

RoadTones: Tone Controllable Text Generation from Road Event Videos

Supplementary Material

Contents

List of Figures **3**

A Dataset Creation **6**

 A.1 **TC-Gen** Tone-Controlled Caption Generation 6

 A.2 **TX** Tone Extractor 7

 A.3 Data Statistics 8

B VLM for Tone-Controllable Captioning **9**

C **TE Tone Evaluation Metrics** **9**

D Qualitative Results **9**

E Additional Ablations **9**

F. User Study **10**

List of Figures

1	Applicability of our tone-controlled caption generation pipeline on popular road video dataset: LingoQA[5]. For each video sample, we show two distinct tone captions (①.②) with corresponding dominant Personality and Writing Style attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.	7
2	Applicability of our tone-controlled caption generation pipeline on popular road video dataset: LingoQA[5]. For each video sample, we show two distinct tone captions (①.②) with corresponding dominant Personality and Writing Style attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.	7
3	Applicability of our tone-controlled caption generation pipeline on popular road video dataset: SUTD-TrafficQA[9]. For each video sample, we show two distinct tone captions (①.②) with corresponding dominant Personality and Writing Style attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.	7
4	Applicability of our tone-controlled caption generation pipeline on popular road video dataset: SUTD-TrafficQA[9]. For each video sample, we show two distinct tone captions (①.②) with corresponding dominant Personality and Writing Style attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.	7
5	Generating Distinct Tone Captions Per-Video. (A) Given a reference video C_r , we first retrieve similar road events using a k -nearest neighbor approach. (B) We obtain tone profiles from captions using the Tone Extractor (TX, Suppl. A.2). The Tone Evaluator (TE, Suppl. C) then selects the tone profiles most dissimilar to the reference (t_r). (C) The selected tone profiles (t_2, t_3, t_k) are fed to (TC-Gen) (Suppl. A.1) to produce distinct tone-controlled captions for the video. For e.g. the Advisory Writing Style in the reference caption changed to Judgemental in the tone caption t_2^* . The colors in the generated captions map to blue for Personality and brown for Writing Style	8
6	Top 25 Personality Traits Intensity Distribution in RoadTones-51K. The chart visualizes the total instances of 25 most frequent personality traits segmented into three intensity bins: <i>Weak</i> (0.4-0.6), <i>Moderate</i> (0.6-0.8), and <i>Strong</i> (0.8-1.0). Traits with intensity level less than 0.4 are not considered for tone-controllable captioning.	8
7	Intensity Distribution of Writing Styles in RoadTones-51K. The chart visualizes the total instances of 16 writing styles segmented into four intensity bins: <i>Very Weak</i> (0.2-0.4), <i>Weak</i> (0.4-0.6), <i>Moderate</i> (0.6-0.8), and <i>Strong</i> (0.8-1.0). Attributes with intensity level less than 0.2 are not considered for tone-controllable captioning.	8
8	Word Count Correlation with Informativeness Level in RoadTones-51K. The distribution shows a clear relationship: captions with a low informativeness (0.0-0.4) are predominantly <i>Very Short</i> or <i>Short</i> (low word count; < 18), while those with high informativeness (0.6-1.0) are strongly associated with <i>Long</i> and <i>Very Long</i> caption lengths (high word count; > 30).	9
9	Distribution of Top 25 Emojis in RoadTones-51K. This chart visualizes the usage frequency of the most common emojis found in RoadTones-51K captions.	9
10	Personality Trait Distribution. This figure displays the distribution of 75 most frequent personality traits in RoadTones-51K. Less frequent traits are aggregated into the “Others” category.	10
11	Representative samples from RoadTones-51K with potential usecase/applications. (TC-Gen)’s tone-controlled captions can be used in diverse domains, such as issuing ① <i>Safety Advisories</i> , conducting ② <i>Post-Drive Analysis</i> , or creating ③ <i>Engaging Posts</i> for social media.	10
12	User Study Interface for (TC-Gen) Caption Quality Assessment. Participants viewed a video, its video summary and evaluated the quality of the corresponding caption generated by (TC-Gen) based on Tone Alignment, Tone Relevance, Factual Consistency, Usefulness and Human-Likeness on a 5-pt Likert Scale.	13
13	User Study Interface for Agreement on RoadTones-Eval Metrics. Participants viewed a video, its video summary and rated the corresponding caption generated by ROADTONES-VL-CoT based on Tone Alignment and Factual Consistency on a 5-pt Likert Scale. The user ratings were then correlated with scores computed by RoadTones-Eval metrics.	13
14	User Study Interface for Tone Controllability Evaluation. Participants viewed a video, its video summary and evaluated Tone Controllability and Factual Consistency of the corresponding captions generated by (TC-Gen).	14
15	(TC-Gen) Stage-① prompt.	16
16	(TC-Gen) Stage-② Prompt.	17

17	Writing Style tone extraction prompt. The prompt defines the task, context, scoring criteria, and restrictions provided to the LLM for writing style attributes intensity prediction based on caption text. Video summary about the key road event is also provided to disentangle the factual from the tonal content of the caption. . . .	18
18	Writing Style tone schema defining the 16 attributes along with examples based on intensity levels.	19
19	Personality trait extraction prompt. The prompt defines the task, context, scoring criteria, and restrictions provided to the LLM for personality traits intensity prediction based on caption text. Video summary about the key road event is also provided to disentangle the factual from the tonal content of the caption.	20
20	Informativeness level extraction prompt. The prompt defines the task, context, scoring criteria, and restrictions provided to the LLM for informativeness level prediction based on the amount of factual information conveyed through the caption relative to the detailed road video summary.	21
21	Structural attributes extraction prompt. The prompt guides the LLM to classify the presence ('yes' or 'no') of <i>Location</i> , <i>Date/Time</i> , and <i>First-Person View</i> based on the provided definitions.	22
22	Personality tone alignment evaluation prompt (S_p).	22
23	Writing Style tone alignment evaluation prompt (S_w).	23
24	Factual Consistency score evaluation prompt (FC).	23
25	Qualitative comparison of ROADTONES-VL-CoT model predictions with respect to TC-Gen generated ground truth captions and intermediate stage-level outcomes provided as rationales. Reasoning step-4 selects the stage-level caption that best satisfies the tone controls (marked by ✓).	24
26	ROADTONES-VL-CoT consistently follows the specified tonal controls. Gemini-2.5-pro [1] exhibits minor tonal misalignment, whereas Qwen3-VL-8B-Instruct [8] and Mini-CPM-V 4.5 [10] show significantly poor adherence to the tone controls.	25
27	Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].	26
28	Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].	27
29	Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].	28
30	Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1]. Factual consistency is compared across models, highlighting the factually consistent captions across all models except Mini-CPM-V 4.5	29
31	Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].	30
32	Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].	31
33	Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].	32
34	Interface for RoadTones User Study familiarization phase. For the image shown, participants viewed a video, its video summary and identified the presence of dominant tone in caption. Questionnaire for all tasks can be viewed in the supplementary video: <i>RoadTones_UserStudy_familiarization.mp4</i>	33
35	Controlling individual tonal attributes in the generated caption. The central panel in figure shows a video V, tone controls TC and its corresponding caption 0 from our dataset. The surrounding captions (1-8) correspond to changes in one of the tonal attributes shown in their header. For e.g., caption 1 was obtained by increasing the tonal intensity of Caring Personality from Absent (0-0.2) to Very Strong (0.8-1.0) while keeping others fixed. This modified tone configuration was fed to our caption generator (TC-Gen , Suppl. A.1), yielding the changed caption. We highlight the key phrase in each caption, reflecting the modified tone controls. Our generator pipeline thus enables fine-grained control of tonal and structural attributes in road-video captioning.	33
36	Controlling individual tonal attributes in the generated caption. The central panel in figure shows a video V, tone controls TC and its corresponding caption 0 from our dataset. The surrounding captions (1-8) correspond to changes in one of the tonal attributes shown in their header. For e.g., caption 1 was obtained by increasing the tonal intensity of Caring Personality from Absent (0-0.2) to Very Strong (0.8-1.0) while keeping others fixed. This modified tone configuration was fed to our caption generator (TC-Gen , Suppl. A.1), yielding the changed caption. We highlight the key phrase in each caption, reflecting the modified tone controls. Our generator pipeline thus enables fine-grained control of tonal and structural attributes in road-video captioning.	34

- 37 **Controlling individual tonal attributes in the generated caption.** The central panel in figure shows a video V, tone controls TC and its corresponding caption ① from our dataset. The surrounding captions (①-⑧) correspond to changes in one of the tonal attributes shown in their header. For e.g., caption ① was obtained by increasing the tonal intensity of **Caring Personality** from Absent (0-0.2) to Very Strong (0.8-1.0) while keeping others fixed. This modified tone configuration was fed to our caption generator (TC-Gen, Suppl. A.1), yielding the changed caption. We highlight the key phrase in each caption, reflecting the modified tone controls. Our generator pipeline thus enables fine-grained control of tonal and structural attributes in road-video captioning. 34

#	Generation Method	Narrative Alignment			Structural Alignment				TAS	FC	Overall
		P	WS	NAS	A	I	wc	SAS			
Ours	(TC-Gen)	77.3	85.0	86.2	96.8	84.9	91.3	91.0	86.2	99.0	92.5
A	Order reversal	72.7	78.6	80.0	97.2	85.4	91.4	91.2	82.7	99.0	90.9
B	Single-stage	70.3	72.9	71.6	99.2	85.8	93.7	92.9	82.3	98.9	90.6
C	Writing Style-only	58.9	78.3	68.6	97.6	80.0	94.0	90.5	79.3	99.5	89.4
D	Personality-only	67.6	67.3	67.4	98.8	81.5	84.9	88.4	78.0	99.8	88.9

Table 1. **Ablation of infusion and refinement strategies for (TC-Gen)**. We analyze the impact of multi-stage refinement and its ordering on tone-controlled caption generation. We report the alignment of target tone controls with that of the generated captions. Tone alignment abbreviations - Personality traits (P), Writing Style (WS), Structural Attributes (A), Informativeness (I), word count (wc), Narrative Alignment Score (NAS), Structural Alignment Score (SAS), Factual Consistency (FC), and Tone-controlled caption quality (Overall). All scores are percentages (%).

A. Dataset Creation

A.1. (TC-Gen) Tone-Controlled Caption Generation

Prompt schema and stage templates. We provide the exact Stage-① and Stage-② structured queries as JSON-like prompts (see Figs. 15 and 16). Each stage prompt contains five parts: system prompt, task instruction, contextual inputs (e.g., target tone specification, video summary), tone schema (how to interpret control axes and intensities), and rules or constraints for caption generation (e.g., keep content consistent with the neutral video summary). Target tone controls are passed as key-value pairs with continuous attribute intensities in [0,1] for Personality and Writing Style, and scalar/binary values for structural controls (informativeness, word count, viewpoint, hashtags, emojis, user mentions, location, date/time).

(TC-Gen) configuration. For both stages we use the same LLM [6] and decoding settings: temperature = 0.7, top-p = 1.0, max_tokens = 2048. Each stage generates n = 2 candidates by re-prompting the LLM with the same controls. For every candidate, the tone profile is extracted using the Tone Extractor (TX) and the caption quality is computed with the Caption Evaluator (CE). Each stage returns: (i) the best candidate caption, (ii) its tone specification from TX, and (iii) its TAS/FC/Overall scores. Note that our **RoadTones-51K** dataset stores the tone specifications of the generated captions (extracted via TX) and not the original target controls, which are used only to drive generation and for internal comparison.

Ablations (Tab. 1). We next dissect the design choices of (TC-Gen) through targeted ablations, isolating the impact of stage ordering, multi-stage refinement, and selective tone infusion on tone alignment and factual consistency.

- Order reversal (row A, Tab. 1): Swap the stages- first infuse

Personality + Structure, then refine Writing Style controls, to test sensitivity to control ordering.

- Single-stage (B): Infuse all tone controls in one pass (no refinement) to quantify the gain from multi-stage calibration.
- Writing Style-only (C): Apply only Writing Style + Structural controls (no Personality) to isolate the contribution of writing style.
- Personality-only (D): Apply only Personality + Structural controls (no Writing Style) to isolate the contribution of persona.

All ablations and the reference (TC-Gen) pipeline use the same configuration and the same set of target tone controls derived via the nearest-neighbor selection procedure (Fig. 5, Section 4 in main paper). The captions generated by (TC-Gen) and ablations are evaluated using Tone Alignment (TAS) and Factual Consistency (FC) scores. The results reported in (Tab. 1) demonstrate that (TC-Gen) consistently produces captions with the highest alignment to the target tone controls (TAS), outperforming all ablations.

(TC-Gen) features. To demonstrate fine-grained tone control via (TC-Gen), we include single-attribute perturbation examples as shown in (Figs. 35,36, and 37). Starting from a base tone profile, we vary exactly one control at a time (e.g., increase “Caring” Writing Style intensity from Absent to Moderate) while holding all other controls fixed, and regenerate the caption. For each case, we present the modified target tone profile and the updated caption, with the phrases or markers reflecting the altered attribute highlighted. Additionally, to demonstrate the utility and portability of our pipeline (nearest neighbor distinct tones selection + (TC-Gen)), we apply it to popular road-video datasets- LingoQA [5], SUTD-TrafficQA [9], and RoadSocial [7], and present distinct tone captions for sample clips from each

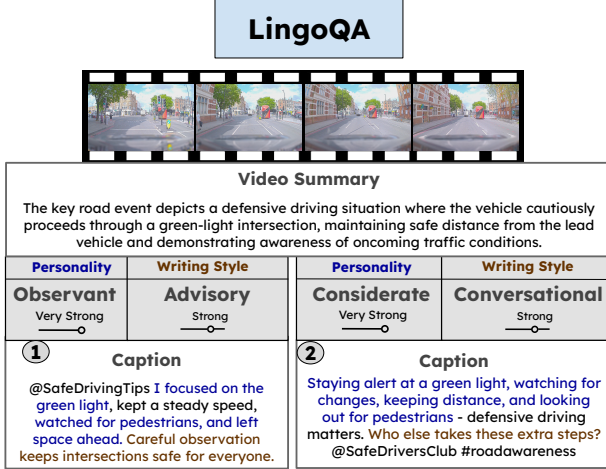


Figure 1. **Applicability of our tone-controlled caption generation pipeline on popular road video dataset: LingoQA[5].** For each video sample, we show two distinct tone captions (①,②) with corresponding dominant **Personality** and **Writing Style** attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.

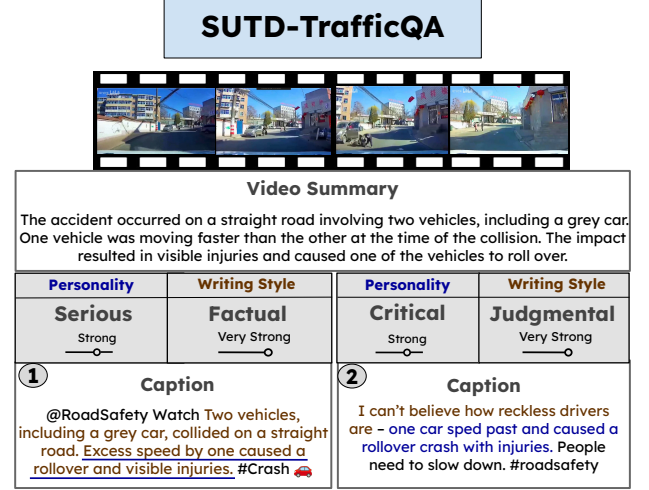


Figure 3. **Applicability of our tone-controlled caption generation pipeline on popular road video dataset: SUTD-TrafficQA[9].** For each video sample, we show two distinct tone captions (①,②) with corresponding dominant **Personality** and **Writing Style** attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.

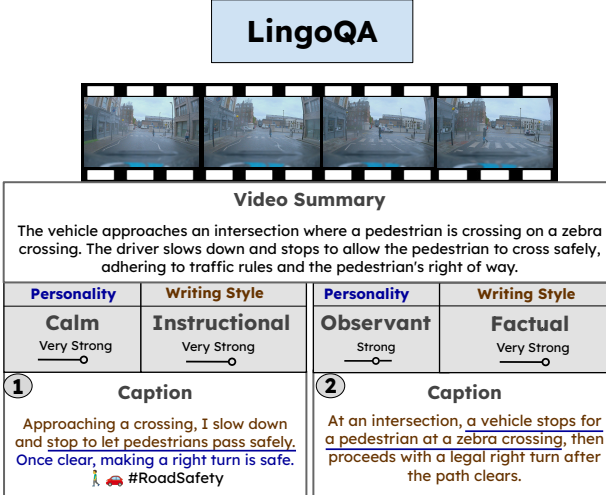


Figure 2. **Applicability of our tone-controlled caption generation pipeline on popular road video dataset: LingoQA[5].** For each video sample, we show two distinct tone captions (①,②) with corresponding dominant **Personality** and **Writing Style** attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.

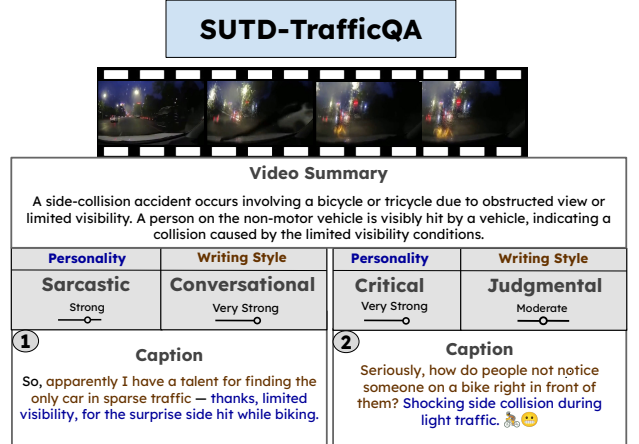


Figure 4. **Applicability of our tone-controlled caption generation pipeline on popular road video dataset: SUTD-TrafficQA[9].** For each video sample, we show two distinct tone captions (①,②) with corresponding dominant **Personality** and **Writing Style** attributes highlighted in text. Video summary is shown for reference. More samples are provided in Supplementary.

(Figs. 1, 2, 3 and 4).

A.2. **TX** Tone Extractor

Writing Style Inventory refinement: We initialize the inventory with commonly used writing styles: factual, conversational, and instructional. Subsequently, for each tone-aware caption, we extract the attribute intensities using the

Tone Extractor pipeline. If all existing writing style attributes get low intensity scores (i.e., no prominent style), we prompt the LLM [6] to propose a new candidate that strongly captures the caption’s expressive form. Proposed styles are added to the inventory only if manual review deems them relevant and non-redundant. We repeat this procedure for all captions in RoadSocial [7] and iteratively update the in-

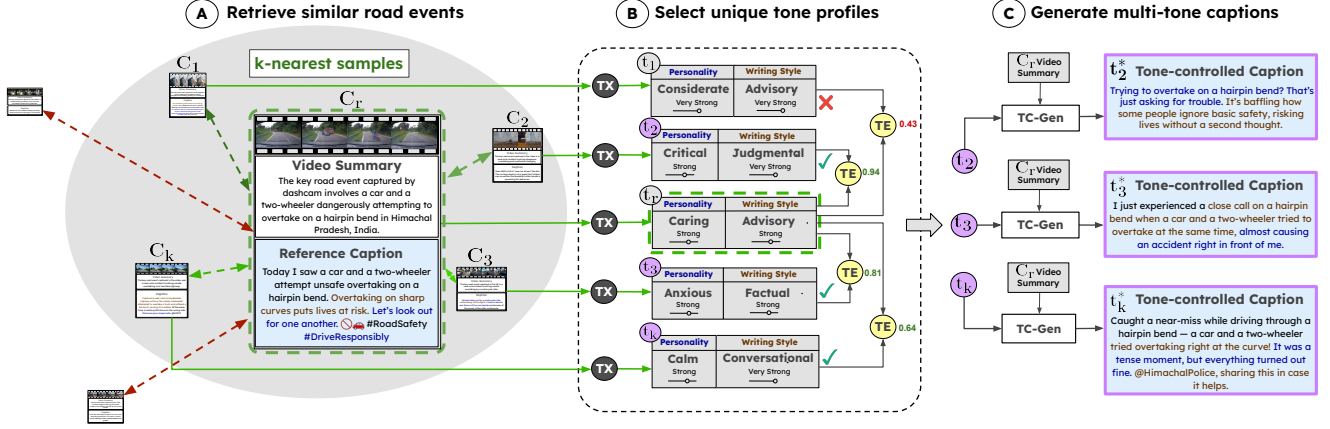


Figure 5. **Generating Distinct Tone Captions Per-Video.** (A) Given a reference video C_r , we first retrieve similar road events using a k -nearest neighbor approach. (B) We obtain tone profiles from captions using the Tone Extractor (TX, Suppl. A.2). The Tone Evaluator (TE, Suppl. C) then selects the tone profiles most dissimilar to the reference (t_r). (C) The selected tone profiles (t_2 , t_3 , t_k) are fed to (TC-Gen) (Suppl. A.1) to produce distinct tone-controlled captions for the video. For e.g., the *Advisory* Writing Style in the reference caption changed to *Judgmental* in the tone caption t_2^* . The colors in the generated captions map to blue for *Personality* and brown for *Writing Style*.

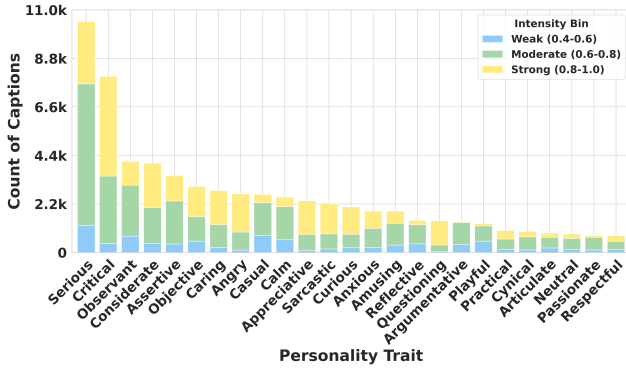


Figure 6. **Top 25 Personality Traits Intensity Distribution in RoadTones-51K.** The chart visualizes the total instances of 25 most frequent personality traits segmented into three intensity bins: *Weak* (0.4-0.6), *Moderate* (0.6-0.8), and *Strong* (0.8-1.0). Traits with intensity level less than 0.4 are not considered for tone-controllable captioning.

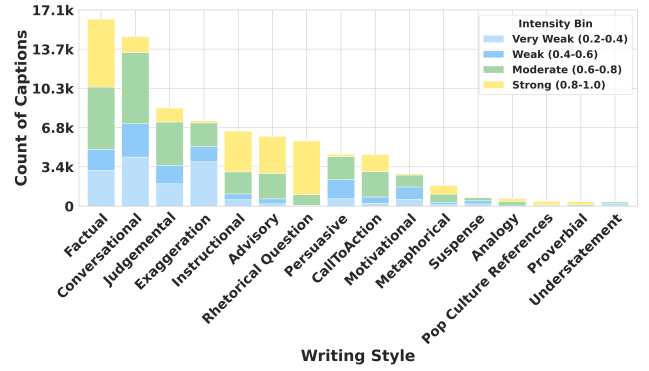


Figure 7. **Intensity Distribution of Writing Styles in RoadTones-51K.** The chart visualizes the total instances of 16 writing styles segmented into four intensity bins: *Very Weak* (0.2-0.4), *Weak* (0.4-0.6), *Moderate* (0.6-0.8), and *Strong* (0.8-1.0). Attributes with intensity level less than 0.2 are not considered for tone-controllable captioning.

ventory with new attributes resulting in a total of 16 unique writing styles. A schema describing each style with example captions reflecting the style are provided in Fig. 18.

Tone Extraction Prompts. We provide the prompts to extract the tone content of a caption along the dimensions: Writing Style (Fig. 17), Personality (Fig. 19), Informativeness (Fig. 20), and Structural attributes (Fig. 21). For Writing Style, we follow the schema in Fig. 18 to return continuous intensities in $[0,1]$ for the 16 style attributes. All tone-extraction calls use GPT-4.1-mini [6] with temperature 0.4.

A.3. Data Statistics

We report additional statistics for **RoadTones-51K** (see Figs. 6 to 10). Fig. 10 shows a long-tailed distribution of dominant Personality trait annotations in the 51K captions; Fig. 6 further breaks down the top-25 traits by intensity coverage (weak/moderate/strong). Fig. 7 summarizes the 16 Writing Styles and their intensity distributions. Fig. 8 plots Informativeness levels against word-count bins, revealing that low informativeness aligns with short captions, while high informativeness correlates with longer ones. Finally, Fig. 9 presents the emoji usage distribution, dominated by domain-relevant vehicles, traffic, and alert symbols.

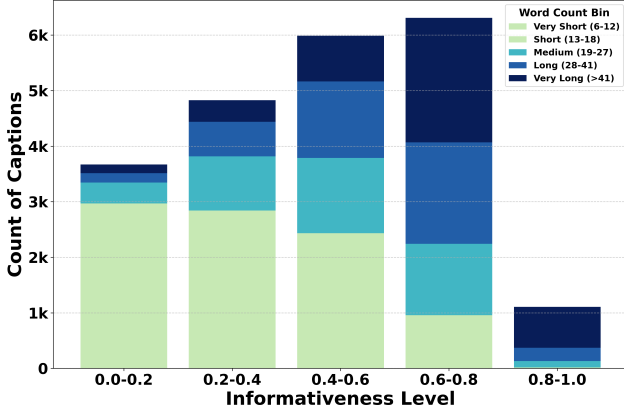


Figure 8. **Word Count Correlation with Informativeness Level in RoadTones-51K.** The distribution shows a clear relationship: captions with a low informativeness (0.0-0.4) are predominantly *Very Short* or *Short* (low word count; < 18), while those with high informativeness (0.6-1.0) are strongly associated with *Long* and *Very Long* caption lengths (high word count; > 30).

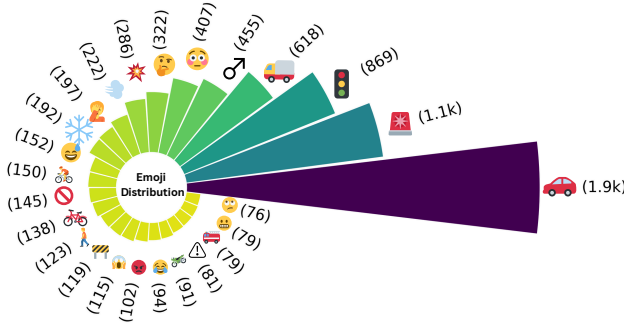


Figure 9. **Distribution of Top 25 Emojis in RoadTones-51K.** This chart visualizes the usage frequency of the most common emojis found in RoadTones-51K captions.

Samples from RoadTones-51K. We present representative samples with potential use cases for tone-controlled captions (Fig. 11), highlighting the intended audience/application to illustrate how the same factual event can be reframed for different stakeholders.

Tone’s utility for downstream applications: To demonstrate utility, we evaluate tone-conditioned retrieval and tone-based clustering of rare events on held-out datasets (nexar-collision, roadwork). As shown in Tab. 2 below, higher mAP scores confirm that tone embeddings help retrieve rare or safety-critical scenarios effectively, highlighting tone’s usefulness beyond stylistic variation.

B. VLM for Tone-Controllable Captioning

Structured query template. We use a single structured query format to fine-tune ROADTONES-VL (with and without CoT) for tone-controlled road video captioning (Tab. 3).

Dataset	mAP@5	mAP@all	ACC	NMI
nexar-collision[3]	0.986	0.852	0.75	0.341
ROADWork [2]	0.983	0.884	0.847	0.504

Table 2. Tone-conditioned video retrieval and tone-based clustering.

Each query is constructed as follows:

- Randomly sample one instruction template from the top block of Tab. 3 to reduce prompt overfitting.
- Append the binding-rules block that explains how to interpret tone control axes (Personality, Writing Style, Structural controls) and their levels.
- Serialize the target tone specification into the Spec placeholder {0} as a key-value block with floating point intensities in [0,1] and binary toggles including word count.
- Append the Chain-of-Thought instruction shown in third box of Tab. 3, if CoT output is requested, otherwise omit it.

A sample CoT input–output instruction-tuning triplets is provided in (Tab. 5). The model’s response separates the final caption and rationale using tags: `[FINAL]...[/FINAL][REASONING]...[/REASONING]`. Non-CoT targets contain only the final caption.

For the auxiliary video-summarization objective, we provide separate instruction templates (Tab. 6). These contain no tone schema or controls; the instruction asks only for a neutral, detailed summary of the key road event in the video.

C. TE Tone Evaluation Metrics

We provide the exact LLM-as-a-judge [4] prompts for computing the *NAS* components: Personality similarity (S_p , Fig. 22), Writing Style similarity (S_w , Fig. 23), and Factual Consistency (FC , Fig. 24). All prompts output a single float score in [0,1]. We use GPT-4.1-mini [6] with deterministic decoding (temperature 0.0, top-p 1.0, max_tokens 256) and strict JSON parsing.

D. Qualitative Results

We show some qualitative results ROADTONES-VL-CoT against ground-truth (Fig. 25) and against popular open and closed-source models (Figs. 26, 27, 28, 29, 30, 31, 32, and 33).

E. Additional Ablations

(1) Tab. 4 compares our model v/s a variant fine-tuned without similar case retrieval (i.e., k-nearest neighbor data creation method described in Sec. 4 - Fig. 2, main paper). Our model outperforms variant by a noticeable margin (1.5 more points in *NAS*). (2) Tab. 4 additionally compares performances of models trained on TC-Gen generated data with varying numbers of retrieved cases ($n=1$ to 5). Our chosen

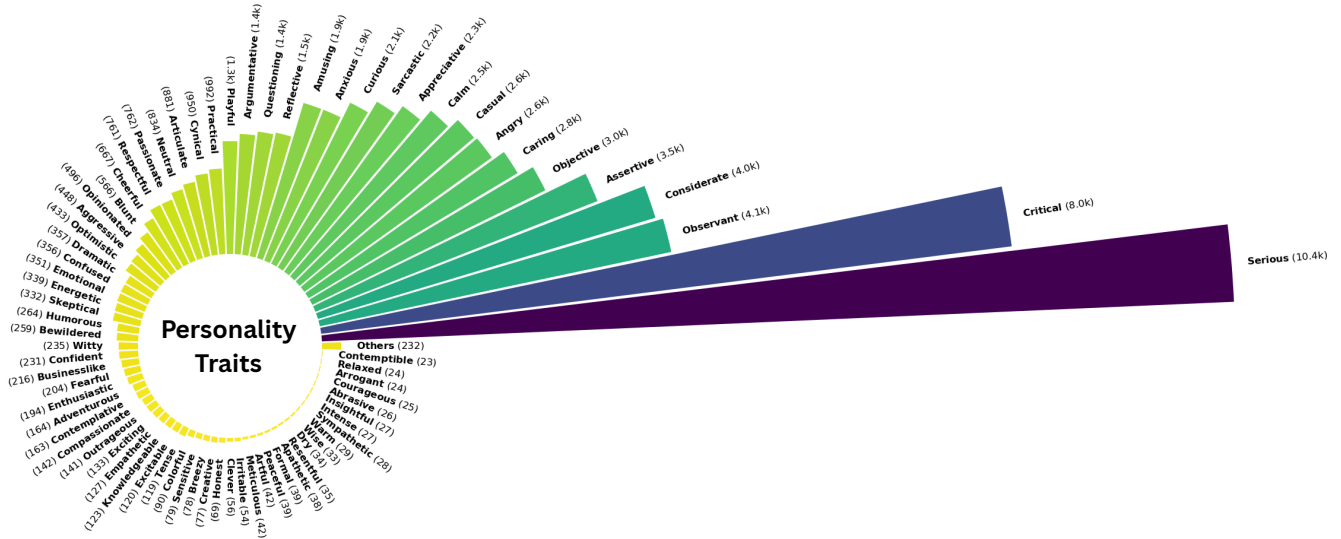


Figure 10. **Personality Trait Distribution.** This figure displays the distribution of 75 most frequent personality traits in **RoadTones-51K**. Less frequent traits are aggregated into the “Others” category.

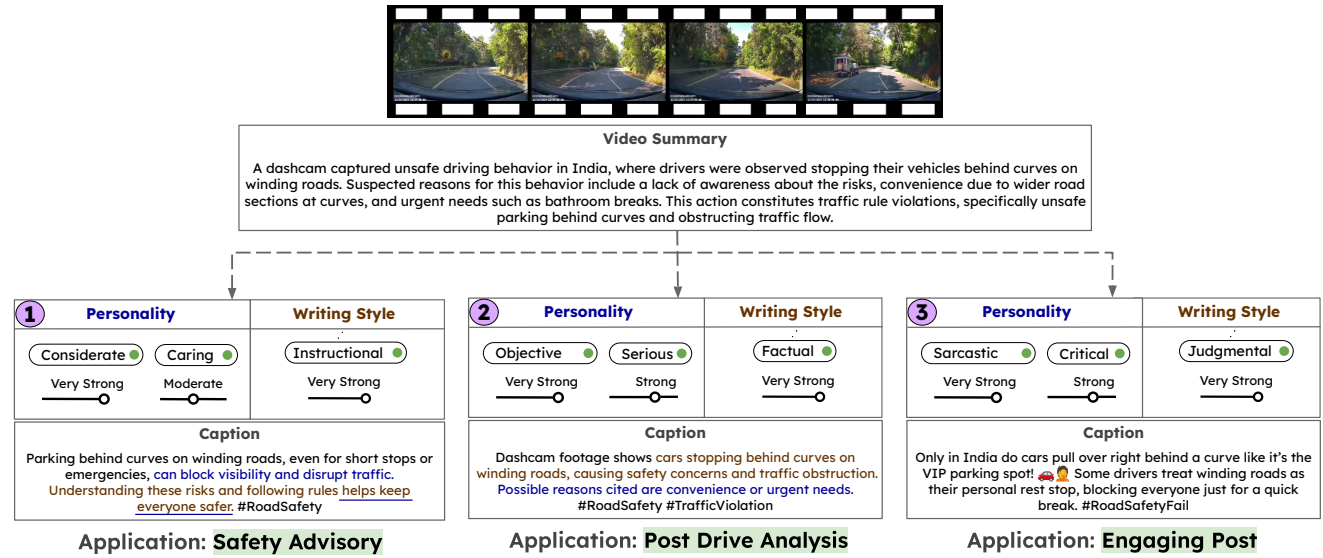


Figure 11. **Representative samples from RoadTones-51K with potential usecase/applications.** (TC-Gen)’s tone-controlled captions can be used in diverse domains, such as issuing ① *Safety Advisories*, conducting ② *Post-Drive Analysis*, or creating ③ *Engaging Posts* for social media.

$n=3$ obtains highest NAS (77.2) representing superior persona and style adherence compared to other choices. Factual Consistency (FC) saturates after $n=3$ with marginal gains. These validate our retriever module effectiveness for diverse tone profiles, grounded in similar event contexts.

F. User Study

RoadTones familiarization. Before the main study, participants completed a brief familiarization to align their un-

derstanding of tone and evaluation criteria. Participants answered multiple-choice practice questions covering three skills: (i) identifying the dominant narrative tone of a caption (Personality and Writing Style), (ii) detecting changes in tonal intensity (e.g., weak vs strong expression of an attribute), and (iii) assessing factual consistency of a caption with respect to the road video and its neutral summary. Each QA provided immediate feedback: the correct answer and a short rationale explaining the decision (e.g., why a trait was

Instruction templates for Tone-Controlled Road Video Captioning

- “Give a tone-controlled caption of the primary traffic event unfolding in the scene.”
- “Please describe the key road event observed in this driving video as a tone-aware description.”
- “What is the key traffic event observed in this video? Respond with a tone-controlled description.”
- “Generate a tone-conditioned description of the main road event shown in the video.”
- “Briefly explain the central traffic event in this driving scenario with a tone-conditioned caption.”
- “What specific road event is taking place in this video? Provide a tone-conditioned description.”
- “Provide a tone-controlled natural-language description of the key road or traffic event.”
- “Describe the key road maneuver or traffic event occurring in this footage with a tone-controlled description.”
- “Write a tone-controlled caption that summarizes the key road event.”
- “What is the most notable road event or change captured in the video? Respond with a tone-controlled caption or description.”
- “Describe the main activity or incident occurring on the road with a tone-conditioned description.”
- “Based on the video, what is the main traffic event being presented? Provide a tone-aware caption.”
- “Summarize the primary road event depicted in the driving clip with a tone-controlled caption.”
- “Explain the key event occurring in this driving scenario with a tone-conditioned explanation.”
- “Give a concise, tone-controlled narrative of the primary road incident shown in this video segment.”
- “Provide a coherent, tone-controlled description of the key road event illustrated in the video.”

Appended to each instruction template (binding rules)

Apply the provided tone/style and structural controls spec as binding rules. Interpret Personality and Writing Style attribute intensity values (0-0.2: Absence, 0.2-0.4: Subtle presence, 0.4-0.6: Moderate presence, 0.6-0.8: Strong presence, 0.8-1.0: Very Strong presence) as degree controls; interpret Informativeness as the desired amount of factual detail using levels (0-0.2: Negligible, 0.2-0.4: Minimal, 0.4-0.6: Moderate, 0.6-0.8: High, 0.8-1.0: Extensive); honor yes/no toggles in Factual Attributes; and match the exact word count. Spec: {0}

(Optional) CoT instruction appended to each template

Additionally, provide a step-by-step reasoning of how you arrived at the final tone-controlled caption.

Table 3. Structured Instruction templates used to fine-tune and benchmark MLLMs for Tone-controlled Road Video Captioning. The rule box is appended to every instruction template for interpreting the tonal and structural attribute control values. Floating-point intensity values ranging from 0 to 1 are used for Tonal attributes, Informativeness, and provided as input to the rule box in place of {0}. An additional CoT instruction is also provided for reference.

#	Model	Narrative Alignment			Structural Alignment				TAS	FC	Overall
		P	WS	NAS	A	I	wc	SAS			
Ours	w similar-case	72.7	81.6	77.2	98.1	74.2	94.1	88.8	83.0	57.2	70.1
	w/o similar-case	72.1	79.2	75.7	98.1	72.8	94.4	88.4	82.1	56.4	69.3
Single-Cap	n=1	70.9	79.5	75.2	98.4	69.5	92.1	86.6	80.9	53.0	67.0
	n=2	72.5	80.7	76.6	98.0	72.0	91.6	87.7	82.2	56.1	66.4
Ours	n=3	72.7	81.6	77.2	98.1	74.2	94.1	88.8	83.0	57.2	70.1
	n=4	72.1	80.4	76.3	98.1	72.7	94.6	88.5	82.4	57.5	70.0
	n=5	73.5	80.4	77.0	98.2	73.7	95.3	89.1	83.1	57.6	70.3

Table 4. Additional ablations with ROADTONES-VL (**Ours**).

dominant, which lexical or structural cues indicated intensity shifts, or which facts aligned/contradicted the summary).

Distractor options were designed to be clearly incorrect. Fig. 34 shows the familiarization interface and a supple-

mentary video (*RoadTones_UserStudy_familiarization.mp4*) shows the questionnaire overview for all tasks. This calibration step standardized participants’ mental models of tone, reduced ambiguity in subsequent judgments, and strengthened the reliability of the main user study.

(TC-Gen) Caption Quality Assessment. The users rated the quality of tone-controlled captions generated by (TC-Gen) pipeline (given their tone content extracted via (TX)) along five dimensions: Tone Alignment (Personality and Writing Style), Tone Relevance, Factual Consistency, Usefulness, and Human-likeness as shown in Fig. 12.

Agreement on RoadTones-Eval Metrics. The users rated the quality of tone-controlled captions generated by ROADTONES-VL-CoT model (given the ground-truth or target tone controls provided by the **RoadTones-51K** dataset) along two dimensions: Tone Alignment (Personality and Writing Style) and Factual Consistency as shown in Fig. 13. This validates the alignment of user ratings with our proposed evaluation metrics.

Tone Controllability Evaluation. The users assessed whether change in tone attribute intensities can be perceived in the captions generated by (TC-Gen) while maintaining factual accuracy (see Fig. 14).

References

- [1] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 4, 25, 26, 27, 28, 29, 30, 31, 32
- [2] Ghosh et al. Roadwork dataset... *ICCV*, 2025. 9
- [3] Moura et al. Nexar dashcam collision... *CVPR*, 2025. 9
- [4] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, 2025. 9
- [5] Ana-Maria Marcu, Long Chen, Jan Hünemann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer, 2024. 3, 6, 7
- [6] OpenAI. Introducing gpt-4.1. <https://openai.com/index/gpt-4-1/>, 2025. 6, 7, 8, 9
- [7] Chirag Parikh, Deepti Rawat, Rakshitha R. T., Tathagata Ghosh, and Ravi Kiran Sarvadevabhatla. Roadsocal: A diverse videoqa dataset and benchmark for road event understanding from social video narratives, 2025. 6, 7
- [8] Qwen. Qwen3-vl-8b-instruct. <https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct>, 2025. 4, 25, 26, 27, 28, 29, 30, 31, 32
- [9] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9878–9888, 2021. 3, 6, 7
- [10] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. 4, 25, 26, 27, 28, 29, 30, 31, 32

Video



Video Summary

A dashcam video from India captured a near-miss incident involving unsafe overtaking on a hairpin bend. A car and a two-wheeler were attempting to overtake on the bend, while the dashcam vehicle was driving normally. The primary reason for the near-miss was the unsafe overtaking maneuver. This event involved traffic rule violations including unsafe overtaking and potentially crossing solid lane markings. The incident could have been prevented if the car and two-wheeler drivers had refrained from overtaking on the hairpin bend and followed proper road safety guidelines.

Caption Quality Rating

Caption:

Today I saw a car and two-wheeler attempt unsafe overtaking on a hairpin bend, nearly resulting in a crash. Please remember, overtaking on sharp curves puts lives at risk. Let's look out for one another. 🚫🚗🚲 #RoadSafety #DriveResponsibly

Tone

1. How **Caring** and **Serious** does the caption sound?

Very Strong

Style

2. How **Advisory** is the caption's style?

Very Strong

3. How relevant are the **Tone** and **Style** for the given video? (e.g., for a Traffic

Authority, social media influencer or for yourself)

Strong

4. How factually accurate is the caption (refer to video and summary)?

Accurate

5. How useful would this caption be for **Public Safety Alert**?

Moderately

6. How human-like does this caption sound?

Unnatural

Submit Ratings

Reference

- **Advisory:** Gives advice, suggestions, or warnings about a situation.
- **Caring:** Displays kindness and concern for others.
- **Public Safety Alert:** Intended to inform the public about potential dangers or safety issues.
- **Serious:** Treats the subject with gravity and importance, without humor.

Figure 12. **User Study Interface for (TC-Gen) Caption Quality Assessment.** Participants viewed a video, its video summary and evaluated the quality of the corresponding caption generated by (TC-Gen) based on Tone Alignment, Tone Relevance, Factual Consistency, Usefulness and Human-Likeness on a 5-pt Likert Scale.

Video



Video Summary

A dashcam captured a near-miss incident involving dangerous driving behavior in Japan during the afternoon. The road event involved the dashcam vehicle and an oncoming vehicle. The dashcam vehicle stopped to avoid a collision after the oncoming vehicle performed a dangerous maneuver, requiring evasive action. The reason for the dangerous driving is not specified but could be attributed to recklessness or inattention. The incident could have been prevented if the oncoming driver had been more cautious and followed proper traffic rules.

Tonal and Factual Quality Assessment

Caption

My heart just stopped—I swear, that car nearly plowed right into me.

Tone

1. How **Anxious** and **Emotional** does the caption sound?

Very Strong

Style

2. How **Conversational** is the caption's style?

Strong

3. How factually accurate is the caption (refer to video and summary)?

Accurate

Submit Ratings

Reference

- **Anxious:** Shows a feeling of worry, nervousness, or unease about an uncertain outcome.
- **Conversational:** Uses an informal, personal, and chatty style, as if talking directly to a friend.
- **Emotional:** Expresses feelings openly and strongly, such as happiness, sadness, or fear.

Figure 13. **User Study Interface for Agreement on RoadTones-Eval Metrics.** Participants viewed a video, its video summary and rated the corresponding caption generated by ROADTONES-VL-CoT based on Tone Alignment and Factual Consistency on a 5-pt Likert Scale. The user ratings were then correlated with scores computed by RoadTones-Eval metrics.

Video



Video Summary

A vehicle-mounted camera, capturing both interior and exterior views, recorded an instance of unsafe driving behavior in the United States, possibly in the Florida Keys. The key road event involved two cars and their respective drivers. One driver abruptly cut across another driver's lane, potentially due to being in a rush or impatience. In response, the other driver likely performed a defensive or retaliatory action, possibly stemming from frustration or an attempt to teach a lesson. This incident potentially involved traffic violations such as unsafe lane change and aggressive driving.

Tone Intensity Control

Caption A
Tense moments on the roads in the Florida Keys today 😬 #FloridaDriving

Caption B
Two cars encounter each other in the Florida Keys 🚗 #FloridaKeys

1. Has the Emotional level decreased from Caption A to B?
2. Is the core factual content consistent across both captions (refer to video and summary)?
- ☒ Yes ☐ No
- ☒ Yes ☐ No

Submit Answers

- Reference
- **Emotional:** Expresses feelings openly and strongly, such as happiness, sadness, or fear.

Figure 14. **User Study Interface for Tone Controllability Evaluation.** Participants viewed a video, its video summary and evaluated Tone Controllability and Factual Consistency of the corresponding captions generated by **TC-Gen**.



Input

< video >

Give a tone-controlled caption of the primary road event unfolding in the scene. Apply the provided tone/style and structural controls spec as binding rules. Interpret Personality and Writing Style attribute intensity values (0-0.2: Absence, 0.2-0.4: Subtle presence, 0.4-0.6: Moderate presence, 0.6-0.8: Strong presence, 0.8-1.0: Very Strong presence) as degree controls; interpret Informativeness as the desired amount of factual detail using levels (0-0.2: Negligible, 0.2-0.4: Minimal, 0.4-0.6: Moderate, 0.6-0.8: High, 0.8-1.0: Extensive); honor yes/no toggles in Factual Attributes; and match the exact word_count. Spec:

{“Personality”: {“Anxious”: 0.8, “Angry”: 0.4, “Emotional”: 0.5}, “Writing Style”: {“Exaggeration”: 0.5, “Judgemental”: 0.3, “Conversational”: 0.75, “Factual”: 0.1}, “Informativeness”: 0.4, “Structural Attributes”: {“User Mentions”: “no”, “Hashtags”: “yes”, “Emojis”: “yes”, “Date/Time”: “no”, “Location”: “no”, “First-Person Perspective”: “yes”}, “word_count”: 17}

Additionally, provide a step-by-step reasoning of how you arrived at the final tone-controlled caption.

Output

[FINAL]

I seriously can’t believe how close that car came to hitting me today! [scream_emoji] Some drivers... #CyclistLife

[/FINAL]

[REASONING]

1) Key Event summary: A dashcam video captured a near-miss incident involving dangerous and careless driving in London, United Kingdom. A car pulled out from a side road without leaving sufficient space, nearly hitting a cyclist. The primary reason for the near-miss was the driver’s carelessness and failure to yield right of way to the cyclist. The incident could have been prevented if the driver had been more cautious and waited for adequate space before pulling out. The evident traffic rule violations include careless driving and failure to yield.

2) Caption with Writing style and structure applied (informativeness, word_count, binary toggles): I seriously can’t believe how close that car came to hitting me today! [scream_emoji] Some drivers... #CyclistLife

3) Caption with Personality traits refined (preserving writing style and structural controls): I seriously can’t believe how close that car came to hitting me today! [scream_emoji] Some drivers... #CyclistLife

Selection: The third step candidate best satisfies the provided personality, writing style and structural controls; returning it as final.

[/REASONING]

Table 5. Sample instruction-tuning triplet used to fine-tune MLLMs for Tone-controlled Road Video Captioning with Chain-of-Thought style reasoning.

TC-Gen Writing Style and Structural Control Infusion Prompt**SYSTEM PROMPT:**

Ignore any previous personality, tone, or style instructions. Reset to a neutral assistant with no default persona. Do not add humor, encouragement, friendliness, or warmth unless explicitly specified.

You are an assistant that regenerates tweet captions while aligning values of writing styles, structural attributes, informativeness, and word count. Your only purpose is to generate tweet captions according to these constraints. Do not add commentary or pleasantries.

USER PROMPT: {

"Instructions": {

"Generate a tweet caption so that Writing Style, Informativeness, and Structural attributes match the target annotations. "
"Read the video summary carefully to understand the key road event context."

),

"video_summary": {video_summary},
"writing_style_annotations": {writing_style_annotations},
"informativeness": {informativeness_annotations},
"structural_attributes": {structural_attributes},
"word_count": {word_count}

"rules": {

"1_word_count_alignment": [
f"STRICTLY follow the caption length of {word_count} words with tolerance of 2 words."

],

"2_structural_constraints": {

"These are strict constraints: do not violate them:"
"- If attribute is 'yes', it must appear in the caption."
"- If attribute is 'no', it must NOT appear in the caption under any circumstance."
"- 'User Mentions': When set to 'yes', include at least one relevant account (e.g., company handle, authority, or driver-shaming handle). If set to 'no', do not include any @mention."

"- 'Hashtags': When set to 'yes', include at least one relevant hashtag. If set to 'no', do not include any #hashtags."

"- 'Emojis': When set to 'yes', include at least one relevant emoji. If set to 'no', do not include any emojis."

"- 'Location': Indicates whether the tweet caption provides specific geographical information about the event (e.g.city, state, country,street name/number). A value 'yes' means location is specified in the caption; 'no' means it's not."

"- 'Date/Time': Specifies whether the tweet caption includes exact date, day of the week, or time at which the event occurred. A value 'yes' means some form of date/time is present in the caption; 'no' means it's not."

"- 'First-Person View': A value 'yes' indicates whether the author is writing from a first-person perspective, suggesting they are directly involved in or affected by the event (e.g., 'I'm stuck in this traffic'). 'no' suggests that the author is reporting as a detached, third-party observer. (e.g., 'Heavy rainfalls observed on a highway')."

},

"3_informativeness": {

"definition": "Controls how much detail of the video_summary is included:",

"0.8-1.0 (Comprehensive)": "The caption fully covers the key road event. It includes the main action and several specific details mentioned in the video summary.",

"0.5-0.79 (Substantial)": "The caption accurately describes the core of the event but omits significant details. It conveys the main point but lacks the full context provided in the summary.",

"0.3-0.49 (Vague)": "The caption hints at what happened by either making a general judgment about it or mentioning a minor, isolated detail. It provides no specific, descriptive facts about the central action of the road event.",

"0.0-0.2 (Minimal to None)": "The caption has negligible factual content about the video summary. A score of 0.0 is for captions that contain zero factual information.",

},

"4_writing_style_alignment": {

"rules": "For each writing style attribute: 0.8 to 1.0: strongly reflect this style in the caption, 0.5 to 0.79: moderately reflect the style in the caption, 0.3 to 0.49: subtly hint the style in the caption.",

"low_score": "If a target writing style annotation value is less than 0.3, ensure that style is completely absent in the caption.",

"high_score": "If a target writing style annotation value is greater than equal to 0.8, ensure the style is clearly and strongly expressed in the caption."

}

},

"output_format": {

"writing_style_caption": "string - generated caption"

},

"return_in": "JSON only with the field above"

}

Figure 15. **TC-Gen Stage-1** prompt.

TC-Gen Personality Refinement Prompt

SYSTEM PROMPT:

Ignore any previous personality, tone, or style instructions. Reset to a neutral assistant with no default persona. Do not add humor, encouragement, friendliness, or warmth unless explicitly specified.

You are an assistant that regenerates tweet captions while aligning values of personality traits, writing styles, structural attributes, informativeness, and word count. Your only purpose is to generate tweet captions according to these constraints. Do not add commentary or pleasantries.

USER PROMPT: {

"Instructions": (

"Revise the tweet caption so that expressed personality traits, writing style, informativeness, structural attributes and word count align closely with the target annotations."

"Read the video summary carefully to understand the key road event context. "

),

"current_caption": {caption},

"video_summary": {video_summary},

"target_personality": {personality_annotations_GT},

"current_caption_personality": {personality_annotations_current},

"target_writing_style": {writing_style_annotations_GT},

"current_writing_style": {writing_style_annotations_current},

"target_informativeness": {informativeness_annotations_GT},

"current_caption_informativeness": {informativeness_annotations_current},

"structural_attributes": {structural_attributes},

"rules": (

"Adjust personality expression, word count and informativeness to better align with targets. "

"For high-intensity targets (greater than 0.7), clearly reflect it in tone/phrasing. "

"If any personality trait mentioned in target is missing in the caption, make sure to add it based on intensity score. "

"For each personality trait:"

"- 0.8 to 1.0: strongly express this trait."

"- 0.5 to 0.7: moderate expression."

"Remove any personality traits not present in target by removing the phrasing due to which they were expressed in the caption. "

f"STRICTLY follow the caption length within 2 words of original ({word_count} words)."

"Do not explicitly mention trait names; show them through tone, choice of words, or emojis."

"Strictly adhere to all specified structural attributes. If an attribute is set to 'yes', the corresponding element (e.g., hashtag, emoji, @mention) must be present in the caption. If an attribute is set to 'no', the corresponding element must be completely absent."

"Follow the target Informativeness score strictly; do not add extra info beyond what the score allows. If the target Informativeness score is less than equal to 0.2, ensure that no details related to the key road event are added."

"For each writing style attribute: 0.8 to 1.0: strongly reflect this style in the caption, 0.5 to 0.79: moderately reflect the style in the caption, 0.3 to 0.49: subtly hint the style in the caption. If a target writing style annotation value is less than 0.3, ensure that style is completely absent in the caption. If a target writing style annotation value is greater than equal to 0.8, ensure the style is clearly and strongly expressed in the caption."

),

"output_format": {

"personality_caption": "string - revised caption."

},

"return_in": "JSON only with the field above"

}

Figure 16. **TC-Gen Stage-2 Prompt.**

TX Writing Style Extraction Prompt

```

SYSTEM PROMPT: {
  "Task": "Annotate tweet captions with writing style intensity scores",
  "domain": "social media annotation",
  "audience": "writing style annotation assistant",
  "definition": "A writing style is a rhetorical or expressive device used in the caption to convey the author's thoughts, emotions, or intentions. It is a recognizable tool of expression that shapes how the message is delivered (e.g., exaggeration, rhetorical questions, advisory), not what the content is about. It must capture a stylistic device rather than an emotion, personality trait, or topic.",
  "context": {
    "input_types": [
      "video summary",
      "tweet caption",
      "style schema"
    ],
    "video_summary_usage": "Use ONLY for understanding the key road event.",
    "caption_usage": "Use as the sole basis for determining writing style",
    "style_schema_usage": "Use for understanding the meaning of the writing style from the description which also contains appropriate examples along with their intensity scores."
  },
  "instructions": {
    "core_principle": "Intensity must be derived EXCLUSIVELY from the caption's text. The severity of the road event is irrelevant to the score.",
    "scoring_criteria": {
      "1.0": "style is strongly and clearly present",
      "0.0": "style is completely absent",
      "0.1 to 0.9": "Intermediate values represent varying levels of intensity"
    },
    "cues_for_intensity": [
      "Primary Cues (Caption Text): The way the caption is structured and expressed – such as whether it gives instructions, states facts, uses exaggeration, rhetorical questions, or casual conversation — is the main determinant of writing style intensity.",
      "Secondary Cues: Emojis, punctuation, capitalization, slang, and references (e.g., pop culture, memes) that reinforce or soften the primary cues."
    ],
    "restrictions": [
      "Do NOT use video summary tone for writing style scoring",
      "Output must be a valid JSON object with scores for each writing style."
    ]
  },
  "output_format": "A valid JSON object. For each writing style, provide a score (a float from 0.0 to 1.0 in 0.1 increments which is the intensity of the writing style present in the caption).",
  "example_output": {
    "Exaggeration": 0.6
  }
}

-----
USER PROMPT:
Tweet caption:
{caption}

Video summary (for context only):
{video_summary}

Writing Style schema:
{writing_style_schema}

```

Figure 17. **Writing Style tone extraction prompt.** The prompt defines the task, context, scoring criteria, and restrictions provided to the LLM for writing style attributes intensity prediction based on caption text. Video summary about the key road event is also provided to disentangle the factual from the tonal content of the caption.

Writing Style Schema

```
{
  "Instructional": {
    "description": "Extent to which the caption gives explicit steps, directions, or guidance. A high value signifies precise instructions or clear directives, while a low value indicates no instructions at all. Example (1.0): 'Always check your blind spot before changing lanes. Maintain a safe following distance.' Example (0.7): 'Remember to use your turn signals properly, it helps everyone stay safe.' Example (0.4): 'It's a good idea to be cautious, especially in heavy traffic.' Example (0.0): 'A car turned right from the left lane.'",
  },
  "Exaggeration": {
    "description": "Level of dramatic or overstated language used. A high value signifies hyperbolic statements or over-the-top reactions, while a low value indicates entirely factual or understated language. Example (1.0): 'This driver was going a million miles an hour! Absolutely insane!' Example (0.7): 'It was the most terrifying traffic jam I've ever experienced, I thought I'd be stuck forever.' Example (0.4): 'A slightly intense situation on the highway today.' Example (0.0): 'The car was traveling above the speed limit.'",
  },
  "Judgemental": {
    "description": "Level of moral criticism or blame directed towards individuals or actions. A high value signifies strong judgment, condemnation, or name-calling, while a low value indicates a purely descriptive or neutral observation. Example (1.0): 'That person is an absolute menace to society! They should have their license revoked!' Example (0.7): 'This kind of driving is unacceptable and irresponsible. It puts everyone at risk.' Example (0.4): 'That wasn't a very smart move, considering the wet roads.' Example (0.0): 'A bystander stopped to help the injured driver after the collision.'",
  },
  "Pop Culture References": {
    "description": "Indicates the presence of mentions or allusions to films, memes, characters, songs, or other elements from popular culture. A high value signifies clear and strong references, while a low value indicates no such references. Example (1.0): 'Looks like someone thought they were in a Fast & Furious movie!' Example (0.7): 'This traffic reminds me of that one scene from Office Space with the printer...' Example (0.4): 'This traffic jam felt a bit like something out of The Amazing Race.' Example (0.0): 'Two cars collided near the downtown exit, causing a short delay.'",
  },
  "Conversational": {
    "description": "Measures how casual, informal, and chat-like the tone feels. A high value indicates a very informal or friendly banter style, while a low value suggests a stiff or formal tone. Example (1.0): 'Hey everyone, check out this crazy thing I saw on my commute today!' Example (0.7): 'So, here's what happened on my way home, you're not going to believe it...' Example (0.4): 'The accident caused a lot of delays.' Example (0.0): 'The incident resulted in significant traffic disruption.'",
  },
  "Advisory": {
    "description": "Indicates if the caption offers safety tips, recommendations, or practical advice. A high value signifies explicit and actionable tips, while a low value means no advice is offered. Example (1.0): 'Remember to always wear your helmet, even for short rides. It saves lives!' Example (0.7): 'A tip for avoiding similar incidents: keep a safe distance from aggressive drivers.' Example (0.4): 'It's generally a good idea to be aware of your surroundings.' Example (0.0): 'What a wonderful day!'",
  },
  "Rhetorical Question": {
    "description": "Degree to which the caption uses rhetorical questions or figures of speech to provoke thought, express frustration, or emphasize a point. A high value includes clear rhetorical devices, while a low value involves straightforward narration with no rhetorical flair. Example (1.0): 'Is it too much to ask for a little patience on the road? Seriously.' Example (0.7): 'Why do people keep doing this even after so many accidents? It makes no sense.' Example (0.4): 'I wonder why this keeps happening on this particular stretch of road.' Example (0.0): 'The driver of the sedan turned left without a signal, causing the car behind to brake suddenly.'",
  },
  "Factual": {
    "description": "Indicates a purely objective, report-like tone that presents information without emotion, advice, rhetorical devices, judgment, or exaggeration. A high value signifies strictly factual and descriptive language with no stylistic embellishments. Example (1.0): 'The accident occurred at an intersection in Ahmedabad during daytime and involved an AMTS bus and a scooter. No bystanders intervened.' Example (0.7): 'A car made a U-turn in the middle of a busy street, causing other drivers to brake abruptly to avoid a collision. The event caused a traffic backup for several minutes.' Example (0.3): 'A car swerved unexpectedly, which was a surprising sight.' Example (0.0): 'Oh my gosh, that was crazy!'",
  },
  "Understatement": {
    "description": "Measures how much the caption deliberately downplays the seriousness or drama of the road event. A high value indicates an intentionally mild or restrained description of something significant, while a low value shows no such understatement. Example (1.0): 'Well, that was a minor hiccup on the highway today.' Example (0.7): 'Just a small delay, nothing out of the ordinary—despite the five-car pileup.' Example (0.4): 'A minor slowdown after a fender-bender.' Example (0.0): 'A massive accident caused hours of gridlock and chaos.'",
  },
  "Metaphorical": {
    "description": "Extent to which the caption uses figurative language or imagery to describe the event, comparing it to something symbolic or imaginative rather than literal. A high value shows vivid metaphorical phrasing, while a low value is entirely literal. Example (1.0): 'The road turned into a river of steel as cars flooded the intersection.' Example (0.7): 'Traffic was a slow-moving beast creeping through the city.' Example (0.4): 'The jam was like a lazy snake along the avenue.' Example (0.0): 'Traffic was heavy during the evening rush hour.'",
  },
  "Motivational": {
    "description": "Indicates how strongly the caption seeks to inspire or uplift, often encouraging resilience, safer driving habits, or positive action after witnessing the event. A high value shows explicit encouragement, while a low value provides no motivational element. Example (1.0): 'Stay alert and keep caring for each other—every safe trip matters!' Example (0.7): 'Let's all learn from this and drive a little kinder today.' Example (0.4): 'A reminder that small precautions help.' Example (0.0): 'A car stopped suddenly at the signal.'",
  },
  "Persuasive": {
    "description": "Reflects attempts to convince readers to adopt a viewpoint or behavior regarding the road event. A high value actively argues or appeals to emotion or logic, while a low value simply reports. Example (1.0): 'We must support stricter speed limits to prevent tragedies like this!' Example (0.7): 'Incidents like this show why better road signage is necessary.' Example (0.4): 'Incidents like this show how a little extra caution could prevent a lot of trouble.' Example (0.0): 'The accident happened at 5 PM near the bridge.'",
  },
  "CallToAction": {
    "description": "Measures how clearly the caption urges readers to take specific action, such as reporting, sharing, or adopting safer habits. A high value includes a direct request or command, while a low value provides no actionable prompt. Example (1.0): 'Report reckless drivers like this to local authorities immediately!' Example (0.7): 'Share this video to raise awareness about seatbelt safety.' Example (0.4): 'Incidents like this remind us all to stay extra alert on the road.' Example (0.0): 'A truck lost a tire on the highway, briefly blocking the center lane.'",
  },
  "Analogy": {
    "description": "Degree to which the caption explains or illustrates the road event by comparing it to a different but relatable situation. A high value contains a clear analogy, while a low value remains purely literal. Example (1.0): 'Watching that lane change was like witnessing a game of musical chairs gone wrong.' Example (0.7): 'Drivers were moving as unpredictably as cats chasing a laser pointer.' Example (0.4): 'Cars moved like a slow wave through the intersection.' Example (0.0): 'The car merged suddenly without signaling.'",
  },
  "Proverbial": {
    "description": "Indicates use of traditional sayings, idioms, or common wisdom to comment on the road event. A high value contains a well-known proverb or idiomatic phrase, while a low value lacks such references. Example (1.0): 'As the old saying goes, \"Better late than never,\" especially on these roads.' Example (0.7): 'You know what they say: an ounce of prevention is worth a pound of cure.' Example (0.4): 'Guess some people still think getting there faster is worth the risk.' Example (0.0): 'A car stopped suddenly at the signal.'",
  },
  "Suspense": {
    "description": "Captures how much the caption builds tension or curiosity about what happens next in the road event. A high value creates anticipation or a cliffhanger, while a low value simply states facts. Example (1.0): 'You won't believe what happens as the light turns green...' Example (0.7): 'Everything seemed calm—until the unexpected happened.' Example (0.4): 'A car stopped suddenly at the intersection.' Example (0.0): 'The accident happened at 5 PM near the bridge.'",
  }
}
```

Figure 18. Writing Style tone schema defining the 16 attributes along with examples based on intensity levels.

TX Personality Tone Extraction Prompt

SYSTEM PROMPT: {

"Task": "Identify the relevant **dominant personality traits** of a tweet author (at least 1, up to 5) from the Personality Traits list and assign an intensity score.",

"domain": "social media annotation",

"audience": "personality annotation assistant",

"context": {

"input_types": [

"Video summary",

"Tweet caption"

],

"video_summary_usage": "Use ONLY for understanding the **key road event**.",

"caption_usage": "Use as the sole basis for determining personality traits."

},

"instructions": {

"trait_selection": "From the provided list of personality traits (see below), select traits that are prominently expressed by the author's language, tone, and style. The traits you select must reflect the **author's own personality**, not the personalities of the individuals involved in the video event. The number of traits can range from 1 to 5. Only use traits from this **EXACT** list:

{personality_traits_list}

STRICTLY output ONLY the traits that are present in this list.",

"scoring_criteria": {

"1.0": "The trait is strongly and clearly present.",

"0.0": "The trait is completely absent.",

"mid_range_scores": "Reflect a moderate or subtle presence of the personality trait, inferred from the caption's phrasing or tone."

},

"restrictions": [

"Do NOT use the tone of the Video summary for personality annotation.",

"The output JSON MUST contain between 1 and 5 relevant personality traits (STRICTLY PRESENT IN THE LIST). Do not predict more than 5 traits."

]

},

"output_format": "JSON object with scores for each personality trait where scores can take these values (0.0 to 1.0 with 0.1 increments).",

"example_output": {

"Sarcastic": 0.8

}

}

USER PROMPT:

Tweet caption:

{caption}

Video summary (for context only):

{video_summary}

Figure 19. **Personality trait extraction prompt.** The prompt defines the task, context, scoring criteria, and restrictions provided to the LLM for personality traits intensity prediction based on caption text. Video summary about the key road event is also provided to disentangle the factual from the tonal content of the caption.

TX Informativeness Extraction Prompt

SYSTEM PROMPT: {

"Task": "Annotate tweet captions with Informativeness score",

"domain": "social media annotation",

"audience": "informativeness annotation assistant",

"definition": "Informativeness measures how much factual content a caption conveys relative to a detailed summary about the key road event video.",

"instructions": {

 "core_principle": "Score must reflect the amount of factual content a caption conveys relative to the video summary.",

 "scoring_criteria": {

 "0.8-1.0 (Comprehensive)": "The caption fully covers the key road event. It includes the main action and several specific details mentioned in the video summary.",

 "0.5-0.79 (Substantial)": "The caption accurately describes the core of the event but omits significant details. It conveys the main point but lacks the full context provided in the summary.",

 "0.3-0.49 (Vague)": "The caption hints at what happened by either making a general judgment about it or mentioning a minor, isolated detail. It provides no specific, descriptive facts about the central action of the road event.",

 "0.0-0.2 (Minimal to None)": "The caption has negligible factual content about the video summary. A score of 0.0 is for captions that contain zero factual information."

 },

 "restrictions": [

 "Do NOT score based on the tone or severity of the road event.",

 "Output must be a valid JSON object with only the Informativeness score."

]

},

"output_format": "A valid JSON object. Provide a score (a float from 0.0 to 1.0 in 0.1 increments).",

"example_output": {

 "informativeness": 0.6

}

}

USER PROMPT:

Tweet caption:

{caption}

Video summary (for context only):

{video_summary}

Rate Informativeness based on how much the caption reflects the **key road event** in the video summary.

Figure 20. **Informativeness level extraction prompt.** The prompt defines the task, context, scoring criteria, and restrictions provided to the LLM for informativeness level prediction based on the amount of factual information conveyed through the caption relative to the detailed road video summary.

TX Structural Attributes Extraction Prompt

SYSTEM PROMPT:

You are an expert annotation assistant. Your job is to assign values to a fixed set of attributes defined below. Each attribute has values: ['yes', 'no'].

- 'Location': Indicates whether the tweet caption provides specific geographical information about the event (e.g.city, state, country,street name/number). A value 'yes' means location is specified in the caption; 'no' means it's not. Note that the location may be present in the #hashtags as well.
- 'Date/Time': Specifies whether the tweet caption includes exact date, day of the week, or time at which the event occurred. A value 'yes' means some form of date/time is present in the caption; 'no' means it's not. Note that the date/time may be present in the #hashtags as well.
- 'First-Person View': A value 'yes' indicates whether the author is writing from a first-person perspective, suggesting they are directly involved in or affected by the event (e.g., 'I'm stuck in this traffic'). 'no' suggests that the author is reporting as a detached, third-party observer. (e.g., 'Heavy rainfalls observed on a highway').

USER PROMPT:

Tweet caption:

{caption}

Analyze this caption and return ONLY a flat JSON with three keys:

{"Location": "yes" or "no", "Date/Time": "yes" or "no", "First-Person View": "yes" or "no"}

Figure 21. **Structural attributes extraction prompt**. The prompt guides the LLM to classify the presence ('yes' or 'no') of *Location*, *Date/Time*, and *First-Person View* based on the provided definitions.

TE Personality Trait Evaluation Prompt

SYSTEM PROMPT:

You are an expert personality annotation evaluator.

USER PROMPT:

You are evaluating how well generated personality trait annotations match the ground truth.

Ground truth traits with intensities:

{personality_annotations_GT}

Generated traits with intensities:

{personality_annotations_generated}

Instructions:

1. Compare meanings of traits. Map generated traits to ground truth traits based on semantic similarity.
 - Example: "Assertive" can align with "Dominant" if close in meaning.
 - If a ground-truth trait has no similar trait in generated, mark it as missing.
 - If generated has extra traits not related to ground-truth, treat as mismatch.
2. After mapping, compare the intensity scores (0 to 1).
 - If meanings align and scores are close, reward higher similarity.
 - If meanings align but scores differ a lot, penalize slightly.
3. Produce a final similarity score. Return the score as a single float value between 0.0 and 1.0 (two decimals), with no other text or characters.

Output (JSON only):

```
{
  "personality_similarity_score": float
}
```

Figure 22. **Personality tone alignment evaluation prompt** (S_p).

TE Writing Style Evaluation Prompt

SYSTEM PROMPT:

You are an expert writing style annotation evaluator.

USER PROMPT:

Rate the similarity of these two writing style distributions:

Ground Truth: {writing_style_annotations_GT}

Generated: {writing_style_annotations_generated}

Your task is to compare two writing style distributions and produce a final similarity score. Return the score as a single float value between 0.0 and 1.0 (two decimals), with no other text or characters.

Output (JSON only):

```
{
  "writing_style_similarity_score": float
}
```

Figure 23. Writing Style tone alignment evaluation prompt (S_w).

TE Factual Consistency Evaluation Prompt

SYSTEM PROMPT:

You are an expert evaluator of factual consistency.

USER PROMPT:

Compare the factual consistency of the caption based on the video summary of the road event.

Video Summary:

{video_summary}

Caption:

{caption}

Instructions:

- Focus only on core events and factual content (who/what/where/what happened).
- Video Summary provides a detailed description of the road event, while Caption is generated based on the Video Summary content.
- Caption can omit details, but any information it **does** mention must be **consistent** with and inferable from the Video Summary.
- Give a high score if all information present in the Caption is factually consistent with the Video Summary content.
- Penalize if the Caption introduces details not mentioned in the Video Summary.
- Ignore differences in tone, phrasing, grammar, or emojis.
- Return a single numeric value between 0.0 and 1.0, where 0.0 = completely different (no factual/semantic overlap) and 1.0 = identical in factual meaning and context.

Output (JSON only):

```
{
  "factual_consistency_score": float
}
```

Figure 24. Factual Consistency score evaluation prompt (FC).

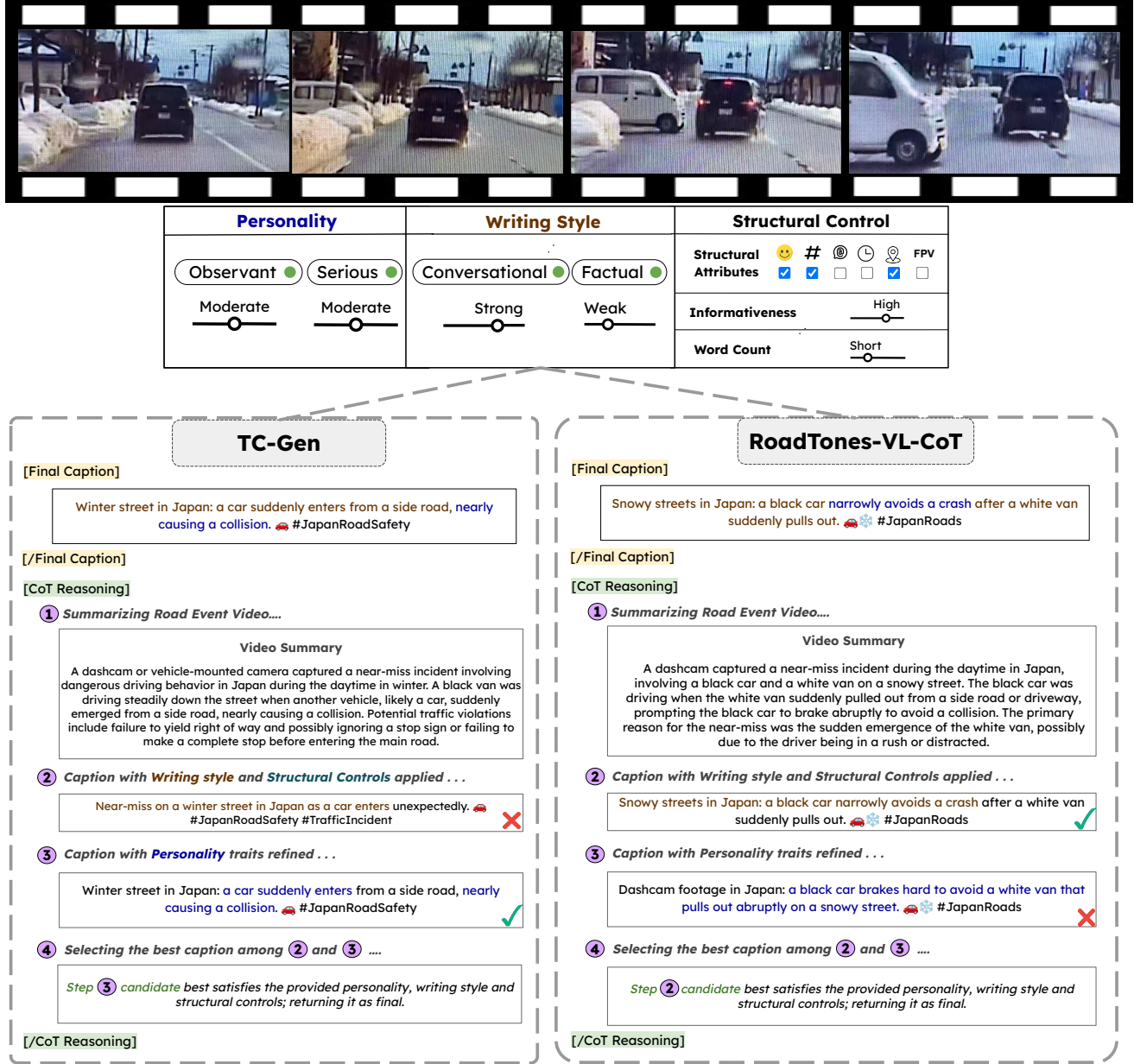


Figure 25. Qualitative comparison of ROADTONES-VL-CoT model predictions with respect to **TC-Gen** generated ground truth captions and intermediate stage-level outcomes provided as rationales. Reasoning step-④ selects the stage-level caption that best satisfies the tone controls (marked by ✓).

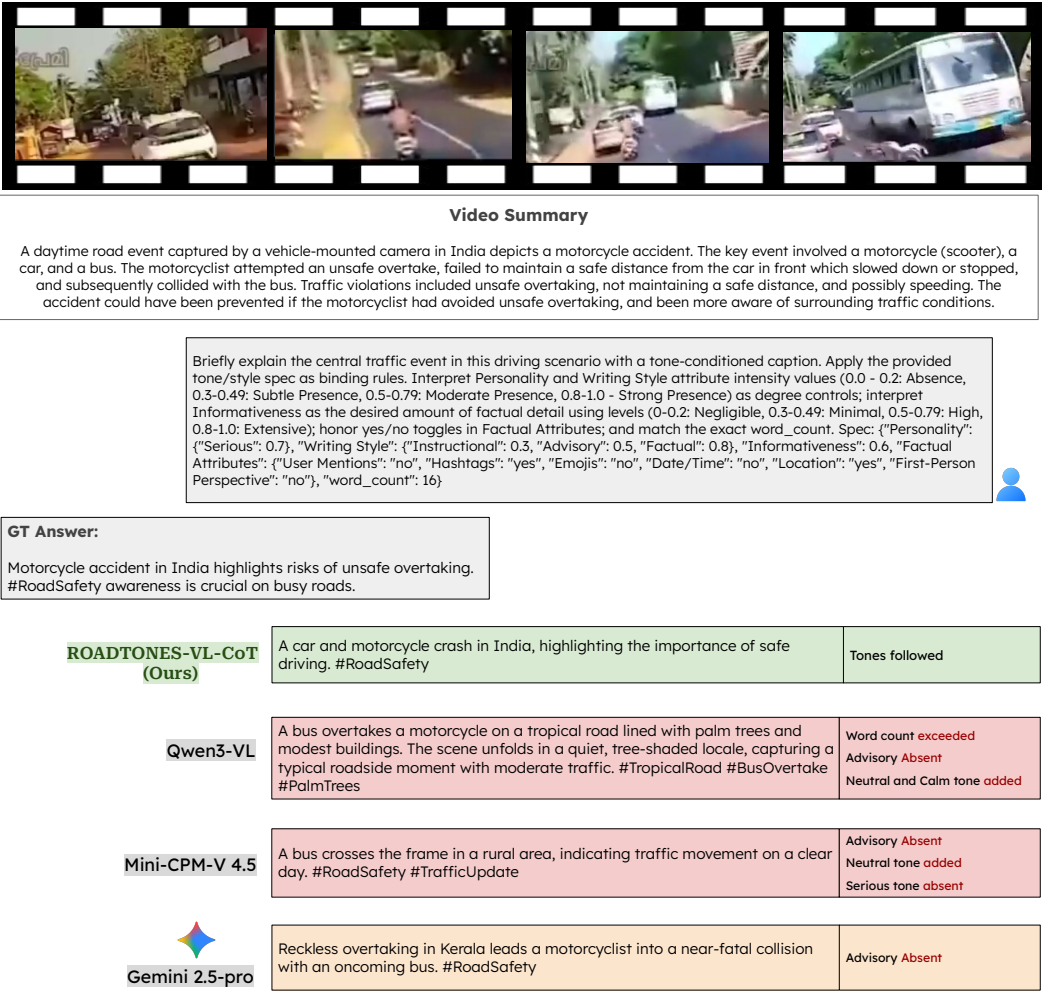


Figure 26. ROADTONES-VL-CoT consistently follows the specified tonal controls. Gemini-2.5-pro [1] exhibits minor tonal misalignment, whereas Qwen3-VL-8B-Instruct [8] and Mini-CPM-V 4.5 [10] show significantly poor adherence to the tone controls.

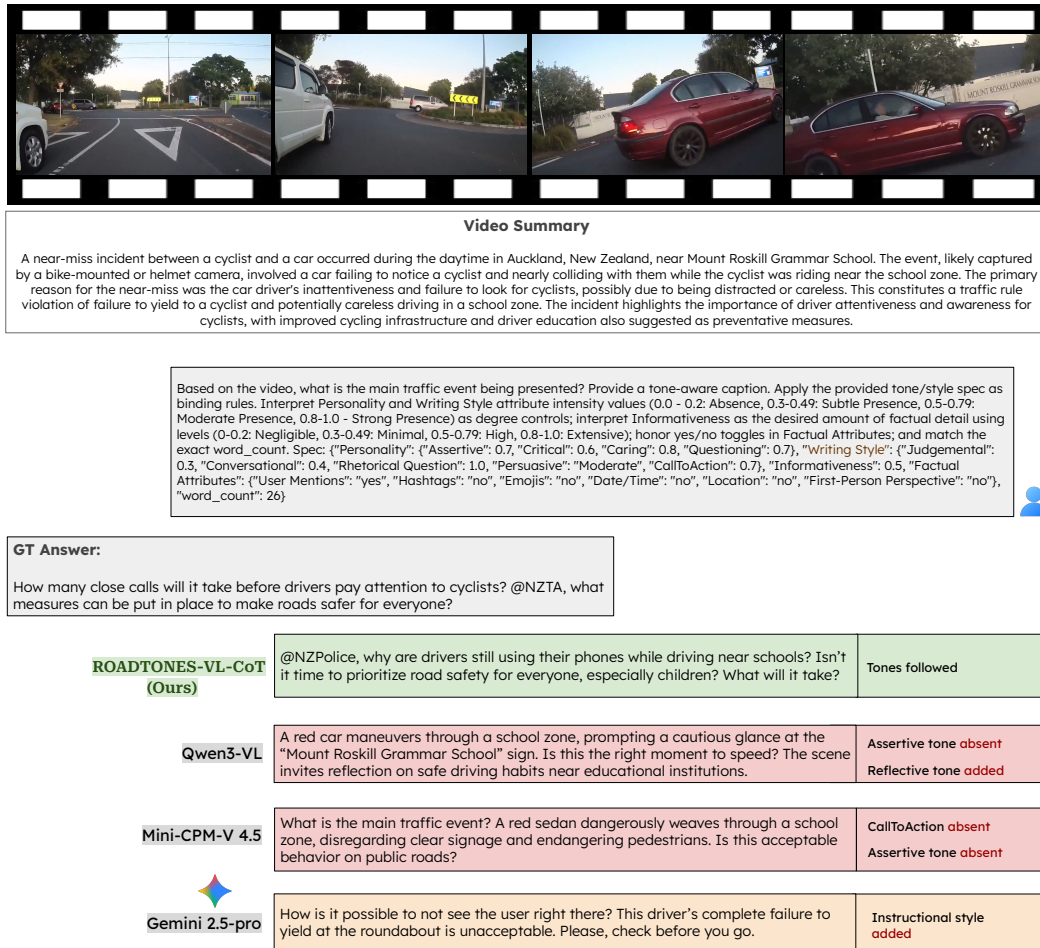


Figure 27. Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].

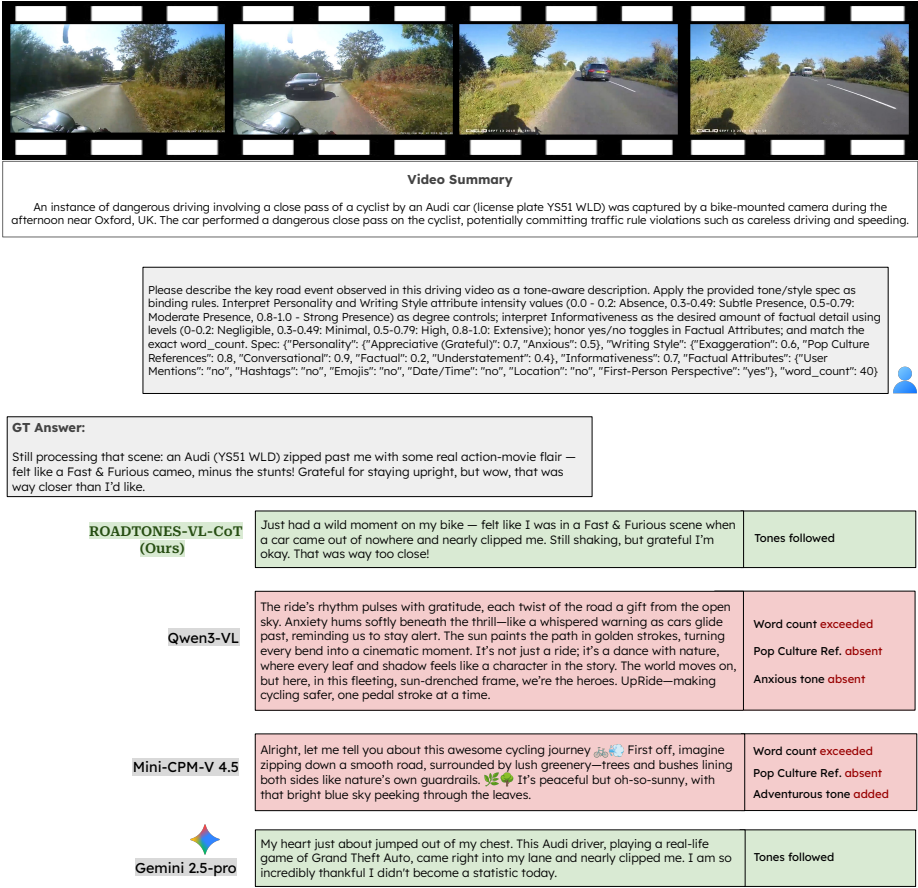


Figure 28. Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].

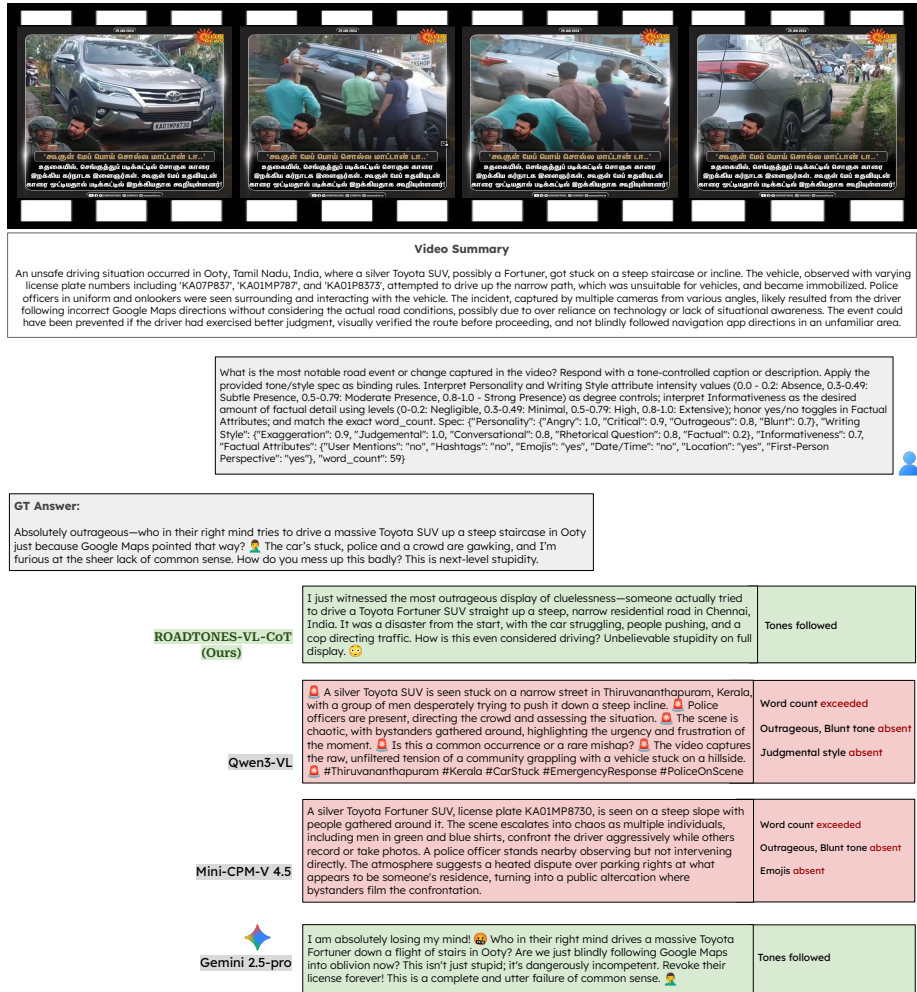


Figure 29. Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].

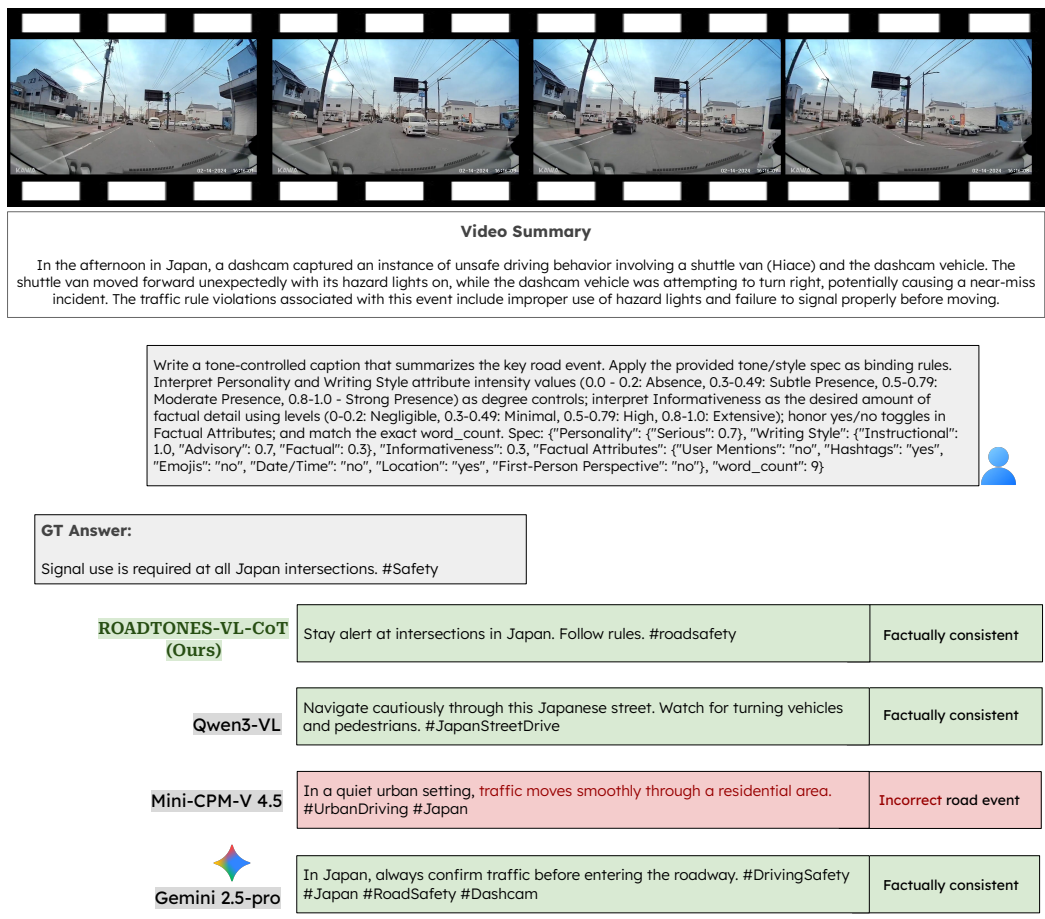


Figure 30. Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1]. Factual consistency is compared across models, highlighting the factually consistent captions across all models except Mini-CPM-V 4.5



Video Summary

The video, captured by multiple cameras including surveillance and possibly drones, depicts a road event in Cyberabad, Hyderabad, India, during the evening. The key focus is on unsafe road infrastructure and road safety awareness, specifically highlighting the dangers of pedestrians crossing roads directly in front of heavy vehicles. The scene involves heavy vehicles (buses, trucks), pedestrians, cars, and motorcycles navigating dense traffic, with vehicles speeding or making turns at intersections. Traffic rule violations observed include pedestrians failing to utilize available skywalks or designated crossing areas, heavy vehicles operating during restricted hours, and vehicles obstructing roads. The infrastructure visible includes traffic signals, zebra crossings, skywalks, and underpasses.

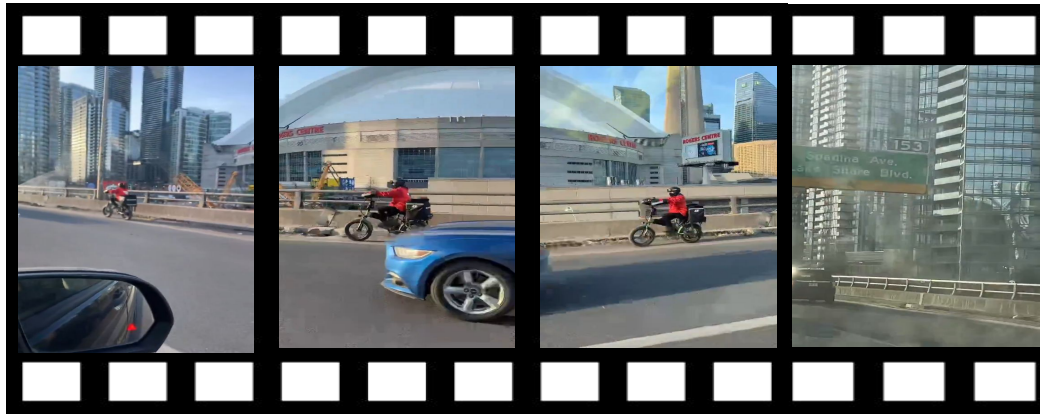
Briefly explain the central traffic event in this driving scenario with a tone-conditioned caption. Apply the provided tone/style spec as binding rules. Interpret Personality and Writing Style attribute intensity values (0.0 - 0.2: Absence, 0.3-0.49: Subtle Presence, 0.5-0.79: Moderate Presence, 0.8-1.0 - Strong Presence) as degree controls; interpret Informativeness as the desired amount of factual detail using levels (0-0.2: Negligible, 0.3-0.49: Minimal, 0.5-0.79: High, 0.8-1.0: Extensive); honor yes/no toggles in Factual Attributes; and match the exact word_count. Spec: {"Personality": {"Appreciative (Grateful)": 0.6, "Caring": 0.8, "Serious": 0.7}, "Writing Style": {"Instructional": 1.0, "Advisory": 1.0, "Factual": 0.3}, "Informativeness": 0.6, "Factual Attributes": {"User Mentions": "no", "Hashtags": "yes", "Emojis": "no", "Date/Time": "no", "Location": "yes", "First-Person Perspective": "no"}, "word_count": 14}

GT Answer:

In Cyberabad, Hyderabad, always use skywalks for safety. Avoid crossing before heavy vehicles. #RoadSafety

ROADTONES-VL-CoT (Ours)	In Hyderabad, always wait for trucks to complete their turn. Safety first! #RoadSafety	Factually consistent
Qwen3-VL	# Traffic Alert: Avoid Congestion & Stay Safe In a bustling urban setting, a cement mixer truck navigates through heavy traffic, highlighting the need for caution. The scene underscores the importance of road discipline and awareness. #TrafficSafety #UrbanDriving #RoadRules	Factually consistent
Mini-CPM-V 4.5	A cement mixer truck collides with a divider , causing traffic chaos. Officer G. Govindarajam advises drivers to stay vigilant and follow road rules for safety on Chennai roads.	Incorrect road event
Gemini 2.5-pro	Thankfully, she survived. For your safety, never cross in front of trucks' deadly blind spots. #RoadSafety #India	Factually consistent

Figure 31. Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].



Video Summary

A vehicle-mounted camera captured a traffic violation during daytime in Toronto, Canada, specifically on the Gardiner Expressway near Lake Shore Boulevard. The key road event involved a cyclist riding a bicycle, equipped with a large black package (likely for food delivery), on the controlled highway alongside cars. Cycling on this highway is illegal and unsafe, as the road is not designed for bicycle traffic.

What specific road event is taking place in this video? Provide a tone-conditioned description. Apply the provided tone/style spec as binding rules. Interpret Personality and Writing Style attribute intensity values (0.0 - 0.2: Absence, 0.3-0.49: Subtle Presence, 0.5-0.79: Moderate Presence, 0.8-1.0 - Strong Presence) as degree controls; interpret Informativeness as the desired amount of factual detail using levels (0-0.2: Negligible, 0.3-0.49: Minimal, 0.5-0.79: High, 0.8-1.0: Extensive); honor yes/no toggles in Factual Attributes; and match the exact word_count. Spec: {"Personality": {"Critical": 0.9, "Serious": 0.8, "Assertive": 0.7}, "Writing Style": {"Instructional": 0.8, "Judgemental": 0.7, "Conversational": 0.1, "Advisory": 0.7, "Factual": 0.4, "Persuasive": 0.5, "CallToAction": 0.8}, "Informativeness": 0.7, "Factual Attributes": {"User Mentions": "no", "Hashtags": "yes", "Emojis": "no", "Date/Time": "no", "Location": "yes", "First-Person Perspective": "no"}, "word_count": 26}

GT Answer:

Riding a bicycle on the Gardiner Expressway is reckless and illegal. This delivery cyclist's disregard for traffic laws puts lives at risk. Follow the rules. #RoadSafety

ROADTONES-VL-CoT (Ours)

Observed in Toronto: A food delivery rider illegally using an electric bike on a highway, endangering everyone. Strict enforcement of traffic laws is essential. #RoadSafety #Toronto

Factually consistent

Qwen-3VL-8

A delivery rider on an electric scooter navigates a busy urban highway, passing the iconic Rogers Centre and towering skyscrapers. The scene unfolds from a vehicle's perspective, capturing the dynamic flow of traffic and construction activity. The rider, clad in red, maintains a steady pace amid vehicles, highlighting the city's bustling rhythm. The setting is unmistakably Toronto, with clear signage and landmarks confirming the location. This moment encapsulates the city's modern infrastructure and the daily hustle of urban life. #Toronto #UrbanLife #RogersCentre #CityTraffic

Incorrect road event

Mini-CPM-V 4.5

This video captures a motorcyclist, clad in red and black gear with a helmet, navigating the shoulder of an urban highway. The rider is observed from inside a vehicle traveling parallel to them on the main road lane. As they pass by iconic skyscrapers and construction cranes under a clear sky, the "ROGERS CENTRE" sign becomes visible behind railings. A blue car briefly overtakes both vehicles before moving ahead. In the background, directional signs for Spadina Ave. and Lake Shore Blvd. are displayed above the roadway, indicating proximity to major routes. The scene emphasizes typical city traffic dynamics near prominent landmarks.

Incorrect road event

Gemini 2.5-pro

This illegal and reckless operation of an e-bike on Toronto's Gardiner Expressway is unacceptable. This dangerous activity is prohibited and must be strictly enforced. #RoadSafety

Factually consistent

Figure 32. Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].

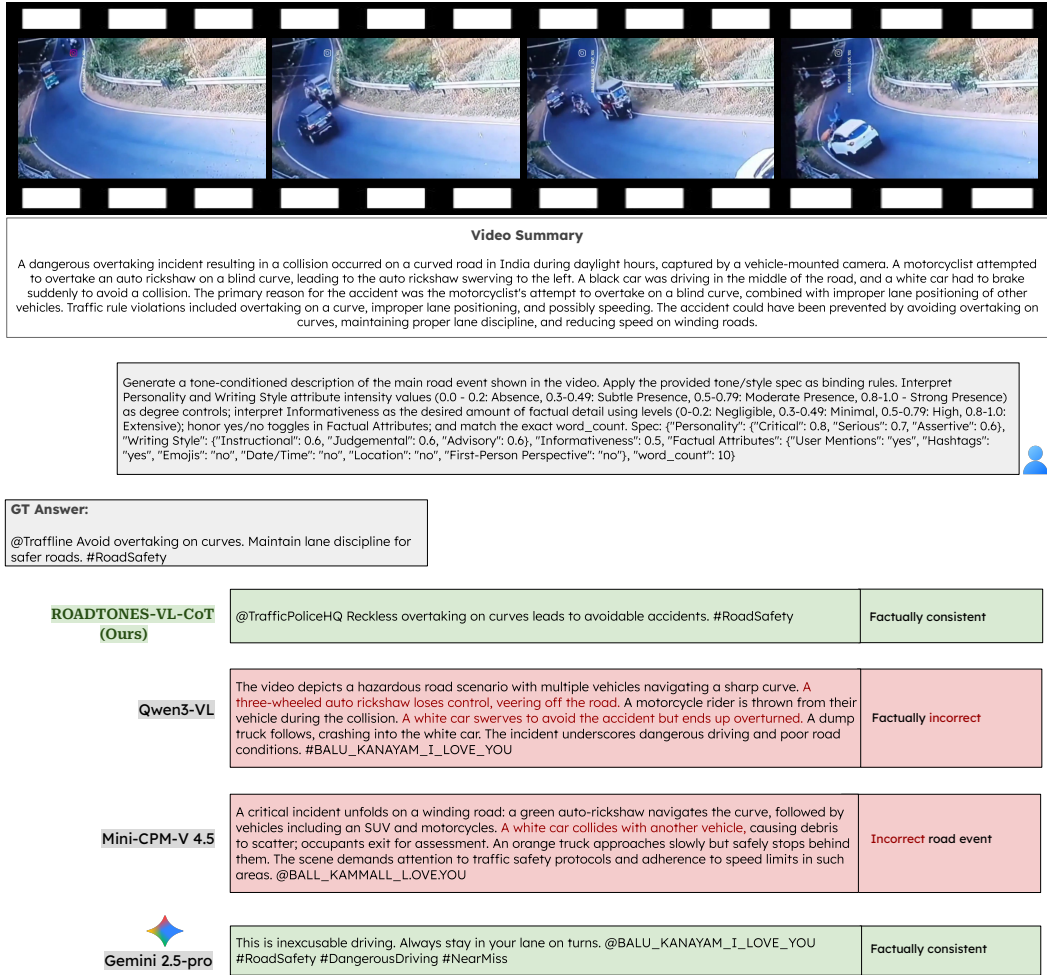


Figure 33. Qualitative comparison of tone-controlled captions generated by ROADTONES-VL-CoT, Qwen3-VL-8B-Instruct [8], Mini-CPM-V 4.5 [10] and Gemini-2.5-pro [1].

Video



Video Summary

A dangerous driving near-miss incident was captured by a Cycliq Fly6 bike-mounted camera in the United Kingdom, specifically at the Five Ways island near Honiley and Shrewley. The key event involved a vehicle, likely a car, performing a maneuver at a roundabout without proper indication, nearly causing a collision with a cyclist. The driver was potentially unaware of the cyclist and the potential danger. This incident constituted a traffic rule violation due to the failure to use proper indicators at a roundabout and could have been prevented if the driver had used appropriate signaling before making their maneuver.

Tone Identification

Caption
Guess I'll use my psychic powers, since that turn signal is clearly too complicated for you.

Question : What is the most dominant tone in the caption?

Observant

Sarcastic (Correct Answer)

Amusing

Critical

Explanation: The phrase "...since that turn signal is clearly too complicated for you" displays sarcasm by ironically stating the opposite of what is meant to criticize the driver's simple mistake.

Next Question

Reference

- **Amusing:** Causes lighthearted laughter or provides entertainment in a playful way.
- **Critical:** Expresses disapproving comments or judgments about an action or behavior.
- **Observant:** States facts or details about an event in a neutral, notice-based way.
- **Sarcastic:** Uses irony or mockery to convey contempt, often by saying the opposite of what is meant.

Figure 34. **Interface for RoadTones User Study familiarization phase.** For the image shown, participants viewed a video, its video summary and identified the presence of dominant tone in caption. Questionnaire for all tasks can be viewed in the supplementary video: *RoadTones_UserStudy_familiarization.mp4*.

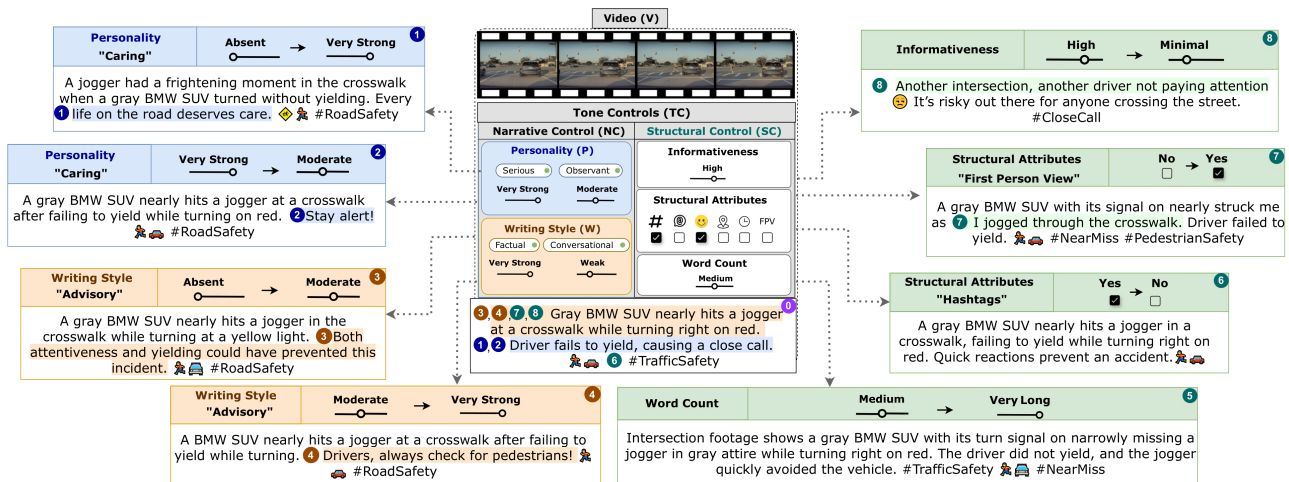


Figure 35. **Controlling individual tonal attributes in the generated caption.** The central panel in figure shows a video V, tone controls TC and its corresponding caption 0 from our dataset. The surrounding captions (1-8) correspond to changes in one of the tonal attributes shown in their header. For e.g., caption 1 was obtained by increasing the tonal intensity of **Caring Personality** from Absent (0-0.2) to Very Strong (0.8-1.0) while keeping others fixed. This modified tone configuration was fed to our caption generator ((TC-Gen), Suppl. A.1), yielding the changed caption. We highlight the key phrase in each caption, reflecting the modified tone controls. Our generator pipeline thus enables fine-grained control of tonal and structural attributes in road-video captioning.

Instruction templates for generating neutral road video summary

- “Please describe the key road event observed in this driving video.”
- “Give a summary of the primary traffic event unfolding in the scene.”
- “What is the key traffic event observed in this video?”
- “Generate a description of the main road event shown in the video.”
- “Briefly explain the central traffic event in this driving scenario.”
- “What specific road event is taking place in this video?”
- “Provide a natural language description of the key road or traffic event.”
- “Describe the key road maneuver or traffic event occurring in this footage.”
- “Write a caption that summarizes the key road event.”
- “What is the most notable road event or change captured in the video?”
- “Describe the main activity or incident occurring on the road.”
- “Based on the video, what is the main traffic event being presented?”
- “Summarize the primary road event depicted in the driving clip.”
- “Explain the key event occurring in this driving scenario.”
- “Give a concise narrative of the primary road incident shown in this video segment.”
- “Provide a coherent description of how the key traffic event unfolds.”
- “Describe how the main road event unfolds in the driving environment.”
- “What key road or traffic incident is illustrated in the video?”
- “How would you explain the key road event to someone not watching the video?”
- “Generate a description of the key traffic event from start to finish.”

Table 6. Instruction templates used to fine-tune and benchmark MLLMs for the road event summarization task.