

Beamr Optimized compression for Cosmos Curate AV pipeline

Achieving 41-57% reduction while preserving model fidelity

Authors: [Ronen Nissim](#), [Boris Fillipov](#), [Ethan Fenakel](#) — AI Team, [BEAMR](#), 23 Feb 2026

1. Executive Summary

We measured embedding distances and perceptual quality within the Cosmos Curate AV pipeline across 9 uncompressed source files from several well-known datasets, analyzing the impact of NVENC with CABR optimization compared to default NVENC compression.

CABR optimization reduced video bitrates by 41–57% while introducing no measurable impact on the captioning pipeline beyond the model's inherent stochastic variability. This invariance was validated through two complementary embedding models (T5-XXL and SBERT) and confirmed via cosine similarity analysis, t-SNE visualization, and Visual Realism Index (VRI) assessment. In several cases, CABR-optimized files produced embedding similarity scores closer to the source than default NVENC — indicating the pipeline is operating below the sensitivity threshold needed to distinguish between the two compression approaches."

VRI field agreement rates for CABR-compressed variants remained between 88.9% and 100% across all tested videos, with most videos achieving 93–98% agreement — comparable to the source-to-default compression baseline and well within expected classification noise.

2. Goal

As part of BEAMR's ongoing efforts to ensure ML-safe compression, the proposed analysis focuses on evaluating the sensitivity of Cosmos Curate's textual descriptions to video compression, and verify their invariance to reduced bitrate, CABR-optimized compression. This evaluation is based on analysis of embedding distances and model perceptual quality metrics.

3. Methodology

3.1 Analysis Rational

The Cosmos Curate captioning model is built on a large multimodal LLM that samples input video at a low frame rate, encodes each frame into visual tokens, combines them with a textual instruction prompt, and produces a natural-language description of the video sequence given the provided prompt context. Since the decoded pixel values of input frames are the model's only visual input, any effect compression has will propagate through the entire pipeline to manifest in the output text.

Because the model is highly complex, it is not possible to identify a single internal layer or control point whose sensitivity to input perturbations would reliably correlate with contextual variation in the final text output. Consequently, the only faithful way to evaluate the effect of compression on model behavior is to evaluate the pipeline end-to-end.

However, quantifying textual contextual similarity is not straightforward. We address this through four complementary approaches:

a. **Embedding-based similarity** — Encode captions using established text-embedding models that map semantically similar texts to nearby points in vector space, where contextual differences are captured by cosine distance. Multiple embedding models (T5-XXL, SBERT) are used to confirm that findings are not artifacts of a single embedding architecture/model.

b. **Stochastic noise-floor characterization** — Establish the model's inherent output variability by performing repeated non-deterministic inference runs on uncompressed source videos. This stochastic distribution defines the noise floor, or underlying noise level, against which any compression-induced difference must be compared.

c. **Intra- vs. inter-video cluster separation** — Evaluate whether per-video embedding clusters are sufficiently tight and mutually separated to make this analysis method informative. If stochastic embeddings for a single video form a compact cluster well-separated from other videos' clusters, the method has discriminative power. If clusters overlap to the extent that a stochastic sample cannot be attributed to its source video, the evaluation method itself would lack the resolution to detect compression effects.

d. **Structured classification agreement (VRI)** — Complement the free-text embedding analysis with a field-level categorical assessment. The Visual Realism Index (VRI) independently classifies 45 scene attributes (infrastructure, traffic, fine details) from the video, providing a structured cross-check that can reveal localized semantic disagreements — such as a misclassified traffic light state or pedestrian count — that aggregate embedding similarity might mask.

The core hypothesis is: **if compression-induced embedding shifts fall within the stochastic noise floor of the uncompressed source, and that noise floor is narrow enough to distinctly differentiate individual videos in embedding space, then the model is insensitive to that level of compression — and the method has sufficient resolution or accuracy to make that determination.** By following these guidelines, even for a limited test corpus, we can determine whether this sensitivity analysis method is viable and establish an initial assessment of the model's sensitivity to compression.

3.2 Pipeline Overview

The Cosmos Curate AV pipeline processes a subsampled video sequence through a vision-language model (Qwen2.5-VL-7B via vLLM) along with an instructive prompt, generating a natural-language text caption per video clip. The caption is then encoded into a high-dimensional embedding via T5-XXL for downstream use in video-text alignment, retrieval, and curation tasks.

The pipeline stages are:

1. **QwenInputPreparationStage** — Decode video, sample frames at 2.0 fps, resize to Qwen-compatible resolution (multiples of 28px within a pixel budget), normalize with CLIP mean/std, and construct chat-template LLM inputs.
2. **QwenCaptionStage** — Generate captions with Qwen2.5-VL-7B (vLLM). The "default" prompt uses a two-stage self-refinement loop.
3. **T5Stage** — Encode the refined caption with T5-XXL (encoder-only, 1024-dim output).
4. **EnhanceCaptionStage** — (Optional) Classify VRI attributes using ChatLM.

3.3 Test Corpus

A set of 9 uncompressed AV videos were selected from the following sources: [A2D2](#), [KITTI](#), [PANDaset](#), and [BEAMR](#)-proprietary footage.

Two additional reference videos — "[Big Buck Bunny](#)" and an all-black-frames sequence — were included to illustrate embedding behavior for genuinely distinct content.



Analysis dataset

Video	Source Dataset	Scene Type	Resolution	Duration	FPS
1aa17d25-3009-41cd-bb69-f3291d7754da	A2D2	Suburban, daytime, clear sky	1920×1080	5.0s	30
95c10498-2dfb-43ff-b469-3671dc5b8b07	A2D2	Urban intersection, cloudy day	1920×1080	5.0s	30
2011_09_26_drive_0015_extract_000	KITTI	Suburban road, clear day	1392×512	5.0s	25
99a66be6-8ee9-41c0-83e1-3e5353f90273	A2D2	Nighttime highway, streetlit	1920×1080	5.0s	30
70a21c76-78e4-481c-bad3-c0db5d48834d	A2D2	Nighttime urban intersection	1920×1080	5.0s	30
2025-04-29_15-49-00-front_000	BEAMR	Urban intersection, clear day	1280×960	5.0s	36
DJI_0001_015_003_5s	PANDA	Aerial / urban intersection	1280×720	5.0s	30
gai_cam_front_center	PANDA	Urban street, modern architecture	1920×1208	5.0s	30
mun_cam_front_center	PANDA	City street, daytime, parked cars	1920×1208	5.0s	30

While the sample size is limited, videos were selected to span diverse scene types, lighting conditions, and bitrate ranges.

3.4 Compression Configurations

For each uncompressed source video, two compressed variants were generated at QP=20, inline with Cosmos Curate default encoding configuration:

- **CABR_0**: NVENC ConstQP=20 baseline encoding - similar to the default NVENC encoding in Cosmos Curate.
- **CABR_1**: the corresponding CABR-optimized compression, applying content-adaptive bitrate optimization on top of the baseline NVENC encoding.

3.5 Captioning Protocol

Two captioning modes were used to separate deterministic behavior from stochastic variability:

Deterministic captioning — Caption generation for the source and both compressed variants using Qwen2.5-VL-7B (via vLLM) with the Cosmos Curate default prompt. Configuration: strictly deterministic greedy decoding (temperature=0, do_sample=False) with a fixed seed (seed=42), producing identical output across repeated runs.

Non-deterministic captioning — 19 or 20 repeated runs per source video with the same configuration to characterize the model's intrinsic caption variability. This distribution serves as the underlying noise level, or noise floor, against which compression-induced differences are measured.

3.6 Embedding Models

Two embedding models were used to ensure findings are robust and not artifacts of a single encoder:

- **T5-XXL (1024-dim)** — Mean-pooled encoder output capturing deep semantic structure. T5-XXL is the native embedding model within the Cosmos Curate pipeline, so its embeddings directly reflect how the production system represents video content for downstream tasks such as video-text alignment and retrieval.
- **SBERT (768-dim, all-mpnet-base-v2)** — Sentence-level embedding for semantic similarity. SBERT provides a complementary, model-independent baseline trained on large-scale semantic similarity data, validating that findings generalize beyond the T5 representation.

3.7 Evaluation Metrics

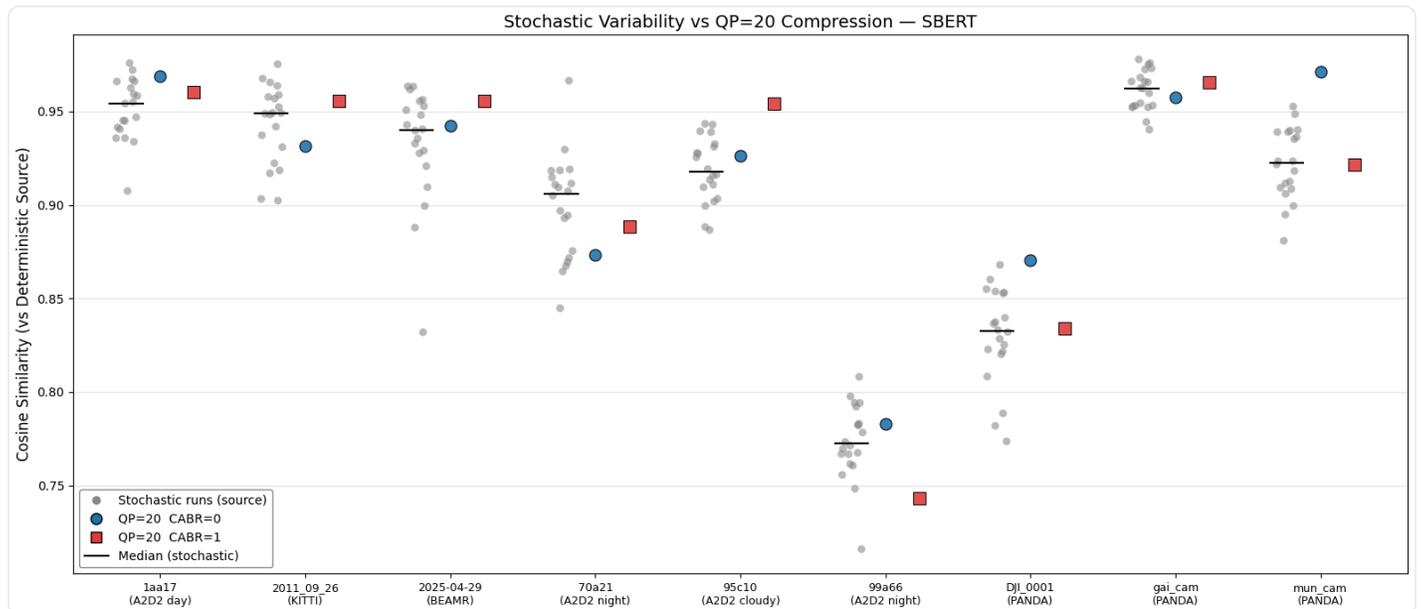
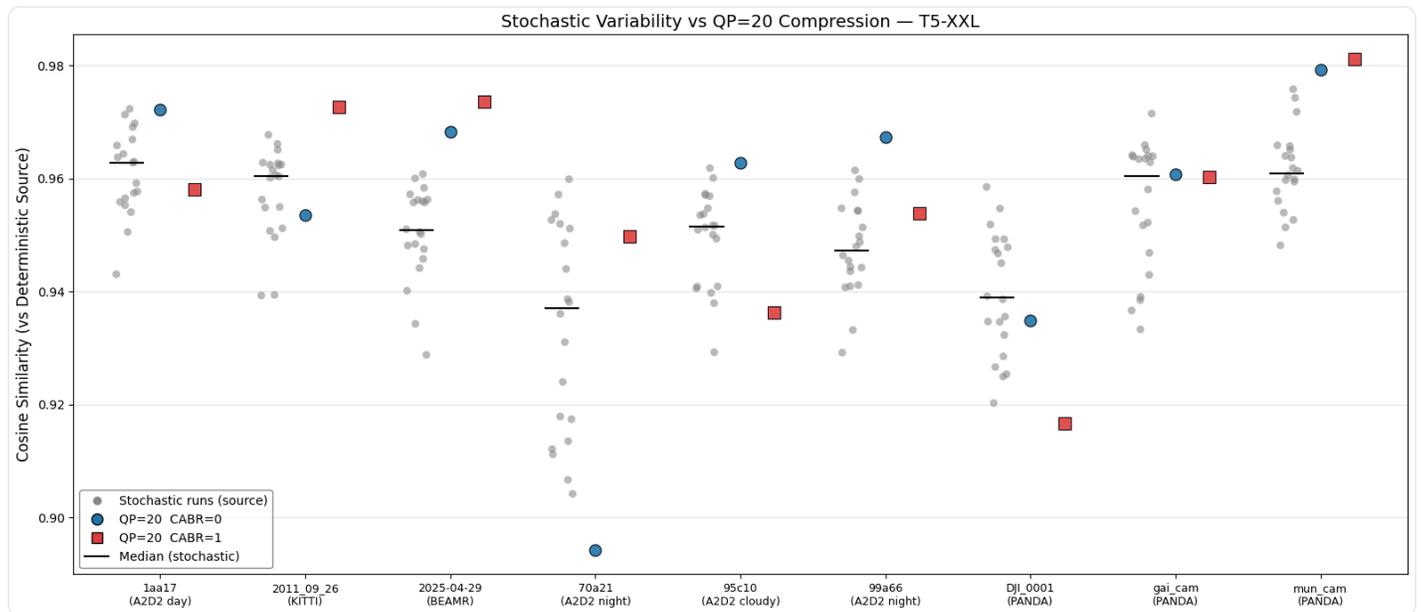
- **Cosine similarity** between the deterministic source caption embedding and each compressed variant's deterministic caption embedding, compared against the stochastic distribution.
- **t-SNE visualization** of high-dimensional embeddings to qualitatively assess clustering behavior across videos and compression variants.
- **Visual Realism Index (VRI)** — 45-field structured classification (15 static infrastructure, 15 dynamic traffic/hazard, 15 fine detail/count fields) comparing source and compressed outputs field-by-field.

4. Results

4.1 Stochastic Baseline Characterization

The model exhibits considerable run-to-run caption variability even with near-deterministic sampling parameters. This stochastic distribution defines the noise floor: any compression-induced difference smaller than this variability is indistinguishable from the model's inherent randomness.

The figures below show the cosine similarity distribution for the evaluated videos. for encoding both with and without CABR:



As will be further described ahead, for both default NVENC and optimized CABR-NVENC encodes the embeddings consistently fall within the stochastic range for each video across both evaluated embedding models. This suggests model invariance to both the default compression configuration and that CABR maintains equivalent contextual similarity (see sections 4.2 and 4.3 ahead)

The variation in median similarity across videos is partly attributable to where the deterministic source embedding falls within the stochastic distribution. Because the deterministic caption is produced with a single fixed seed, it represents one particular draw from the model's output space. When that draw happens to land near the centroid of the stochastic embedding cluster, all stochastic runs will exhibit high cosine similarity to it, yielding a high median. Conversely, when the deterministic embedding falls closer to the periphery of the cluster, measured similarities are lower and more dispersed — not because the model is less consistent for that video, but because the reference point is less central. This is a property of the chosen reference, not of the compression pipeline.

Below are the two embeddings stochastic similarity statistics tables, showing the cosine similarities to the deterministic source caption:

T5-XXL Stochastic Similarity Statistics

Video	Runs	T5 Mean	T5 Std	T5 Min	T5 Max
1aa17d25-3009-41cd-bb69-f3291d7754da	19	0.96	0.01	0.94	0.97
2011_09_26_drive_0015_extract_000	20	0.96	0.01	0.94	0.97
2025-04-29_15-49-00-front_000	20	0.95	0.01	0.93	0.96
70a21c76-78e4-481c-bad3-c0db5d48834d	20	0.93	0.02	0.9	0.96
95c10498-2dfb-43ff-b469-3671dc5b8b07	20	0.95	0.01	0.93	0.96
99a66be6-8ee9-41c0-83e1-3e5353f90273	20	0.95	0.01	0.93	0.96
DJI_0001_015_003_5s	20	0.94	0.01	0.92	0.96
gai_cam_front_center	20	0.96	0.01	0.93	0.97
mun_cam_front_center	20	0.96	0.01	0.95	0.98

SBERT Stochastic Similarity Statistics

Video	Runs	SBERT Mean	SBERT Std	SBERT Min	SBERT Max
1aa17d25-3009-41cd-bb69-f3291d7754da	19	0.95	0.02	0.91	0.98
2011_09_26_drive_0015_extract_000	20	0.94	0.02	0.9	0.98
2025-04-29_15-49-00-front_000	20	0.93	0.03	0.83	0.96
70a21c76-78e4-481c-bad3-c0db5d48834d	20	0.9	0.03	0.84	0.97
95c10498-2dfb-43ff-b469-3671dc5b8b07	20	0.92	0.02	0.89	0.94
99a66be6-8ee9-41c0-83e1-3e5353f90273	20	0.77	0.02	0.72	0.81
DJI_0001_015_003_5s	20	0.83	0.03	0.77	0.87
gai_cam_front_center	20	0.96	0.01	0.94	0.98

mun_cam_front_center	20	0.92	0.02	0.88	0.95
----------------------	----	------	------	------	------

Observations:

- T5-XXL stochastic means range from 0.933 to 0.961, with standard deviations of 0.007–0.018. The tightest distributions appear for well-lit daytime scenes (mun_cam, 1aa17), while the widest spread occurs for nighttime scenes (70a21, std=0.018).
- SBERT shows wider stochastic spread overall (std 0.010–0.031), particularly for nighttime video 99a66 (mean=0.774, std=0.020) and aerial footage DJI_0001 (mean=0.830, std=0.025). This reflects SBERT's higher sensitivity to caption wording variations in ambiguous scenes.

4.2 Cosine Similarity Analysis

The following tables summarize bitrate savings and cosine similarity to the deterministic source caption for both compression variants, alongside the stochastic baseline bounds for direct comparison.

T5-XXL Similarity Results

Video	BR Savings	CABR0 BR (*)	CABR1 BR (*)	CABR0 Sim	CABR1 Sim	Stoch. Range (min-max)
1aa17d25-3009-41cd-bb69-f3291d7754da	42.6%	25.93	14.88	0.97	0.96	0.9430-0.9723
2011_09_26_drive_0015_extract_000	42.4%	21.22	12.22	0.95	0.97	0.9393-0.9677
2025-04-29_15-49-00-front_000	41%	6.5	3.83	0.97	0.97	0.9288-0.9608
70a21c76-78e4-481c-bad3-c0db5d48834d	46%	4.85	2.62	0.89	0.95	0.9042-0.9599
95c10498-2dfb-43ff-b469-3671dc5b8b07	42.6%	25.54	14.66	0.96	0.94	0.9292-0.9618
99a66be6-8ee9-41c0-83e1-3e5353f90273	45.5%	7.84	4.27	0.97	0.95	0.9292-0.9614
DJI_0001_015_003_5s	54.2%	11.19	5.13	0.93	0.92	0.9202-0.9585
gai_cam_front_center	57%	40.85	17.57	0.96	0.96	0.9333-0.9715
mun_cam_front_center	50%	63.66	31.8	0.98	0.98	0.9482-0.9758

SBERT Similarity Results

Video	BR Savings	CABR0 BR (*)	CABR1 BR (*)	CABR0 Sim	CABR1 Sim	Stoch. Range (min-max)
1aa17d25-3009-41cd-bb69-f3291d7754da	42.6%	25.93	14.88	0.97	0.96	0.9074-0.9758
2011_09_26_drive_0015_extract_000	42.4%	21.22	12.22	0.93	0.96	0.9023-0.9752
2025-04-29_15-49-00-front_000	41%	6.5	3.83	0.94	0.96	0.8320-0.9634
70a21c76-78e4-481c-bad3-c0db5d48834d	46%	4.85	2.62	0.87	0.89	0.8448-0.9664
95c10498-2dfb-43ff-b469-3671dc5b8b07	42.6%	25.54	14.66	0.93	0.95	0.8866-0.9434
99a66be6-8ee9-41c0-83e1-3e5353f90273	45.5%	7.84	4.27	0.78	0.74	0.7160-0.8081
DJI_0001_015_003_5s	54.2%	11.19	5.13	0.87	0.83	0.7736-0.8679
gai_cam_front_center	57%	40.85	17.57	0.96	0.97	0.9402-0.9778

(*) - BR = Bitrate, units: Mbps

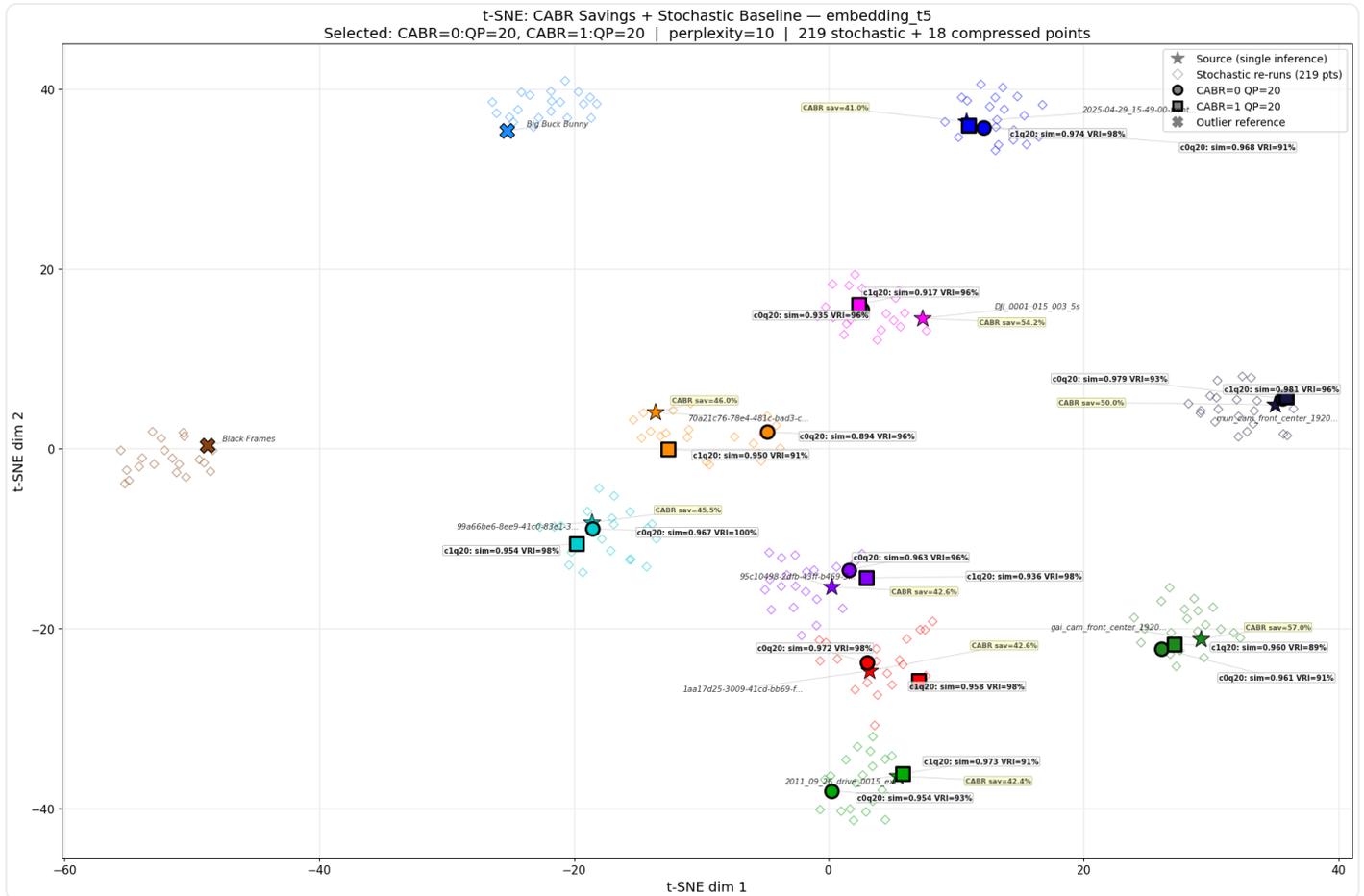
Notable observations:

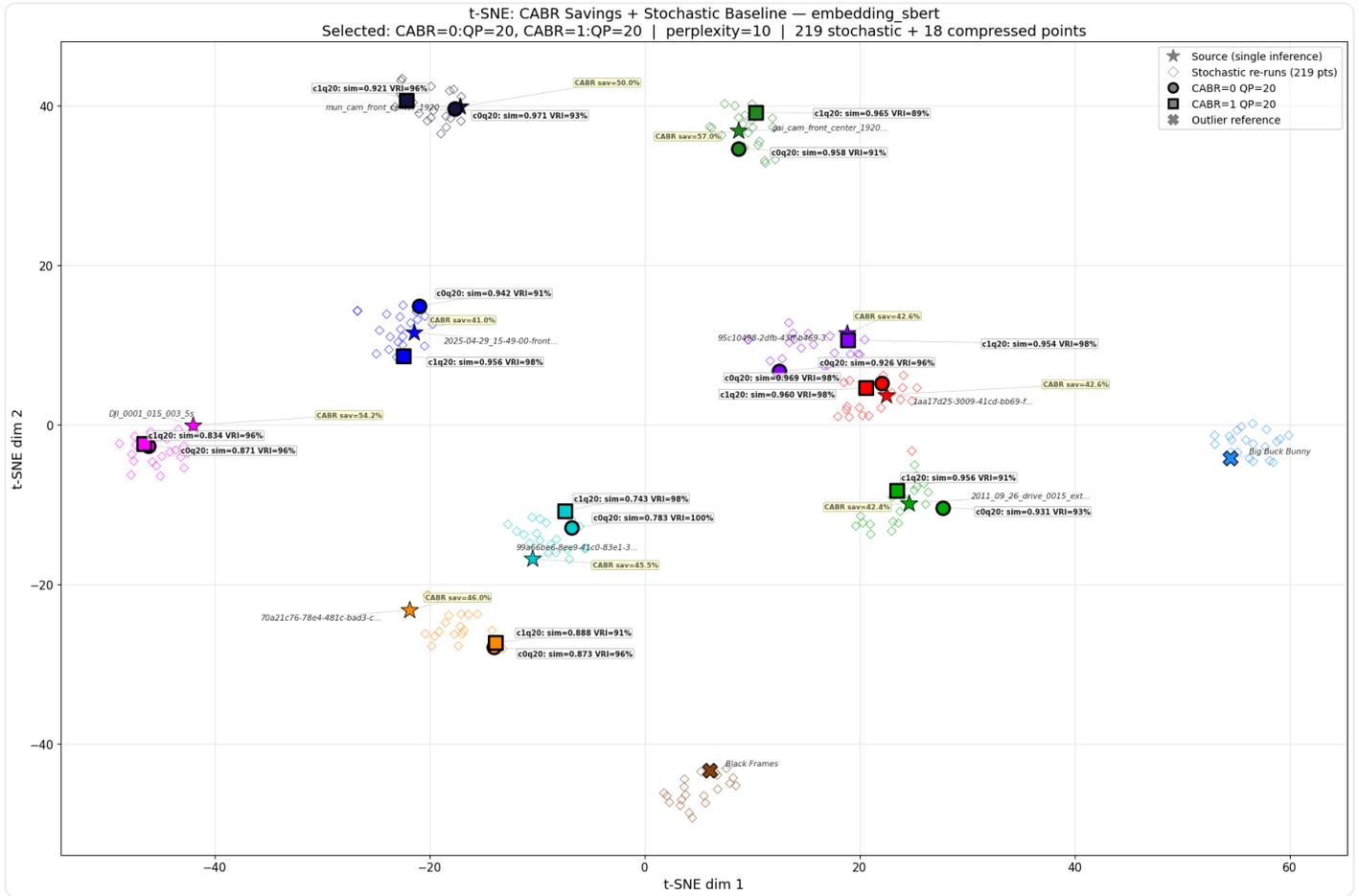
- Both CABR0 and CABR1 similarity values consistently fall within the stochastic range for each video across both embedding models. No compressed variant produces a similarity score below the stochastic minimum for its corresponding source video.
- In several cases (e.g., 70a21 T5: 0.894→0.950, 2011_09_26 T5: 0.954→0.973) the CABR-optimized variant is *closer* to the source caption than the unoptimized NVENC. This is consistent with expected stochastic variability in deterministic captioning of near-identical visual inputs — the model is not sensitive enough to distinguish between these compression levels.
- The lowest SBERT similarity scores appear for 99a66 (0.783/0.743) and DJI_0001 (0.871/0.834). However, these values fall within the stochastic ranges for those videos (99a66: 0.716–0.808; DJI_0001: 0.774–0.868), confirming the differences are within model noise.

4.3 t-SNE Cluster Visualization

To qualitatively assess inter-video and intra-video embedding structure, high-dimensional embeddings were projected to 2D using t-SNE. Note: t-SNE preserves local neighborhood structure but does not preserve absolute distances; this visualization is intended to complement, not replace, the quantitative cosine similarity analysis above.

Ahead are the t-SNE projections for both types of evaluated embeddings (T5, SBERT):





Clear and separate contextual clusters are visible for the videos in both T5 and SBERT projections. Compression variants (CABR_0 and CABR_1) consistently cluster with their respective source video's stochastic cloud rather than forming separate clusters, visually confirming that compression-induced embedding shifts are indistinguishable from the model's run-to-run variability.

Videos with high inter-video similarity:

Some video clusters sit closer together, corresponding to shared scene characteristics. Videos 1aa17, 95c10, and 2011_09_26 share similar daytime driving environments, lighting, and surroundings, resulting in adjacent but distinguishable clusters.

Similarly, nighttime videos 99a66 and 70a21 share contextual similarity evident in the t-SNE plot, though their surroundings differ enough to maintain distinct clusters (highway vs. urban intersection).

Videos with high contextual distinctiveness:

DJI_0001 (aerial footage) and mun_cam (European city street) show clear separation from all other clusters in both t-SNE projections, reflecting their unique visual content.

Reference videos:

The reference videos (Big Buck Bunny and all-black-frames) appear far from all driving video clusters in the t-SNE space, confirming that the embedding space meaningfully separates genuinely distinct content from compression-level variations.

4.4 Visual Realism Index (VRI) Analysis

While sections 4.1–4.3 assess whether CABR optimization changes what the captioning model says about a video, VRI assesses whether compression affects the model's *structured scene understanding* — addressing the possibility that semantic degradation could occur in specific classification categories that aggregate embedding similarity does not capture.

VRI overview:

VRI is a structured 45-field classification system covering three categories:

- **VRI Static (15 fields)** — Infrastructure and context: road type, lane configuration, crosswalks, surface condition, boundaries, overhead structures, etc.
- **VRI Dynamic (15 fields)** — Traffic and hazards: traffic lights, ego-vehicle maneuver, pedestrians, cyclists, emergency vehicles, etc.
- **VRI Details (15 fields)** — Counts and fine details: lane count, vehicle counts, sign readability, pedestrian distance, road debris, etc.

Each field is independently classified from the video by the Qwen2.5-VL model. Agreement rate measures the fraction of fields where the compressed variant's classification matches the source video's classification.

VRI Agreement Results at QP=20

Video	BR Savings	CABR0 VRI Agree	CABR1 VRI Agree
1aa17d25-3009-41cd-bb69-f3291d7754da	42.6%	97.8%	97.8%
2011_09_26_drive_0015_extract_000	42.4%	93.3%	91.1%
2025-04-29_15-49-00-front_000	41%	91.1%	97.8%
70a21c76-78e4-481c-bad3-c0db5d48834d	46%	95.6%	91.1%
95c10498-2dfb-43ff-b469-3671dc5b8b07	42.6%	95.6%	97.8%
99a66be6-8ee9-41c0-83e1-3e5353f90273	45.5%	100%	97.8%
DJI_0001_015_003_5s	54.2%	95.6%	95.6%
gai_cam_front_center	57%	91.1%	88.9%
mun_cam_front_center	50%	93.3%	95.6%

Average VRI agreement: CABR0 = 94.8%, CABR1 = 94.8%.

Observations:

- VRI agreement rates are consistently high (88.9–100%) for both compression variants, indicating that the structured scene classification is largely invariant to both baseline compression and CABR optimization.
- CABR1 agreement is comparable to CABR0 agreement across all videos — there is no systematic VRI degradation from CABR optimization. In fact, for several videos (2025-04-29, 95c10) CABR1 achieves *higher* agreement than CABR0.
- The videos with lowest embedding similarity (99a66 in SBERT, DJI_0001 in SBERT) show strong VRI agreement (97.8–100% and 95.6% respectively), suggesting the embedding distance is driven by caption wording variation rather than substantive semantic disagreement.

5. Discussion & Limitations

Key findings:

The contextual differences between CABR-compressed variants and their uncompressed sources fall within the captioning model's baseline stochastic variability across all tested videos, as measured by both T5-XXL and SBERT embeddings. VRI analysis further confirms that structured scene classification is preserved under CABR optimization, with average agreement rates of ~95% for both CABR0 and CABR1.

CABR achieved 41–57% bitrate savings with no measurable impact on any evaluated metric.

Limitations and caveats:

- **Sample size (n=9):** While videos span diverse scene types (daytime suburban, nighttime urban, aerial, multi-camera configurations) and bitrate ranges (2.6–63.7 Mbps compressed), a larger corpus would strengthen generalizability claims.
- **QP selection:** Tests were conducted at QP=20 which is not the exact Cosmos Curate default, however the results are expected to hold which will be verified in the next testing round.
- **Single CABR configuration:** Only the default CABR optimization mode was tested, representing the intended production configuration.
- **Semantic granularity:** Cosine similarity on mean-pooled embeddings captures overall semantic alignment but may not detect differences in fine-grained details (e.g., specific object counts, safety-critical scene elements). The VRI analysis partially addresses this concern by evaluating 45 individual classification fields, but some sub-field nuances may still be missed.
- **t-SNE visualization:** t-SNE does not preserve global distances and is sensitive to hyper-parameter choices (perplexity, learning rate). The t-SNE plots in this report are intended as qualitative support only; all quantitative claims are based on cosine similarity.

6. Conclusion

Across all 9 evaluated videos, CABR achieved 41–57% bitrate savings with no measurable impact on the Cosmos Curate AV captioning pipeline beyond the model's inherent output variability. This was validated independently through two embedding models (T5-XXL and SBERT) and confirmed by VRI perceptual quality assessment across 45 structured classification fields. Both compressed variants (CABR0 and CABR1) produced embedding similarities and VRI agreement rates that fall squarely within the stochastic noise floor of the captioning model itself.

Given this invariance, compression strategy can prioritize maximum bitrate savings while optimizing for a broader set of downstream requirements such as perceived visual quality.

7. Next Steps

- Evaluate Cosmos Curate Embed1 pipeline (distinct from the AV pipeline) sensitivity to compression.
- Expand test corpus to additional videos, scene types, and encoding configurations
- Evaluate more aggressive CABR configurations to determine the compression envelope within which pipeline invariance is maintained.

8. Appendix

A.1 Video Descriptions (Source Captions)

While most descriptions are thorough and accurate, some small set of captions contain text in Chinese, which while containing similar contextual information might affect embedding model similarity assessment. In the dataset Observed, this was not an issue but it is worth noting for future follow-up work.

Video

Caption

1aa17d25-3009-41cd-bb69-f3291d7754da	<p>The video captures a serene drive through a suburban area on a clear, sunny day. From the perspective of a dashboard camera mounted inside a vehicle, the viewer sees a two-lane road stretching ahead, flanked by lush green trees and utility poles. The road is relatively quiet, with only a few parked cars visible along the curbside, suggesting a calm residential or commercial zone. The sky is a vibrant blue without any clouds, and the bright sunlight casts distinct shadows on the road, enhancing the clarity and warmth of the scene. The surroundings include a mix of greenery and man-made structures, such as buildings and street signs, which add to the suburban ambiance. The overall atmosphere is peaceful, with no other vehicles or pedestrians in sight, allowing for an uninterrupted journey through this tranquil setting.</p>
95c10498-2dfb-43ff-b469-3671dc5b8b07	<p>The video captures a serene drive through an urban neighborhood on a cloudy day. From the perspective of a dashboard camera inside a car, the viewer sees the vehicle navigate a traffic intersection where all traffic lights are green, allowing smooth passage in both directions. The road stretches ahead, flanked by residential houses with well-maintained lawns and mature trees that provide a sense of tranquility. On the right side of the road, a gas station stands out, its structure blending into the suburban landscape. The sky overhead is overcast, casting a soft, diffused light that enhances the muted colors of the environment. The road itself is mostly empty, with only a few cars visible in the distance, adding to the calm and quiet atmosphere of the scene. The overall impression is one of a peaceful drive through a typical suburban area on a cloudy day.</p>
2011_09_26_drive_0015_extract_000	<p>The video captures a serene suburban scene under a clear blue sky, suggesting a bright and sunny day. A two-lane road stretches into the distance, flanked by lush green grass and tall trees that line both sides, creating a natural border. Parallel to the road on the left, railway tracks run alongside, adding an industrial touch to the otherwise tranquil setting. A black car is prominently featured, driving directly towards the camera, its sleek form contrasting against the lighter asphalt. The road markings are clearly visible, guiding the viewer's eye down the straight path. In the background, several other vehicles can be seen either parked or moving at a leisurely pace, contributing to the calm flow of traffic. The environment is peaceful, with no signs of pedestrian activity, allowing the focus to remain on the vehicles and the surrounding nature. The interplay of light and shadow from the trees and the road enhances the depth and texture of the scene, making it visually engaging. The overall atmosphere is one of quiet movement and natural beauty, inviting viewers to appreciate the simplicity and harmony of the setting.</p>
99a66be6-8ee9-41c0-83e1-3e5353f90273	<p>The video captures a serene nighttime drive on a multi-lane road, bathed in the soft glow of streetlights. The road stretches ahead, mostly deserted, with no other vehicles or pedestrians in sight. On the right side, tall buildings with illuminated windows stand prominently, their lights reflecting off the wet pavement, hinting at recent rainfall. To the left, a single-story house with warm lighting adds a touch of domesticity to the urban landscape. The sky above is pitch black, emphasizing the quiet solitude of the night. The road markings are crisp and well-defined, guiding the viewer's eye forward.偶尔，雨刷器轻轻动弹，表明道路上有薄薄的雾气或细雨。</p>
70a21c76-78e4-481c-bad3-c0db5d48834d	<p>The scene unfolds at night on a bustling urban street intersection, bathed in the soft glow of streetlights and building illumination. Multi-story buildings with lit windows stand tall on either side of the street, their facades reflecting the ambient light. Bare trees line the sidewalks, their branches reaching out into the dark sky, adding a stark contrast to the urban landscape. A few parked cars are scattered along the curbside, their metallic surfaces gleaming under the streetlights. Traffic lights hang above the intersection, displaying alternating red and green signals. The camera perspective suggests it is mounted on a vehicle, providing a first-person view as it navigates through the intersection. A black SUV is captured in motion, crossing the intersection from left to right, its headlights cutting through the darkness. The street itself is relatively quiet, with minimal pedestrian activity, enhancing the sense of calmness that pervades the scene. The overall atmosphere is serene yet vibrant, with the interplay of artificial lights creating a dynamic and engaging visual experience.</p>

2025-04-29_15-49-00-front_000

The video captures a bustling traffic intersection under a clear blue sky, suggesting a bright and sunny day. The intersection is equipped with multiple lanes and traffic lights, guiding the flow of vehicles. Several cars, including a white van and a black SUV, are seen navigating through the intersection. Notably, a large tour bus with "Ramon Tours" prominently displayed on its side drives from right to left across the frame, adding a dynamic element to the scene. The background showcases a picturesque hilly landscape covered in lush greenery, enhancing the serene yet lively atmosphere of the setting. The absence of pedestrians emphasizes the focus on vehicular movement within this urban environment.

DJI_0001_015_003_5s

The image captures a bustling urban intersection viewed from inside a vehicle, likely during the early evening or late afternoon as indicated by the soft, diffused lighting and long shadows. The road is moderately busy with several cars, including a white van labeled "Levant" and a silver SUV, all stopped at a red traffic light. The vehicle from which the photo is taken is positioned in the center lane, facing forward towards the intersection. On the left side of the road, a prominent building with "WeWork" signage stands out, indicating a commercial district. Adjacent to it is another building with a sign reading "Paz Bank," suggesting a mix of office and financial services. The right side of the intersection features a large, modern building with numerous windows, possibly a commercial or residential complex. Overhead, a bridge structure is visible, adding to the urban setting. The sky above is clear, and the overall ambiance is calm, with no immediate signs of heavy traffic or congestion. The scene is framed by the dashboard of the vehicle, reflecting the lights of the surrounding cars and the road ahead, enhancing the sense of being immersed in the moment.

gai_cam_front_center

The video captures a serene urban street scene, characterized by its modern architecture and orderly layout. On either side of the street, contemporary buildings rise, featuring clean lines and a mix of neutral tones. The street itself is relatively quiet, with several cars parked neatly along the curbs and a few vehicles moving at a slow pace. The presence of trees planted along the sidewalks adds a touch of nature to the otherwise concrete surroundings, offering patches of greenery that contrast with the urban environment. The sky above is predominantly clear with a few scattered clouds, suggesting a bright and pleasant day. The overall ambiance is calm and peaceful, hinting at a time of day when activity is minimal, possibly midday or early afternoon. Despite the absence of visible pedestrians, the scene conveys a sense of tranquility and order, typical of a well-maintained cityscape.

mun_cam_front_center

The image captures a bustling city street scene during the daytime under a clear blue sky. The street is flanked by parked cars on both sides, with a notable black Volvo XC60 prominently positioned in the foreground. Pedestrians are scattered along the sidewalks, some near the crosswalks, adding to the lively urban atmosphere. Traffic lights are visible at various intersections, with one showing a red signal, indicating a momentary pause in traffic flow. The trees lining the street are bare, suggesting that the season is likely late autumn or early spring. The architecture is characterized by multi-story buildings with classical facades, adorned with shops that have awnings providing shade for pedestrians. The buildings exhibit a mix of neutral tones, with some featuring light-colored stone and others painted in warm hues like yellow. The overall ambiance is calm yet vibrant, reflecting the dynamic nature of an urban environment during daylight hours. The presence of bicycles and the general layout of the street suggest a well-organized and pedestrian-friendly area.

A.2 Default Cosmos-Curate (AV pipeline) prompt:

```
_DEFAULT_PROMPT = ""
```

You are describing the driving video. Ensure the description is concise, precise, and focuses only on visible and verifiable details.

The description should include as many of the following elements as applicable:

1. Objects in the scene: Identify visible objects (e.g., cars, pedestrians, traffic lights, trees, houses).
2. Actions of objects: Highlight actions of objects in the scene (e.g., a car turning left, a pedestrian crossing the road).
3. Scene setting: Provide a general description of the environment (e.g., a busy street, a quiet neighborhood).
4. Weather conditions: Mention the visible weather (e.g., sunny, cloudy, rainy).
5. Time of day: Note the time of day based on the lighting (e.g., day, night).

Output:

[Your concise and informative description]

Please respond only in English and do not include Chinese words.

""

A.3 Cosmos-Curate (AV pipeline) prompt used for the custom VRI analysis:

```
_VRI_STATIC_PROMPT = ""
```

Analyze the driving video and classify the following 15 attributes based strictly on visible evidence.

CRITICAL INSTRUCTION FOR "UNKNOWN" vs "NONE":

- Select 'None' (or 'No') if the feature is CLEARLY NOT PRESENT in the scene.
- Select 'Unknown' ONLY if the visual conditions (darkness, blur, occlusion, glare) prevent you from determining the attribute.
- Select exactly ONE value per attribute. Do NOT combine multiple values.

Infrastructure & Layout

- Road Type: [Highway, Urban Street, Residential, Parking Lot, Rural Road, Tunnel, Bridge/Overpass, None, Mixed, Unknown]
- Lane Configuration: [One-way, Two-way, Divided Highway, None, Unknown]
- Lane Marking Style: [Solid Lines, Dashed Lines, Double Lines, None, Mixed, Unknown]
- Intersection Geometry: [4-Way, T-Junction, Roundabout, Merge/Split, None, Unknown]
- Crosswalk Presence: [Marked, Unmarked, None, Unknown]
- Sidewalk Location: [Left Side, Right Side, Both Sides, None, Unknown]
- Bike Lane Infrastructure: [Dedicated Lane, Shared/Sharrows, Protected Path, None, Unknown]
- Road Boundary: [Curb, Guardrail, Concrete Barrier, Grass/Ditch, Wall, None, Mixed, Unknown]
- Overhead Structures: [Traffic Light Gantry, Sign Bridge, Power Lines, Tunnel/Bridge, None, Unknown]
- Parking Arrangement: [Parallel, Perpendicular/Angled, Open Lot, Mixed, None, Unknown]

Context & Conditions

- Special Road Zones: [Construction Zone, School Zone, Bus Stop, Railway Crossing, None, Unknown]
- Road Surface Condition: [Smooth/New, Damaged/Potholes, Wet, Icy, Construction/Gravel, Snow-Covered, None, Mixed, Unknown]
- Vegetation Context: [Urban Trees, Dense Forest, Open Field, Desert/Barren, Mixed, Unknown]
- Building Density: [High-rise/Urban, Low-rise/Suburban, Industrial, Rural/Sparse, Mixed, Unknown]

- Visual Obstructions: [Sun Glare, Rain on Lens, Fog/Haze, Darkness, Clear, Unknown]

RESPONSE FORMAT:

Please output the results strictly as a list of Key: Value pairs.

Road Type: [Value]

Lane Configuration: [Value]

... (and so on)

""

_VRI_DYNAMIC_PROMPT = ""

Analyze the driving video and classify the following 15 dynamic attributes based strictly on visible evidence.

CRITICAL INSTRUCTION FOR "UNKNOWN" vs "NONE":

- Select 'None' (or 'No') if the object is CLEARLY NOT PRESENT.
- Select 'Unknown' ONLY if the object might be present but is too blurry or obstructed to identify.
- Select exactly ONE value per attribute. Do NOT combine multiple values.

Traffic Signals & Ego

- Traffic Light Status: [Red, Green, Yellow, Transitioning, Flashing, Off, None, Unknown]
- Ego-Vehicle Maneuver: [Going Straight, Turning, Lane Changing, Curving, Stopped/Stationary, Accelerating, Decelerating, Unknown]
- Lead Vehicle Action: [Moving Constant, Accelerating, Braking, Stopped, Turning, None Present, Unknown]

Vehicle Interactions

- Oncoming Traffic: [Yes - Moving, Yes - Stopped, No, Unknown]
- Vehicle Cutting-In: [Yes - From Left, Yes - From Right, No, Unknown]
- Vehicle Overtaking: [Yes - Passing Ego, No, Unknown]
- Turn Signal Usage: [Left Signal, Right Signal, Hazard Lights, None Visible, Unknown]
- Brake Light Activity: [Active on Lead Car, Active on Other Car, None Visible, Unknown]

Hazards & Vulnerable Users

- Pedestrian Activity: [Walking on Sidewalk, Crossing Road, Standing/Waiting, None, Unknown]
- Cyclist Activity: [Riding in Lane, Riding in Bike Lane, Crossing, None, Unknown]
- Stopped Vehicle Hazard: [Double Parked, Broken Down, Loading/Unloading, None, Unknown]
- Bus Activity: [Driving, Stopped at Stop, Boarding Passengers, None, Unknown]
- Truck/Heavy Machinery: [Driving, Working/Operating, Parked, None, Unknown]
- Emergency Vehicle: [Active Lights/Siren, Passive Driving, Parked, None, Unknown]
- Animal Hazard: [Crossing Road, On Roadside, Flying/Bird, None, Unknown]

RESPONSE FORMAT:

Please output the results strictly as a list of Key: Value pairs.

Traffic Light Status: [Value]

Ego-Vehicle Maneuver: [Value]

... (and so on)

""

```
_VRI_DETAILS_PROMPT = """
```

Analyze the driving video and classify the following 15 detailed attributes based strictly on visible evidence.

CRITICAL INSTRUCTION FOR COUNTS:

- Select 'None' implies a count of ZERO (0).
- Select 'Unknown' implies you cannot count due to visibility issues (too dark/blurry).
- Select exactly ONE value per attribute. Do NOT combine multiple values.

Counts (Visible Objects)

- Lane Count: [1, 2, 3, 4, 5+, Unknown]
- Same-Direction Vehicle Count: [None, 1-2, 3-5, 6-10, Heavy Traffic, Unknown]
- Opposite-Direction Vehicle Count: [None, 1-2, 3-5, 6-10, Heavy Traffic, Unknown]
- Parked Vehicle Count: [None, 1-2, 3-5, Row/Lot Full, Unknown]
- Pedestrian Count: [None, 1, 2-3, Group/Crowd, Unknown]
- Traffic Light Count: [None, 1, 2, 3, 4+, Unknown]
- Traffic Sign Count: [None, 1-2, 3-5, 6+, Unknown]
- Construction Cone/Barrel Count: [None, 1-3, 4-10, 10+, Unknown]

Fine Details & Reading

- License Plate Visibility: [Clearly Readable, Detectable but Blurry, None Visible, Unknown]
- Street Name Legibility: [Readable, Blurry/Unreadable, None Present, Unknown]
- Speed Limit Sign Value: [10-25, 30-45, 50-65, 70+, None Visible, Unknown]
- Pavement Text Markings: [Yes - Readable, Yes - Unreadable, None, Unknown]
- Lead Vehicle Color: [White/Silver, Black/Dark, Gray, Red/Blue/Color, None, Unknown]
- Nearest Pedestrian Distance: [Close (<5m), Medium, Far, None, Unknown]
- Road Debris/Object: [Tire/Trash, Rock/Branch, Puddle, Water/Flooding, None, Unknown]

RESPONSE FORMAT:

Please output the results strictly as a list of Key: Value pairs.

Lane Count: [Value]

Same-Direction Vehicle Count: [Value]

... (and so on)

```
"""
```