q favors answer changes that are common between consecutive phases in SynWTS and penalizes ones that almost never happen, so decoding prefers sequences that evolve the way attributes usually do. When SynWTS has no examples for a phase pair, the term stays neutral and lets the model score alone decide. Decoding is applied only when a question recurs in at least two phases of the same scenario, since a single-phase chain has no transition to decode. The optimal answer sequence is recovered by backtracking from the highest-scoring candidate at the final phase, and the selected candidate is promoted above the remaining options so that the model score remains the primary signal.

3.6 Guided Caption Generation

The refined answers are normalized into structured facts about the environment, pedestrians, and vehicles. For each event phase, we aggregate the facts inferred from all associated questions. These facts, together with sampled traffic frames, are given to Qwen3-VL-8B [3] to generate the caption. This design is supported by prior work on fact-augmented captioning. Kachhadiya et al. [10] likewise reformulate QA outputs as facts and use these fact-augmented inputs for structured traffic captioning. From the broader perspective of autonomous driving, DriveVLM [21] organizes VLM reasoning into scene description, scene analysis, and hierarchical planning, demonstrating that a guided scene description can capture task-relevant information that supports downstream driving decisions. Accordingly, our structured facts direct the model toward benchmark-relevant attributes, while the visual frames supply complementary contextual details. The pedestrian and vehicle captions are produced together from a single prompt, whose overall layout is shown in Fig. 3.

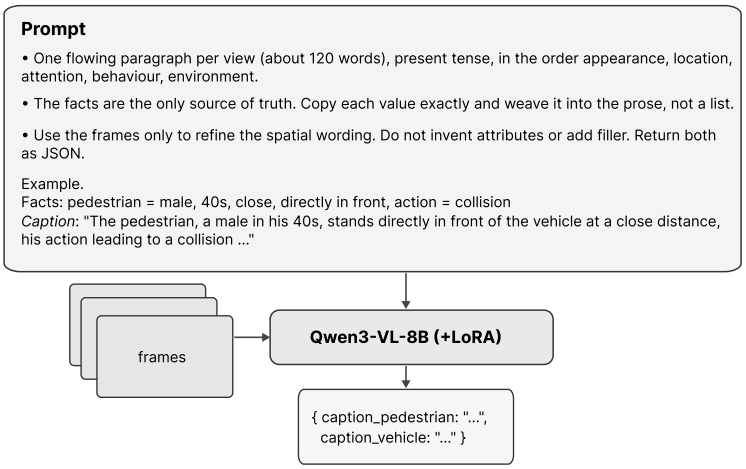

Fig. 3: Structure of the caption generation prompt. The refined VQA facts and the sampled frames are both fed to Qwen3-VL-8B, which returns the pedestrian and vehicle captions as a single JSON object.

The caption generator is a LoRA-adapted Qwen3-VL-8B [3,9]. Its inputs are three frames sampled uniformly across the event phase, the refined VQA facts, and a guided prompt in which the facts are authoritative and the frames only fill in details the facts omit, such as scene layout and weather. The LoRA is trained for three epochs on SynWTS data only, using rank 16, scale 32, and dropout 0.05, with the base weights kept fixed.

The generated caption is checked against the answer values in the prompt and regenerated up to four times when required information is missing, after which deterministic grammar and formatting cleanup is applied. If a segment has no VQA facts, the sampled overhead and ego-camera frames are provided without the structured fact blocks as an image-only fallback.

4 Experiments

4.1 Datasets

Track 2 targets the WTS traffic safety benchmark [11], a real dataset of pedestrian–vehicle interactions recorded simultaneously from several overhead (CCTV) cameras and a vehicle-mounted camera. Each interaction is split into five event phases (pre-recognition, recognition, judgement, action, and avoidance) and annotated per phase with multiple-choice questions on pedestrian, vehicle, and environment attributes and with reference pedestrian and vehicle captions. Bounding boxes accompany the videos. The public test set contains 84 scenarios, 414 caption segments, and 4,501 questions, and its labels are withheld for the official ranking.

To provide labels at scale, the challenge releases SynWTS, a fully labeled synthetic digital twin that reproduces the same interactions under the same annotation schema (questions, captions, phases, camera views, and bounding boxes) [1]. It comprises 249 scenarios (167 for training and 82 for validation) with 11,733 questions, each again captured from up to four overhead views.

4.2 Evaluation Metrics

The organizers score the two tasks with separate official metrics. VQA is evaluated by multiple-choice accuracy, and captions by BLEU-4, METEOR, ROUGE-L, and CIDEr. The final ranking uses a single S2 score, defined as the average of the VQA result and the caption result, so the two tasks contribute equally.

4.3 Implementation Details

Training configuration. The VQA predictor is trained with AdamW, and the caption adapter with LoRA on SynWTS frames and facts paired with the reference captions. Table 2 lists the full training configuration.

Table 2: Implementation settings.

Component	Setting	Value
VQA predictor	Training	AdamW, 16 epochs
	Learning rate / schedule	1.2×10^{-4} / cosine decay / 300 warmup steps
	Batch size	4
	Dropout (visual / question)	0.12 / 0.08
	InfoNCE temperature	0.07
Caption adapter	Training	LoRA, 3 epochs
	Rank / scale / dropout	16 / 32 / 0.05
	Learning rate	10^{-4}
	Batch size / maximum length	16 / 2,048 tokens
Structured decoder	Training	None
	Statistics	Estimated from SynWTS

Hardware. The VQA predictor is trained and evaluated on a single NVIDIA GeForce RTX 5060 Ti, while the caption adapter is trained and run on a single NVIDIA RTX 6000.

4.4 Official Leaderboard Results

We first compare our system with the other Track 2 teams on the official public leaderboard. As Table 3 shows, it ranks first overall with an S2 of 60.09 and a VQA accuracy of 87.09%, and scores best on every individual metric among the top five. Its lead is 2.75 S2 points over the second team and is clearest on captions, where our CIDEr of 0.84 exceeds the next best value of 0.77.

This suggests that guiding the caption model with the VQA answers, rather than letting it describe the frames freely, improves caption quality. The model still reads the video frames, but the answers steer it toward the queried attributes. Because the system is trained only on synthetic data, ranking first on the real test set also confirms that it transfers well across the sim-to-real gap.

Table 3: Official public leaderboard. The five highest-ranked teams are shown.

Rank	Team	S2	BLEU-4	METEOR	ROUGE-L	CIDEr	Acc. (%)
1	**Latent Painter - UTE**	**60.0853**	**0.2798**	**0.4624**	**0.4969**	**0.8396**	**87.0918**
2	UIT-Kitchen	57.3307	0.2658	0.4276	0.4595	0.5833	84.3812
3	KZ6	56.7949	0.2540	0.4241	0.4471	0.7691	83.5370
4	MobilityAI	55.5901	0.2532	0.4233	0.4466	0.7601	81.2042
5	Snow leopard	55.5768	0.2438	0.4247	0.4446	0.7243	81.5152

4.5 Effect of the Semantic Consistency Refinement

Because the predictor answers each question independently, its outputs can contradict one another or change implausibly across phases. This refinement resolves these inconsistencies at inference time using only SynWTS statistics, and it raises VQA accuracy from 81.28% to 87.09% on the labeled real WTS validation set, a training-free gain of 5.81 points. We use this set for evaluation only, never for training or for estimating the refinement statistics. Table 4 adds the three corrections one by one: the answer prior, which only reflects how often each answer appears in SynWTS, contributes just +0.36 points, whereas the relational and temporal refinement steps do almost all the work, adding +1.63 and +3.82 for the cumulative +5.81.

The three corrections differ in how well they transfer. The answer prior only reflects how frequent each answer is in SynWTS, a dataset-specific signal that is unlikely to hold on real video, so it is weighted low and adds little. The relational and temporal steps instead encode the structure of traffic itself, namely the logical agreement between paired questions such as inverse positions and distance, and the smooth change of an attribute across phases. These hold regardless of appearance and account for almost all of the gain, the temporal step most of it. Because every correction is bounded by the model's own scores and acts only on low-margin predictions, it fixes genuine errors where the model is unsure and leaves confident answers untouched, which is why a purely synthetic refinement still transfers.

4.6 Caption Configurations

Once the refined answers are available, the remaining question is how best to turn them into captions. We compare two options that share the same grounded facts and differ only in the language model. The first prompts a frozen Qwen3-VL-8B

Table 4: Cumulative ablation of the refinement. Each check mark adds one correction, and Δ is the gain over the previous row.

Answer prior	Question relation	Phase transition	Acc. (%)	Δ
			81.28	–
✓			81.64	+0.36
✓	✓		83.27	+1.63
✓	✓	✓	**87.09**	**+3.82**

with a few SynWTS examples that pair facts with their reference captions. The second adds a lightweight LoRA adapter trained on SynWTS frames and facts paired with the reference captions. To keep the comparison about language quality alone, VQA accuracy is held fixed at 87.09% for both rows, so any difference in S2 comes from the captions. As Table 5 shows, LoRA improves every caption metric, most clearly CIDEr from 0.58 to 0.84, and lifts S2 from 57.70 to 60.09.

The adapter is trained on synthetic frames but evaluated on real ones, so a gain measured on the real test set indicates that adaptation did not overfit the simulated appearance. We attribute this to the prompt, which keeps appearance-dependent attributes on the facts and leaves the frames only the scene details that the digital twin reproduces faithfully. The frozen setup alone already scores competitively, which shows that the refined answers, not the caption model, carry most of the content.

Table 5: Caption configurations on the public test set. S2* fixes VQA accuracy at 87.0918% for both rows.

Configuration	Acc. (%)	BLEU-4	METEOR	ROUGE-L	CIDEr	S2*
Qwen3-VL-8B (base)	87.0918	0.2221	0.4131	0.4388	0.5846	57.7025
Qwen3-VL-8B + LoRA	87.0918	0.2798	0.4624	0.4969	0.8396	**60.0853**

5 Conclusion

We presented a two-stage framework for sim-to-real traffic scene understanding trained entirely on SynWTS. A frozen V-JEPA 2.1 encoder and a Llama-based predictor retrieve structured VQA answers. A training-free module refines these answers using statistical, relational, and temporal consistency. The refined answers and sampled video frames are then provided to a LoRA-adapted Qwen3-VL-8B for grounded caption generation.

References

1. AI City Challenge: Track 2: Transportation safety understanding and captioning (Sim2Real). https://www.aicitychallenge.org/2026-track2/ (2026), accessed July 14, 2026
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., et al.: V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. Trans. Mach. Learn Res. (2024)
5. Chen, D., Shukor, M., Moutakanni, T., Chung, W., Yu, J., Kasarla, T., Bolourchi, A., LeCun, Y., Fung, P.: VL-JEPA: Joint embedding predictive architecture for vision-language. In: International Conference on Learning Representations (2026)
6. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics. pp. 310–318 (1996). https://doi.org/10.3115/981863.981904
7. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
10. Kachhadiya, R., Patil, D., Anastasiu, D.C.: Multi-agent cooperation for traffic safety description and analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 5486–5494 (2025)
11. Kong, Q., Kawana, Y., Saini, R., Kumar, A., Pan, J., Gu, T., Ozao, Y., Opra, B., Sato, Y., Kobori, N.: WTS: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. In: Eur. Conf. Comput. Vis. pp. 1–18 (2024)
12. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory **47**(2), 498–519 (2001). https://doi.org/10.1109/18.910572
13. LeCun, Y.: A path towards autonomous machine intelligence. OpenReview (2022)
14. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.243
15. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. pp. 4582–4597 (2021). https://doi.org/10.18653/v1/2021.acl-long.353

16. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.20
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023)
18. Mur-Labadia, L., Muckley, M., Bar, A., Assran, M., Sinha, K., Rabbat, M., LeCun, Y., Ballas, N., Bardes, A.: V-JEPA 2.1: Unlocking dense features in video self-supervised learning. arXiv preprint arXiv:2603.14482 (2026)
19. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
20. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beisswenger, J., Luo, P., Geiger, A., Li, H.: DriveLM: Driving with graph visual question answering. In: Eur. Conf. Comput. Vis. pp. 256–274 (2024)
21. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: DriveVLM: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
22. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 23–30 (2017). https://doi.org/10.1109/IROS.2017.8202133
23. Vera, H.S., Dua, S., Zhang, B., Salz, D., Mullins, R., et al.: EmbeddingGemma: Powerful and lightweight text representations. arXiv preprint arXiv:2509.20354 (2025)
24. Viterbi, A.J.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory **13**(2), 260–269 (1967). https://doi.org/10.1109/TIT.1967.1054010
25. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553 (2023). https://doi.org/10.18653/v1/2023.emnlp-demo.49