SAVER: A BENCHMARK FOR STRUCTURED UNDERSTANDING OF E-COMMERCE ADVERTISEMENT VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Producing high-quality e-commerce advertisement videos is costly. This motivates reusing the creative structure of proven, high-performing advertisements to promote new products. Such reuse requires understanding not only where the original product appears, but also how it is connected to demonstrations, spoken claims, on-screen text, and the overall commercial message. Existing advertisement benchmarks mainly evaluate answers to individual questions, without assessing whether models can produce a consistent, structured representation of an entire video for this purpose. We introduce **SAVER**, a benchmark for structured evaluation and representation of e-commerce advertisement videos, containing 6,224 videos across 37 top-level and 283 second-level industry categories. Each video is paired with a **StructuredScript annotation** that organizes global context, detailed content, and commercial information. Shared entity identities and timestamps connect content across shots and modalities, while evidence references connect commercial interpretations to supporting observations. We further introduce **Qwen-SAVER**, a Qwen2.5-Omni-7B model fine-tuned on five script-generation tasks. The model first describes the video content and then uses these descriptions to infer commercial information. We evaluate Qwen-SAVER alongside eight general-purpose baselines on 500 test videos. Compared with the untuned model, Qwen-SAVER improves the schema-conformance rate and commercial-claim annotation-correspondence micro-F1 by 81.8 and 40.6 percentage points, respectively. Finally, a qualitative study explores how field-level modifications to a StructuredScript can guide product-replacement video generation, illustrating its potential for advertisement reuse.

1 INTRODUCTION

E-commerce advertisements combine rapidly changing visuals, speech, and on-screen text to convey product functions, selling points, and purchase incentives. Because production is costly, successful advertisements are attractive templates for new products, but reuse requires an editable decomposition that preserves the overall creative design. Existing benchmarks such as VideoAds (Zhang et al., 2025b) and E-VAds (Liu et al., 2026) evaluate question-level responses rather than consistent, video-wide organization of entities, actions, speech, text, and commercial intent. Figure 1 illustrates this gap.

We introduce **SAVER**: Structure Ad Video Evaluation and Representation to evaluate the cross-shot identity, cross-modal timing, and evidence grounding needed for structured reuse. Its **StructuredScript annotations** separate global context, detailed content, and commercial information; shared identities, timestamps, and evidence references connect these records. SAVER contains 6,224 videos totaling 67.1 hours across 37 top-level and 283 second-level categories.

We fine-tune Qwen2.5-Omni-7B as **Qwen-SAVER** and compare it with eight general-purpose baselines on 500 test videos. We separately evaluate schema conformance, reference-graph diagnostics, annotation correspondence, and field fidelity. Relative to the original model, Qwen-SAVER improves schema conformance by 81.8 percentage points and product-entity correspondence micro-F1

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

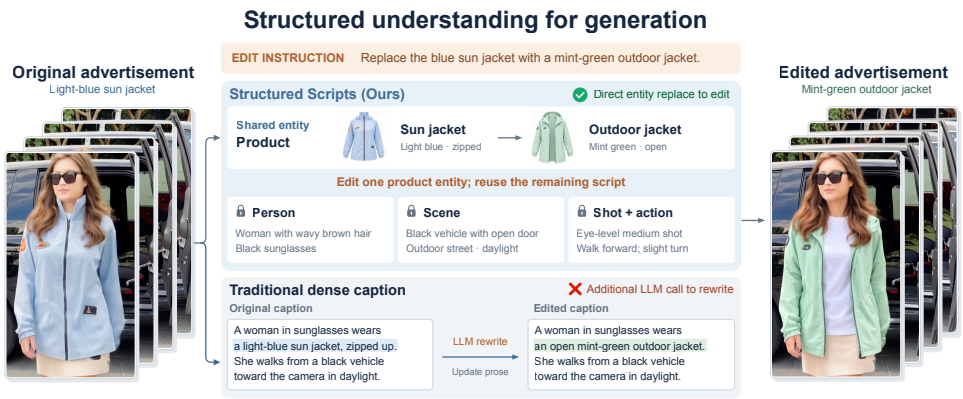


Figure 1: A StructuredScript exposes the product as an editable component, enabling replacement while preserving the person, scene, shots, and audio.

by 63.2 points. We also qualitatively study how direct field edits guide product-replacement video generation.

We summarize our main contributions as follows:

- **SAVER**: 6,224 e-commerce advertisements with StructuredScript annotations linked by identity, time, and evidence references.
- **Qwen-SAVER**: a supervised model compared with eight general-purpose baselines on 500 test videos using output-structure diagnostics and content metrics.
- **Structured editing**: an initial qualitative study of field-level control for advertisement reuse.

2 Related Work

2.1 Video Understanding Benchmarks

General-purpose benchmarks probe complementary aspects of video understanding: ActivityNet-QA (Yu et al., 2019) and NExT-QA (Xiao et al., 2021) test question answering over complex activities and temporal or causal actions; MVBench (Li et al., 2024) targets diverse temporal skills, and EgoSchema (Mangalam et al., 2023) stresses long-range reasoning. Video-MME (Fu et al., 2025) assesses multimodal comprehension across video durations, with audio and subtitles, while WorldSense (Hong et al., 2025) requires joint audio–visual reasoning. Early advertisement work (Hussain et al., 2017) annotated the actions and persuasive rationales conveyed by video ads. More recent benchmarks ask domain-specific questions: VideoAds (Zhang et al., 2025b) tests fast-paced advertisement perception and reasoning, AdsQA (Long et al., 2025) addresses persuasive and social intent, and E-VAds (Liu et al., 2026) evaluates perception and commercial reasoning in e-commerce short videos. The general-purpose and advertising datasets compared in Table 1 assess task-specific outputs rather than the consistency of a complete script. SAVER instead requires a structured script linking entities, temporally aligned events, and commercial interpretations across an entire advertisement.

2.2 Video Large Language Models

Video-language models have progressed from image–video alignment in Video-LLaVA (Lin et al., 2024) to video-specific training in VideoLLaMA3 (Zhang et al., 2025a) and temporal grounding for real-world shorts in ARC-Hunyuan-Video (Ge et al., 2025). Our comparisons cover vision-language baselines Qwen3-VL (Bai et al., 2025), Qwen3.5 (Qwen Team, 2026), InternVL3.5 (Wang et al., 2025), VideoLLaMA3, and Keye-VL-2.0 (Kwai Keye Team, 2026), as well as the audio-capable

Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), and e-commerce-oriented Valley3 (Chen et al., 2026). Other audio-visual systems, such as video-SALMONN 2 (Tang et al., 2025), optimize detailed video description alongside question answering. Audio matters here because spoken claims and sound can carry information absent from frames. Our baselines provide reference points for multimodal capability, not supervised solutions to our output schema. We finetune Qwen2.5-Omni to examine the effect of script-specific supervision against its original checkpoint.

2.3 Video Captioning and Structured Representations

ActivityNet Captions (Krishna et al., 2017) and Vid2Seq (Yang et al., 2023) pair localized events with descriptions, while MovieGraphs (Vicol et al., 2018) and Action Genome (Ji et al., 2020) add graph-structured descriptions of social situations and visual interactions. Audio-visual captioning approaches such as UGC-VideoCaptioner (Wu et al., 2025) and Omni-Captioner (Ma et al., 2025) emphasize detailed coverage of both modalities. TimeChat (Ren et al., 2024) and TimeChat-Captioner (Yao et al., 2026) add temporal grounding and structured scene-level captions; Script-a-Video (Tencent Hunyuan Team, 2026) links reference, shot, and audio streams, while OmniScript (Pu et al., 2026) generates hierarchical cinematic scripts. StructuredScript instead targets evidence-grounded commercial interpretation of e-commerce advertisements within a video-wide script.

3 The SAVER Benchmark

SAVER evaluates whether models can produce consistent, structured descriptions of entire e-commerce advertisement videos. Each video is paired with a StructuredScript annotation, which organizes global context, detailed content, and commercial information through shared identities, timestamps, and evidence references. This section describes the representation, dataset construction and characteristics, and annotation and quality control.

3.1 StructuredScript representation

StructuredScript separates *global context*, *detailed content*, and *commercial information*. `global` summarizes video-level properties and promoted entities. Detailed content comprises video-wide `entities`, visual `shots` containing scenes, subjects, and actions, and four `tracks` for speech, on-screen text, overlays, and non-speech audio. `commercial` records claims, pain–solution relations, target audiences, and calls to action as advertising messages, without verifying their real-world validity. Figure 2 shows an example.

Shared identifiers link entity appearances and mentions across shots and modalities. Shots and independently bounded events share a timeline, while evidence references ground narrative and commercial interpretations in observations. Appendix C details the fields and references.

Field values balance consistency and flexibility: *closed* vocabularies standardize attributes such as shot scale and camera motion, *semi-open* vocabularies support extensible labels such as location and narrative function, and *open* descriptions retain detailed appearances and actions. This design provides a common target without discarding ad-specific detail.

3.2 Dataset construction and characteristics

We construct the dataset through a multi-stage pipeline (Figure 3). We first assemble an initial pool of approximately 4.02 million advertisement videos by combining high-spend advertisements with randomly selected videos. Quality and content-risk screening narrows this pool to roughly 2.28 million candidates, from which we sample within industry and duration strata. Content-complexity filtering yields a final dataset of 6,224 unique videos with higher multimodal information density. Screening criteria, sampling weights, and coverage-exception rules are detailed in Appendix B.

Category coverage and duration. Our dataset covers 37 top-level and 283 second-level industry categories. Figure 3(b) presents the eight most frequent top-level categories and the five most

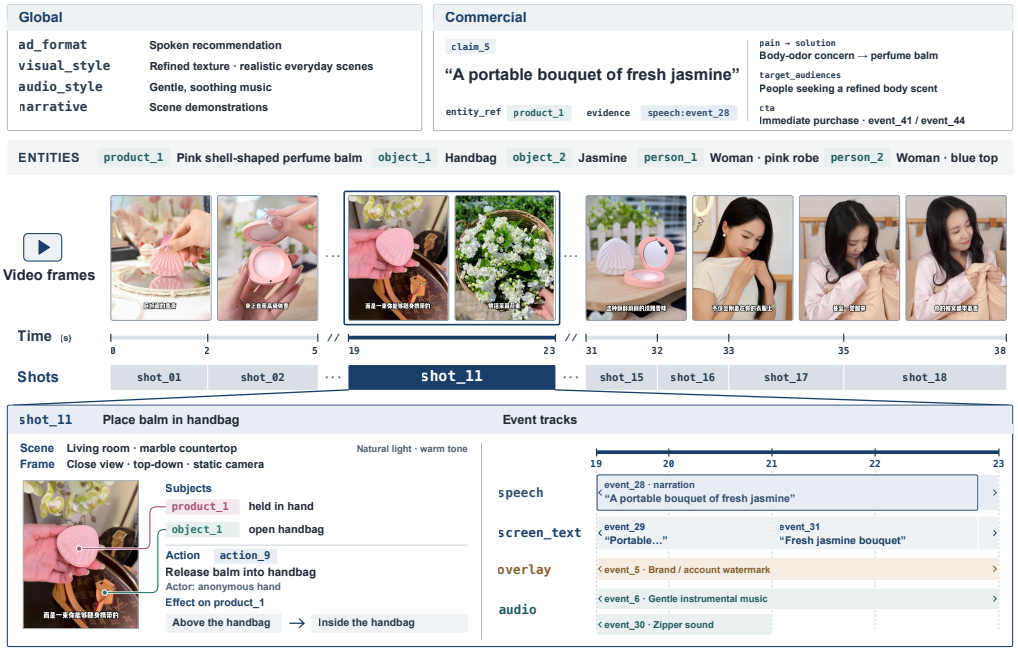


Figure 2: Selected StructuredScript annotations for a perfume-balm advertisement. The global and commercial panels summarize a spoken product recommendation and the claim "A portable bouquet of fresh jasmine," supported by the narration. Shared entity identities connect the pink shell-shaped perfume balm, handbag, jasmine, and two women across shots. The expanded shot shows the balm being released into an open handbag, including the resulting state change and independently timed speech, screen-text, overlay, and audio tracks.

frequent second-level categories within each. Videos have a mean duration of 38.8 seconds and a median of 29.4 seconds. As shown in Figure 3(c), short (< 15 s), medium (15–60 s), and long (60–120 s) videos account for 26.6%, 48.8%, and 24.6% of the dataset, respectively.

Multimodal information density. Following E-VAds (Liu et al., 2026), we characterize the corpus's multimodal information density using visual–audio–OCR (VAO) measures. Visual dynamic density (V_{den}) captures frame-feature dissimilarity, speech density (A_{den}) measures ASR output per second, and on-screen text density (O_{den}) measures OCR output per second, with corpus distributions shown in Figure 3(d–f). Since the original measurement code was not released, we reimplement the procedure and apply it to both E-VAds and SAVER; VAO values for the other benchmarks in Table 1 are taken from the E-VAds paper. These descriptive measurements show that the selected SAVER corpus combines frequent visual changes with dense speech and on-screen text.

3.3 Annotation and Quality Control

Script annotation. We construct annotation scripts in StructuredScript with a Gemini-based annotation pipeline (Figure 4(a)). Global context and entity identification establish a shared basis for the visual and event components. The visual component describes shots, while four event tracks record speech, on-screen text, overlays, and non-speech audio. A shared entity inventory and consistent identifiers link entities, events, and commercial records across annotation stages, while event annotations inform shot-level narrative and commercial interpretation. These components are integrated into a structured script, followed by a model-based review of the full video to revisit global and entity descriptions.

Quality control. Quality control distinguishes internal consistency from support in the source video. We call the conjunction of all construction-time constraints *full structural integrity*. Let

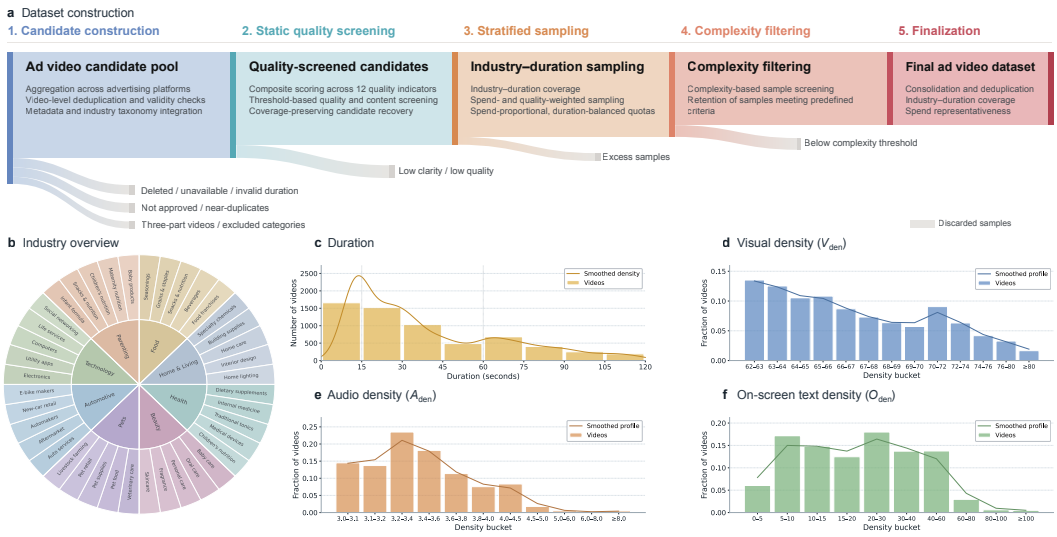


Figure 3: SAVER construction and corpus statistics: (a) multi-stage collection pipeline; (b) frequent top-level categories and subcategories; (c) video duration; and (d–f) visual, speech, and on-screen text density. Gray branches indicate discarded candidates; ribbon widths are schematic.

Table 1: Dataset scale and multimodal information density for general-purpose and advertisement video benchmarks. ‡: estimated total duration.

Benchmark	#Videos	Duration (h)	V_{den}	A_{den}	O_{den}
General-purpose video benchmarks					
VideoMME-short	300	6.7‡	31.65	1.16	4.32
MVBench	3,641	16.2‡	20.69	2.46	3.45
ActivityNet-QA	5,800	179.5‡	29.69	2.77	2.98
EgoSchema	5,063	253.2‡	25.76	0.00	4.24
Advertisement video benchmarks					
VideoAds	200	4.4‡	50.72	1.40	4.02
AdsQA	1,544	22.7	50.13	0.85	5.05
E-VAds	3,961	50.0	62.0	3.07	18.6
SAVER	6,224	67.1	67.4	3.57	24.4

$\mathcal{C}(s)$ denote the set of required structural predicates for an integrated script s, covering required fields and closed-set values, identifier uniqueness and reference resolution, temporal ordering and containment, and selected cross-shot state constraints. We summarize their joint outcome as

$$\text{Struct}(s) = \prod_{c \in \mathcal{C}(s)} \mathbf{1}[c(s) = 1]. \tag{1}$$

Thus, $\text{Struct}(s) = 1$ only when all required predicates hold. Unlike the prediction-time schema-conformance rate in Section 4.2, this construction-time check includes identifier, reference, temporal, and cross-shot constraints. It establishes whether the annotation forms a coherent script, not whether every description is visually correct.

Figure 4(b) illustrates a complementary evidence-based review in two stages. First, a text LLM decomposes scene and action descriptions into atomic visual statements, which are paired with frames from the corresponding shots. Overlay and entity annotations are instead linked directly to visual evidence using event intervals or reference shots. Second, a VLM checks these verification items against the selected frames with task-specific prompts. The illustrated verdicts distinguish contradicted descriptions from insufficient visibility or ambiguity, and overlay checks retain an uncertain outcome alongside presence or absence. For a verification item u,

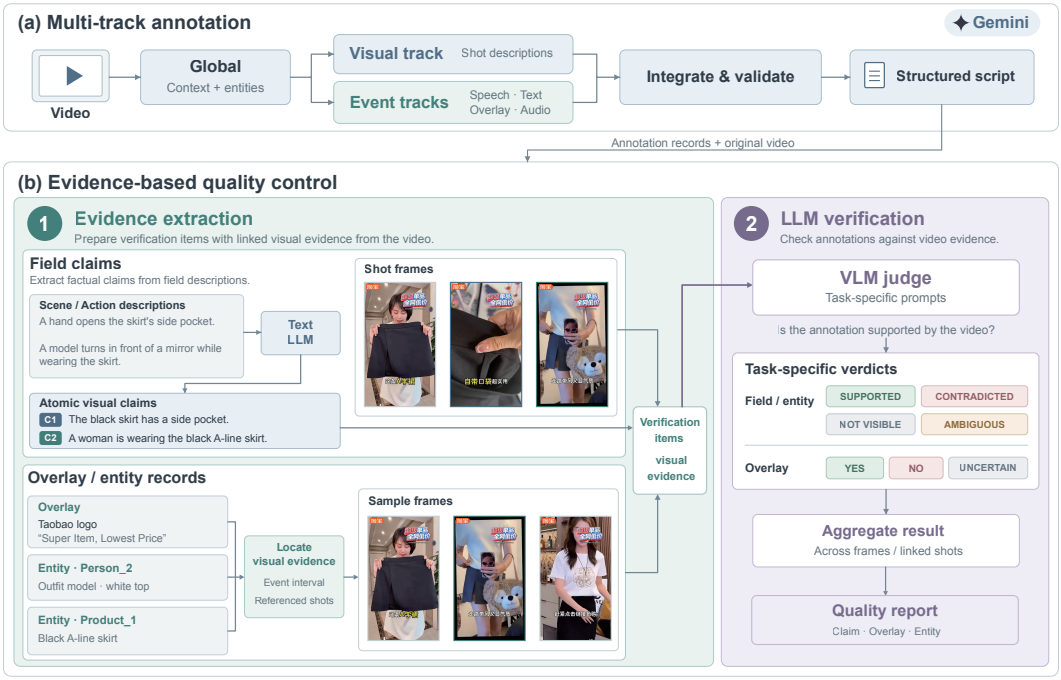


Figure 4: Script annotation and evidence-based quality review. (a) Global context and entities guide shot descriptions and four event tracks, which are integrated and structurally checked. (b) Atomic descriptions and overlay/entity annotations are paired with frames and verified by a VLM; the skirt advertisement illustrates the resulting quality checks.

let $\mathcal{F}_u$ be its selected evidence frames and let $z_{uf} \in \mathcal{A}$ be the verdict on frame f, where $\mathcal{A} = \{\text{support}, \text{contradiction}, \text{uncertainty}\}$. We retain the frame-level outcome proportions

$$\rho_{u,a} = \frac{1}{|\mathcal{F}_u|} \sum_{f \in \mathcal{F}_u} \mathbf{1}[z_{uf} = a], \qquad a \in \mathcal{A}, \qquad \sum_{a \in \mathcal{A}} \rho_{u,a} = 1. \tag{2}$$

These proportions are aggregated across frames or linked shots into a quality report, so insufficient visibility remains distinct from contradictory evidence.

Both $\mathrm{Struct}(s)$ and $\rho_{u,a}$ are construction-time audit quantities for StructuredScript annotations, not model-performance scores. The metrics in Section 4.2 are computed only after annotations are finalized and assess generated scripts separately.

4 Model and Evaluation

We introduce Qwen-SAVER for generating StructuredScript descriptions from advertisement videos and report three metric families: schema conformance, reference-graph diagnostics, and content agreement with StructuredScript annotations. Content agreement is further decomposed into annotation correspondence and field fidelity.

4.1 Qwen-SAVER Model

Qwen-SAVER adapts Qwen2.5-Omni-7B (Xu et al., 2025a) using LoRA. One model is trained on five tasks: global context and entities, speech and audio, screen text and overlays, shot descriptions, and commercial interpretation. The first four extract observable content; the last infers commercial information from it.

Because long descriptions can dominate token-level training, we additionally balance short but critical groups: structural tokens, identifiers, closed-set values, and temporal or character indices.

For example i, let $\mathcal{T}_i$ be all supervised positions and $\mathcal{T}_{ig}$ those in critical group $g \in \mathcal{G}_i$. With $\ell_{it} = -\log p_\theta(y_{it} \mid x_i, y_{i,<t})$, the objective is

$$\mathcal{L}_{\text{train}} = (1-\lambda)\frac{\sum_{i=1}^{B}\sum_{t\in\mathcal{T}_i}\ell_{it}}{\sum_{i=1}^{B}|\mathcal{T}_i|} + \frac{\lambda}{B}\sum_{i=1}^{B}\frac{1}{|\mathcal{G}_i|}\sum_{g\in\mathcal{G}_i}\frac{\sum_{t\in\mathcal{T}_{ig}}\ell_{it}}{|\mathcal{T}_{ig}|}, \qquad \lambda = 0.25. \tag{3}$$

Thus, all script tokens remain supervised while each critical group contributes independently of its length. Training and assembly details are provided in Appendix E.

4.2 Evaluation protocol

Our evaluation follows four stages. Given model predictions for test videos and their fixed StructuredScript annotations, we first measure schema conformance and reference-graph integrity, then align predicted and annotated records, measure record-level correspondence, and finally score fields, timestamps, and text within the established correspondences. We report these metric families separately.

Output-structure diagnostics. We first diagnose the structure of each output. **Schema conformance rate** is the percentage of videos whose outputs satisfy all required fields, types, and constrained values. It is an output-format measure, not a test of complete script validity or editability. **Reference-graph integrity** is reported separately and pools applicable instances across videos to evaluate identifier validity and the resolution of entity and evidence references. Components with no applicable instance are N/A. Schema nonconformance does not remove a prediction from subsequent content scoring, and schema conformance does not imply valid identifiers, resolved references, or correct temporal and semantic links.

Prediction–annotation alignment. We next establish one-to-one prediction–annotation correspondences. Entities are matched semantically within type; shots and multimodal event tracks are matched by time, prioritizing pairs with tIoU at least 0.5; and actions are matched within aligned shots. Commercial records are aligned by category after entity and evidence correspondences are available. This step determines which records are compared but does not score their internal fields.

Record-level annotation correspondence. Aligned pairs are true positives, while unmatched predictions and annotations are false positives and false negatives. Pooling these counts across videos gives a micro-F1 for each record category. The main **shared correspondence** score macro-averages seven category scores: Product, Action, Claim, Pain–solution, Target, Screen text, and Overlay. The broader Entity category is excluded to avoid overlap with Product; Speech and Audio are reported separately because some baselines are vision-only.

Field, temporal, and text metrics. Finally, we score the content of aligned records. **Matched-field agreement** measures supported values reproduced correctly within matched parents: closed-set fields use normalized equality, semi-open fields use a fixed semantic judge, and references are compared after correspondence projection. Missing values are incorrect, while unmatched parents are already penalized by record correspondence. Temporal fields use video-macro tIoU, descriptive text uses video-macro BERTScore-F1, and verbatim text uses video-macro exact-match accuracy. Appendix G provides the full matching and scoring rules and additional field-level results.

5 Experiments

We compare Qwen-SAVER with general-purpose models and examine how changes to individual script fields can guide video generation for advertisement reuse.

Table 2: Output-structure diagnostics and annotation correspondence (%, higher is better). Schema conformance is a video-level rate covering required fields, types, and constrained values only; the three reference-graph columns are separate instance-micro validity rates. None alone measures full structural integrity as defined by $\mathrm{Struct}(s)$. Shared macro-F1 averages seven non-audio categories; Table 10 reports each category.

		Output-structure diagnostics				Annotation correspondence
Method	Size	Schema conform.	ID Valid	Entity Ref.	Evidence Ref.	Shared macro-F1
Qwen3-Omni	30B-A3B	0.0	**79.9**	45.7	**83.1**	42.4
Keye-VL-2.0	30B	0.0	72.3	13.2	14.2	13.2
Qwen2.5-Omni	7B	0.0	38.4	18.7	16.2	28.5
Qwen3-VL	8B	0.2	64.6	73.8	58.3	33.4
Qwen3.5	9B	0.0	79.0	66.1	78.8	24.2
InternVL3.5	8B	7.4	77.7	68.1	0.0	7.9
VideoLLaMA3	7B	0.0	51.6	37.6	40.3	9.9
Valley3	8B	0.0	52.1	66.5	36.8	33.6
Qwen-SAVER (ours)	7B	**81.8**	64.6	**99.0**	48.4	**58.7**

Table 3: Selected field-level agreement and localization results (%, higher is better). The first three columns report agreement on reference-supported values in matched parent records; semi-open fields use the fixed semantic judge. Temporal columns report video-macro tIoU over aligned records. Text columns report video-macro BERTScore-F1 or exact-match accuracy over reference-supported records; missing predictions receive zero. Audio-dependent results are N/A for vision-only models.

		Matched-field agreement			Temporal localization			Descriptive text		Verbatim text	
Method	Size	Form	Predicate	Claim Type	Action	Speech	Screen	Action	CTA	Speech	Claim
Qwen3-Omni	30B-A3B	65.4	9.7	25.5	2.7	2.3	26.2	1.9	**85.9**	0.3	16.9
Keye-VL-2.0	30B	63.0	15.2	N/A	3.2	N/A	20.6	2.4	0.0	N/A	0.0
Qwen2.5-Omni	7B	45.5	11.8	10.5	1.7	0.3	7.0	1.1	68.4	0.1	11.0
Qwen3-VL	8B	58.1	8.9	24.2	2.1	N/A	15.6	1.5	56.2	N/A	14.0
Qwen3.5	9B	52.5	9.8	36.2	4.3	N/A	22.7	3.2	34.2	N/A	0.5
InternVL3.5	8B	44.1	N/A	10.1	0.0	N/A	0.0	0.0	3.7	N/A	0.2
VideoLLaMA3	7B	47.2	N/A	3.3	0.0	N/A	8.4	0.0	13.7	N/A	4.4
Valley3	8B	47.1	8.3	20.2	1.4	3.5	2.6	0.8	70.5	0.5	13.2
Qwen-SAVER (ours)	7B	**69.0**	**40.0**	**61.6**	**43.0**	**22.6**	**30.2**	**32.9**	82.3	**6.9**	**28.0**

5.1 Model comparison

Experimental setup. We evaluate on a common set of 500 SAVER videos using fixed, automatically generated StructuredScript annotations. Qwen-SAVER is fine-tuned with LoRA using separate training and validation partitions; Appendix E reports their sizes and optimization settings. The eight general-purpose baselines receive no task-specific fine-tuning. We use the original Qwen2.5-Omni-7B checkpoint as a direct before-and-after comparison. All models generate StructuredScript predictions, with audio supplied only to audio-capable models. Tables 2 and 3 report structural validity, record-level correspondence, and selected field-level results. Appendix D details the shared input and inference protocol.

Overall advantages. Qwen-SAVER provides the strongest overall balance across the reported diagnostics and content-agreement measures. General-purpose models can produce locally valid identifiers or references while rarely conforming to the complete schema; in contrast, Qwen-SAVER combines high schema conformance with the best shared correspondence and entity-reference validity. Its advantage extends to most selected fields, temporal tracks, and verbatim-text measures, especially for actions and commercial claims. These results suggest that task-specific supervision improves the organization of a complete StructuredScript rather than only the recognition of individual video elements. The comparison with Qwen2.5-Omni evaluates the fine-tuned system as a whole and does not isolate the contribution of the key-field loss. Performance is less consistent

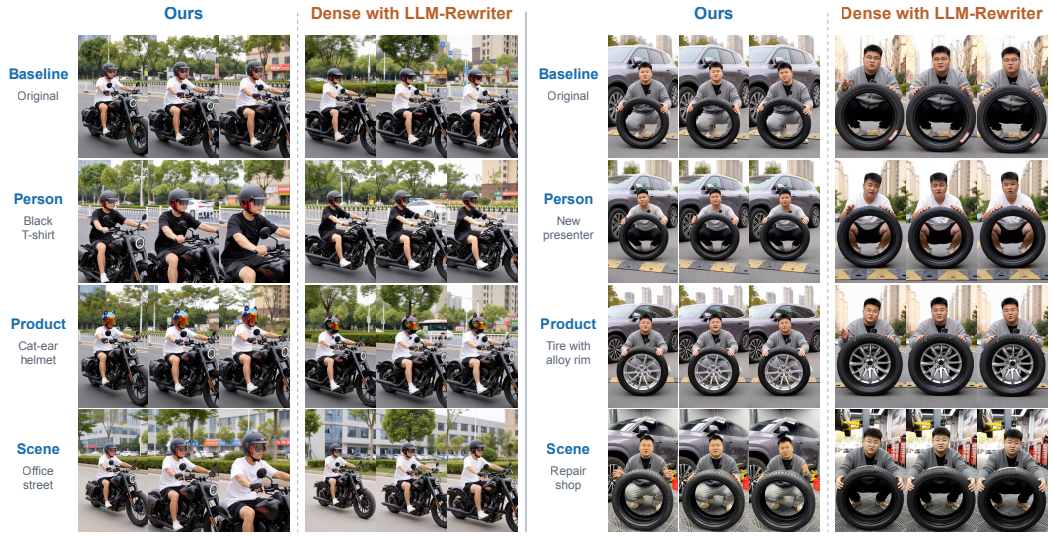


Figure 5: Qualitative editing results for a motorcycle-helmet advertisement (left) and a tire advertisement (right). Rows edit the person, product, or scene relative to the baseline. *Ours* directly modifies the corresponding StructuredScript field, whereas *Dense* uses an LLM-rewritten dense description.

for evidence references and audio-track correspondence, indicating that fine-grained cross-record grounding remains challenging. Detailed category-level results are reported in Appendix 10.

5.2 Qualitative Study: Structured Editing

We study whether StructuredScript can serve directly as an editing interface without an additional LLM rewriter. Figure 5 compares edits to the `person`, `product`, and `scene` fields in two advertisements. Our method modifies only the target field and uses a deterministic template to construct the generation prompt, whereas *Dense* asks an LLM to rewrite a dense description. Both use the same self-hosted distilled version of the MiniMax H3 video generator with identical sampling settings. Detailed settings are provided in Appendix F.

In both the motorcycle-helmet and tire examples, structured editing realizes the requested person, product, and scene changes while largely preserving the remaining composition and temporal content. Its outputs are qualitatively comparable to those produced from rewritten dense descriptions, indicating that explicit field edits can provide sufficient generation control without a separate rewriting stage. These examples are illustrative rather than a general evaluation across products, edits, or generators.

6 Discussion and Conclusion

We introduce SAVER for structured understanding of e-commerce advertisement videos. Its StructuredScript annotations organize global context, detailed content, and commercial information through shared entity identities, timestamps, and evidence references. A layered evaluation protocol separately reports schema conformance, reference-graph diagnostics, and content agreement with pipeline annotations. Content agreement is decomposed into annotation correspondence and field fidelity. Qwen-SAVER learns to understand advertisement videos and generates StructuredScript descriptions. The model comparisons show improvements after supervised fine-tuning, while revealing remaining differences across script components and modalities. StructuredScript also provides an explicit field-level interface for video editing, motivating the initial qualitative study of advertisement reuse.

These findings support structured scripts as a useful target for advertisement understanding, but they should be interpreted within the study's scope. Annotation scripts are automatically generated, audio

comparisons exclude vision-only models, and the editing study is qualitative. Appendix A provides further limitations and future directions.

AI USE STATEMENT

We used generative AI tools, including LLM-based coding and writing assistants, during parts of the research and writing workflow. These tools assisted with implementing and debugging research code; identifying and summarizing potentially relevant literature; supporting experimental execution, including configuration, data processing, and preliminary result analysis; and drafting and polishing portions of the manuscript. The authors checked suggested references against the original sources, reviewed and tested AI-assisted code and experimental procedures, and verified reported quantitative results against the corresponding experiment outputs. All AI-assisted prose was reviewed and revised by the authors for accuracy, attribution, and consistency with the underlying work. The authors take full responsibility for the final content of this paper, including all text, claims, results, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses real-world advertisement videos that may depict identifiable people, brands, products, and potentially sensitive contextual information. The SAVER dataset, including the StructuredScript annotations and other distributable metadata, is made available under the Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license; see `DATA-LICENSE`. It is intended for non-commercial academic research subject to the terms of this license. For commercial licensing or uses beyond non-commercial research, interested parties should contact the authors.

REPRODUCIBILITY STATEMENT

StructuredScript is described in Section 3.1, with the full schema intended for inclusion in the supplementary materials. Schema and reference-graph checks operate on generated scripts, whereas annotation- correspondence and field-fidelity metrics use the automatically generated StructuredScript annotations. Section 4.2 defines the matching and aggregation protocol; the evaluation code is intended for public release. Data construction and annotation are described in Sections 3.2 and 3.3, with additional construction details in Appendix B. Supervised fine-tuning details for Qwen-SAVER are given in Appendix E. Limitations and the planned data-release scope are discussed in Appendix A.

REFERENCES

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zeyu Chen, Guanghao Zhou, Min Yang, Qixiang Yin, Ziwang Zhao, Huanjin Yao, Pengjiu Xia, Cen Chen, and Minghui Qiu. Valley3: Scaling omni foundation models for e-commerce. *arXiv preprint arXiv:2605.01278*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yangyu Tao, Han Hu, Di Wang, and Ying Shan. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts, 2025. URL `https://arxiv.org/abs/2507.20939`.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. *arXiv preprint arXiv:2502.04326*, 2025. URL `https://arxiv.org/abs/2502.04326`.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1705–1715, 2017.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 706–715, 2017.

Kwai Keye Team. Kwai Keye-VL-2.0 technical report. *arXiv preprint arXiv:2606.10651*, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.342.

Xianjie Liu, Yiman Hu, Liang Wu, Ping Hu, Yixiong Zou, Jian Xu, and Bo Zheng. E-VAds: An e-commerce short videos understanding benchmark for MLLMs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, Yang Liu, Jiaheng Ma, and Bowen Zhou. AdsQA: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23396–23407, 2025.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng Siong Chng, and Xie Chen. Omni-Captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*, 2025. URL https://arxiv.org/abs/2510.12720.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Junfu Pu, Yuxin Chen, Teng Wang, and Ying Shan. OmniScript: Towards audio-visual script generation for long-form cinematic video. *arXiv preprint arXiv:2604.11102*, 2026. URL https://arxiv.org/abs/2604.11102.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen Blog, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. URL https://arxiv.org/abs/2506.15220.

Tencent Hunyuan Team. Script-a-Video: Deep structured audio-visual captions via factorized streams and relational grounding. *arXiv preprint arXiv:2604.11244*, 2026. URL https://arxiv.org/abs/2604.11244.

Paul Vicol, Makarand Tapaswi, Lluís Castrejón, and Sanja Fidler. MovieGraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8581–8590, 2018.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Peiran Wu, Yunze Liu, Zhengdong Zhu, Enmin Zhou, and Junxiao Shen. UGC-VideoCaptioner: An omni UGC video detail caption model and new benchmarks. *arXiv preprint arXiv:2507.11336*, 2025. URL https://arxiv.org/abs/2507.11336.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, 2021.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10714–10726, 2023.

Linli Yao, Yuancheng Wei, Yaojie Zhang, Lei Li, Xinlong Chen, Feifan Song, Ziyue Wang, Kun Ouyang, Yuanxin Liu, Lingpeng Kong, Qi Liu, Pengfei Wan, Kun Gai, Yuanxing Zhang, and Xu Sun. TimeChat-Captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions. In *Proceedings of the Forty-Third International Conference on Machine Learning (ICML)*, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. doi: 10.1609/aaai.v33i01.33019127.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Zheyuan Zhang, Wanying Dou, Linkai Peng, Hongyi Pan, Ulas Bagci, and Boqing Gong. VideoAds for fast-paced video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

A Additional limitations and future work

A.1 Limitations

Agreement with automatically generated annotations does not independently establish video accuracy: the annotations may contain errors or omissions. The model-based annotation review is an audit procedure, not a reported human-validated error rate; semi-open field scores also depend on an LLM judge. Audio-dependent results apply only to the four audio-capable systems; the five vision-only baselines do not generate speech or audio tracks. Despite high schema conformance, Qwen-SAVER remains weaker on evidence references and audio-track correspondence, so format-conforming outputs do not guarantee cross-record grounding. The editing study is qualitative and does not establish generality across products, edits, or generators. Data-use terms and the intended evaluation-code release are stated in the ethics and reproducibility statements.

A.2 FUTURE WORK

Future work could combine structural feedback with video-grounded content supervision, and evaluate editing across more categories and requests. Schema conformance or reference validity alone cannot certify the underlying interpretation.

B DATASET CONSTRUCTION DETAILS

This appendix instantiates the collection outline in Section 3.2. we remove repeated identifiers before finalization. The candidate pool combines high-spend advertisements with randomly selected advertisements so that the former retain commercial representativeness while the latter preserve long-tail categories and creative styles. The resulting raw pool contains 4,023,196 candidate records.

B.1 QUALITY AND CONTENT-RISK SCREENING

Each candidate receives a 12-term quality score. Five positive signals (overall quality, high-quality likelihood, visual aesthetics, cover aesthetics, and clarity) enter directly, while seven risk signals enter as their complements. Thus, with positive scores p_i and risk scores r_j, we use

$$q = \frac{1}{12}\left(\sum_{i=1}^{5} p_i + \sum_{j=1}^{7}(1 - r_j)\right). \tag{4}$$

Missing individual signals receive the neutral value 0.5 when computing q. We additionally apply the hard exclusions in Table 4; these guard against clips that a high average score alone would not reliably exclude.

Table 4: Hard quality and content-risk exclusions. Risk scores increase with the indicated undesirable property.

Signal or condition	Exclude when
Low-quality risk	score ≥ 0.50
Visual clarity	score < 0.30
Cover aesthetics	score < 0.20
Visual-aesthetics score	score < 0.20
Sexual-content risk (either of two detectors)	score ≥ 0.60
Clickbait risk	score ≥ 0.85
Intrusive cover-text risk	score ≥ 0.70
Low-quality-inducement risk	score ≥ 0.70
Additional platform content-risk signal	score ≥ 0.60
Missing visual-quality guard	clarity, cover aesthetics, and visual aesthetics all equal 0

The screen retains 2,276,042 candidates. To avoid eliminating a second-level-industry–duration cell solely because all of its candidates triggered a hard filter, we restore the highest-scoring rejected candidate in each empty cell. Ten candidates were restored in this way and are recorded as explicit coverage exceptions. These exceptions may also bypass the VAO complexity gate described in Appendix H.

B.2 STRATIFIED SAMPLING AND REPLENISHMENT

We assign each video to a short (< 15 s), medium (15–60 s), or long (60–120 s) duration bin and sample in two stages. Stage 1 guarantees coverage in each second-level-industry–duration cell. Stage 2 fills the remaining budget within top-level-industry–duration cells, allocating industry quotas in proportion to aggregate spend using largest-remainder apportionment and balancing the quota across duration bins. Stage 1 uses an approximately equal mix of high-spend and random-source candidates; Stage 2 uses an approximately 60/40 mix. A candidate admitted in Stage 1 is not eligible again in Stage 2.

Within each source and stratum, selection is deterministic weighted sampling without replacement. Let C be normalized $\log(1+\text{spend})$, $Q = \text{clip}(q, 0.05, 1)^2$, and $U \in (0,1)$ a deterministic uniform draw. We rank candidates by the deterministic weighted-sampling priority

$$k = -\frac{\log U}{w_C C + w_Q Q + w_U}, \tag{5}$$

selecting smaller priorities first, with source-specific weights in Table 5. The randomized priority avoids a deterministic ordering among otherwise similar candidates while the quality and spend terms balance presentation quality with real-world advertising exposure.

Table 5: Sampling weights (w_C, w_Q, w_U) for spend, quality, and the randomized-selection baseline.

Stratum source	w_C	w_Q	w_U
Stage 1 high-spend	0.40	0.45	0.15
Stage 1 random source	0.10	0.60	0.30
Stage 1 within-cell fallback	0.20	0.60	0.20
Stage 2 high-spend	0.45	0.40	0.15
Stage 2 random source	0.10	0.60	0.30
Stage 2 within-cell fallback	0.20	0.50	0.30

VAO is evaluated only after this inexpensive screen. The ordinary gate requires $A_{\text{den}} > 3$ and V_{den} at $d = 10 > 62$. When a sampled video has missing metrics or fails the VAO gate, we draw its replacement from the same stratum rather than relaxing the gate or borrowing a candidate from another cell. The ten coverage exceptions described above may bypass this gate; all other retained videos pass both the static screen and the VAO criterion. The final corpus contains 6,224 unique videos. This construction makes category and duration coverage explicit while intentionally retaining dense multimodal advertisements.

C StructuredScript field details

StructuredScript distinguishes video-wide identities, local observations, and commercial interpretations (Section 3.1). Table 6 summarizes its annotation units.

Table 6: StructuredScript components and their principal links.

Component	Annotation unit and key content	Links and timing
`meta`	Video ID, duration, frame rate, resolution, schema version	Video-level metadata
`global`	Language, ad format, styles, narrative structure	Promoted entity references
`entities`	Persistent people, products, objects, virtual referents	Video-wide entity IDs
`shots`	Frame, scene, local subjects, actions, narrative	Shot/action times; entity and evidence references
`tracks`	Speech, screen text, overlays, non-speech audio	Independently timed event IDs and optional entity links
`commercial`	Claims, pain–solution links, audiences, CTA	Promoted entities and typed evidence references

A `subject` is an entity in one shot. Commercial records describe what the advertisement communicates, not whether its claims are true.

C.1 Entity and shot records

The video-wide entity inventory has four types: `person`, `product`, `object`, and `virtual`. Each entry has a unique `entity_id`, an optional name and role, and type-specific attributes. Person demographics are annotation categories, not verified personal information; product affordances

describe observable use, not verified advertising benefits. Virtual referents include brands, services, and applications. The same identifier is reused across appearances and mentions.

Each shot has an ID, interval, transitions, and optional mixed-media type. Its `frame` and `scene` describe camera and setting; `subjects` link shot-local visibility and state to video-wide entities. Actions have IDs and within-shot intervals, observable predicates, optional participant references, and any observed state changes. Narrative stage and function labels cite observations through evidence references. A local subject without a stable identity need not have an entity link.

C.2 Event tracks and commercial records

The four tracks contain independently timed `speech`, `screen_text`, `overlay`, and non-speech `audio` events; their intervals may cross shot boundaries. Speech retains the utterance and optional speaker/mention links. Screen text retains literal wording and an optional normalized frame box; overlays describe graphics, not a second transcription. Audio records music, effects, or ambience, with a source link only when identifiable. Speech emphasis spans use zero-based, half-open character offsets within the utterance.

Commercial records comprise claims, pain–solution links, intended audiences, and a call to action (CTA). They connect promoted entities to supporting shots, actions, or events. Claim text records an advertised message, not a verified benefit; audience descriptions indicate who the ad addresses, not measured viewer demographics.

C.3 Value types, timing, and references

Closed values use fixed vocabularies; semi-open labels allow extensions, while open text retains details. Required nullable keys differ from omitted keys, and absent lists are represented as empty arrays; neither convention proves that unrecorded content is absent. Entity, shot, and claim IDs are unique in their video-level collections; action IDs across shots and event IDs across tracks are unique. Time ranges use seconds on a common timeline; actions lie within their shots, but track events may overlap shots. Evidence references name a `source_type` and a resolving `source_ref`; allowed sources are in Table 7. A valid link does not establish that its target actually supports the interpretation.

Table 7: Permitted evidence sources by referring component. Each source must resolve to the corresponding event, action, or shot record.

Referring component	Permitted `source_type` values
Shot narrative	`speech`, `screen_text`, `overlay`, `audio`, `action`, `shot`
Claims and CTA	`speech`, `screen_text`, `overlay`, `action`, `shot`
Pain–solution links	`speech`, `screen_text`, `action`, `shot`
Target audiences	`speech`, `screen_text`, `shot`

D Model comparison protocol

D.1 Shared tasks and media

All nine systems use the same frozen manifest of 500 test videos, the same fixed StructuredScript annotation snapshot, and the same five-stage task graph: `global_entities`, `speech_audio`, `screen_overlay`, `shots`, and `commercial`. The first task establishes global context and entities; the two track tasks then run independently; the shot task receives the prior context; and the commercial task receives only the assembled prior context. Thus, a system receives a sequence of task inputs rather than one unconstrained whole-video prompt. Table 8 gives the shared task-level media profile used to construct these inputs.

For screen and overlay extraction, the video is divided into 20-s core windows with 2-s context on each side. The profile caps in Table 8 do not imply a single frames-per-video or resolution value:

Table 8: Shared task-level media profile for all comparison runs. Frame and pixel quantities are requested upper bounds, not fixed decoded counts or a forced output resolution. A dash denotes a context-only task.

Task	Media scope	fps	Max. frames	Max. pixels per frame	Audio
`global_entities`	Full video	1.0	128	100,352	yes
`speech_audio`	Full video	0.5	64	50,176	yes
`screen_overlay`	20-s core window with 2-s padding	2.0	128	401,408	no
`shots`	Each shot interval	4.0	128	200,704	yes
`commercial`	Prior-task context only	–	–	–	no

Table 9: Native handling of the shared media profile. "Audio" denotes whether the model is eligible for the three audio-enabled tasks in Table 8.

Method family	Native media handling	Audio
Qwen2.5/3-Omni and Ours	Use the Qwen-Omni worker with the shared fps, frame, and pixel fields; its `min_pixels` setting is bounded by 100,352.	yes
Qwen3-VL and Qwen3.5	Use the shared fps and frame cap, but impose a 131,072-pixel-per-frame floor; Qwen3.5 additionally disables its thinking mode.	no
Keye-VL-2.0	Passes the shared fps, frame, and pixel fields to its SGLang video preprocessor.	no
InternVL3.5	Uniformly samples native-resolution frames subject to the frame cap and a context-length-derived cap; it does not apply the pixel cap.	no
VideoLLaMA3	Uses fps and the frame cap, but not the pixel cap.	no
Valley3	Receives the video path through its vendor processor, which uses its default 1-fps setting rather than the shared fps, frame, or pixel caps.	yes

decoded frame counts vary with duration, task window, and the native processor. We record the effective media receipt for every task.

D.2 Native processors and modality policy

Each method uses its native processor, summarized in Table 9. In particular, effective frame count and resolution remain model-dependent. This distinction is essential when interpreting cross-family results.

Audio is enabled only for Qwen2.5-Omni, Qwen3-Omni, Valley3, and Qwen-SAVER, and only for `global_entities`, `speech_audio`, and `shots`. The five vision-only systems skip `speech_audio`; their speech and audio tracks are held empty by the protocol rather than inferred from vision. Valley3 receives a 16-kHz mono PCM waveform, whereas the other audio-capable processors use their native video/audio handlers. We therefore mark audio-dependent results N/A for vision-only systems and do not treat them as directly comparable to audio-capable systems.

D.3 Prompt, decoding, parsing, and retries

All task calls use a fixed JSON-only system role and a task-specific schema derived from the Pydantic model. The user envelope supplies the task, video duration, window and prior context, and output schema. Instructions require explicit nulls and empty lists, source-grounded observations, existing entity/evidence IDs, and seconds relative to the clip start; window outputs are restricted to the core interval. The exact UTF-8 prompts and generated schemas are retained with the experiment configuration.

All models decode greedily. vLLM runs use temperature 0 and top-$p = 1$; Keye-VL uses temperature 0; and Transformer runs use `do_sample=False`. No run uses beam search, top-k, or a repetition penalty. The maximum number of new tokens is 8,192 for every task except `shots`, which is limited to 4,096; the nominal context limit is 32,768 tokens and is never allowed to exceed 65,536.

We preserve raw model responses. Parsing first accepts a strict JSON object. This is syntactic extraction, not semantic repair: we never normalize fields, enums, references, or timestamps. Truncation and schema nonconformance are also recorded but do not trigger a content repair. We allow one additional attempt only after a backend failure or empty response; format errors, schema nonconformance, and truncation do not receive a retry. A failed video is isolated from the remaining run; completed tasks are resumed only under the same run signature. The syntax-only fence extraction was applied consistently across systems, including a regenerated Keye-VL run.

D.4 Comparison interpretation

The Qwen2.5-Omni–Qwen-SAVER comparison holds the Qwen-Omni input pathway, modality policy, task contract, prompt, and decoding limits fixed, so it is a controlled system-level before–after comparison of the untuned and fine-tuned checkpoints. It does not isolate every design choice in the fine-tuned system. Cross-family rows additionally differ in native preprocessing and modality support (Table 9); they are useful reference points, but not a causal test attributing all score differences to "structured understanding."

E Qwen-SAVER training and inference details

Qwen-SAVER follows the five-task protocol in Section 4.1; scoring is specified in Appendix G.

E.1 Optimization and loss assignment

We adapt Qwen2.5-Omni-7B with LoRA for two epochs, using an effective batch size of 32 and a learning rate of 5×10^{-5}. All five component targets are optimized with the group-balanced objective in Equation 3. The objective in Equation 3 supervises all answer tokens and separately averages structural, ID, closed-set, and index-token groups. If no critical group occurs, that term defaults to the mean answer-token loss; $\lambda = 0.25$.

E.2 Autoregressive generation and assembly

At inference, global/entities precedes tracks and shots; their assembled context is passed to the commercial task. The five outputs are assembled before any script-level evaluation.

F Structured editing case-study protocol

For the two examples in Section 5.2, we select one shot from each source script. Each has an unedited baseline and three variants. Person and product variants replace the corresponding entity's name, role, and attributes with those from a same-type donor; the shot-local observed appearance is copied when available, but the original entity ID is retained. The shot variant copies the donor shot's `frame` and `scene` only. Edited scripts are schema-validated and their changes checked against these allowed fields; other shots, actions, and track records are not edited.

For *Ours*, a fixed template renders the selected shot's entities, frame, scene, subjects, and actions into an `integrated_multimodal_description`, with generic soundscape text. Global, commercial, and event-track contents are not rendered; no LLM rewrites the structured prompt. For *Dense*, an LLM rewrites a baseline natural-language caption using the raw donor facts, without receiving the edited structured prompt. Both methods submit their resulting text to the same distilled MiniMax H3 text-to-video generator, with no conditioning image or source video: seed 42, 512×896 pixels, 24 fps, 107 frames (about 4.46 s), and eight sampling steps. The two methods use identical generation settings, not identical prompt formats. These are illustrative re-generations of selected shots, not edits to the original video footage or a quantitative evaluation of editing success.

G Evaluation details

We distinguish schema conformance, reference-graph diagnostics, and content agreement (Section 4.2); content agreement comprises record correspondence and field fidelity. The content mea-

sures compare against pipeline-generated annotations, not independent video truth or verified advertising claims.

G.1 Evaluation scope

Schema conformance is a video-level check of required fields, types, and constrained values. It is an output-format measure, not a test of complete script validity or editability, and nonconformance does not exclude a prediction from content scoring. Separate *reference-graph diagnostics* pool applicable instances: *ID Valid* checks format and uniqueness, *Entity Ref.* checks entity resolution, and *Evidence Ref.* checks type-permitted evidence resolution. Empty applicable sets are N/A, not 100%. Even together, schema conformance and these graph diagnostics do not reproduce the construction-time full structural-integrity check $\mathrm{Struct}(s)$ in Equation 1, which also includes temporal and cross-shot constraints. Resolving a reference does not prove that it supports the associated interpretation.

G.2 Annotation correspondence before field scoring

One-to-one alignment avoids penalizing arbitrary local ID choices. Entities are matched semantically within type; shots are matched by time, then actions within matched shots. Events are aligned separately within each track. Temporal assignment prioritizes the number of pairs with tIoU at least 0.5, then total overlap; text only breaks ties. Actions in unmatched shots count as unmatched records. Commercial categories are matched after entity and evidence correspondences are available. A matched pair establishes correspondence, not field correctness: aligned speech can have an incorrect transcript, and aligned actions can have an incorrect predicate.

G.3 Annotation-correspondence aggregation

For each object or event type c, aligned pairs are true positives and unmatched predictions and annotations are false positives and false negatives. For video v, let P_{vc} and A_{vc} denote the predicted and annotated records, and M_{vc} their one-to-one matches. We compute the category-level micro-F1 by pooling counts over evaluable videos:

$$\begin{aligned} TP_{vc} = |M_{vc}|, \quad FP_{vc} = |P_{vc}| - |M_{vc}|, \quad FN_{vc} = |A_{vc}| - |M_{vc}|, \\ \mathrm{F1}_c^{\mathrm{micro}} = 100 \times \frac{2\sum_v TP_{vc}}{2\sum_v TP_{vc} + \sum_v FP_{vc} + \sum_v FN_{vc}}. \end{aligned} \tag{6}$$

The main-text summary uses the unweighted macro-average

$$\begin{aligned} \mathrm{F1}_{\mathrm{macro}}^{\mathrm{shared}} &= \frac{1}{|\mathcal{C}_{\mathrm{shared}}|} \sum_{c \in \mathcal{C}_{\mathrm{shared}}} \mathrm{F1}_c^{\mathrm{micro}}, \\ \mathcal{C}_{\mathrm{shared}} &= \{\text{product, action, claim, pain–solution,} \\ &\qquad \text{target, screen text, overlay}\}. \end{aligned} \tag{7}$$

The seven categories have equal weight; extra records remain false positives within each. Speech and audio are separate because vision-only systems do not generate those tracks. Entity is separate because Product is its subset. Table 10 retains every category, including weaker audio results.

Table 10: Category-level annotation-correspondence micro-F1 (%, higher is better). The shared macro-average excludes Entity, Speech, and Audio; Product is the product-entity subset. Audio-dependent results are N/A for vision-only models.

Method	Shared macro	Entity	Product	Action	Claim
Qwen3-Omni	42.4	24.4	27.1	2.3	62.8
Keye-VL-2.0	13.2	48.0	59.0	2.4	0.0
Qwen2.5-Omni	28.5	9.3	22.2	1.1	30.6
Qwen3-VL	33.4	37.5	47.2	1.1	46.7
Qwen3.5	24.2	26.7	39.0	2.7	10.9
InternVL3.5	7.9	17.6	32.8	0.0	5.2
VideoLLaMA3	9.9	14.2	25.9	0.0	23.3
Valley3	33.6	35.8	39.4	0.5	46.9
Qwen-SAVER (ours)	**58.7**	**59.2**	**85.4**	**43.5**	**71.2**

Method	Pain-Sol.	Target	Screen	Overlay	Speech	Audio
Qwen3-Omni	75.2	**81.9**	33.1	14.5	4.0	44.8
Keye-VL-2.0	0.0	0.0	24.2	6.8	N/A	N/A
Qwen2.5-Omni	64.5	66.0	11.4	3.8	0.5	0.2
Qwen3-VL	56.0	60.7	18.0	4.1	N/A	N/A
Qwen3.5	36.8	44.6	**33.8**	1.5	N/A	N/A
InternVL3.5	9.2	8.3	0.0	0.0	N/A	N/A
VideoLLaMA3	1.8	3.3	14.3	0.8	N/A	N/A
Valley3	72.6	71.0	2.0	2.9	8.8	**45.8**
Qwen-SAVER (ours)	**83.3**	80.1	31.9	**15.6**	**26.3**	17.8

G.4 FIELD-FIDELITY SCORING

Following Section 4.2, discrete fields use *matched-field agreement*, separately from record-level correspondence. Let $\mathcal{M}_{vf}$ contain the matched prediction–annotation parent pairs for which the annotation supports field f. For a scalar field, let $q_f(p, a)$ be one when the predicted value agrees with the annotation under the field-specific rule and zero otherwise. For a set-valued field, $q_f(p, a)$ is the fraction of annotation values recovered by one-to-one value matching. We pool supported matched-parent instances across videos:

$$\text{Agree}_f = 100 \times \frac{\sum_v \sum_{(p,a)\in\mathcal{M}_{vf}} q_f(p, a)}{\sum_v |\mathcal{M}_{vf}|}. \tag{8}$$

Missing predicted values receive zero. Unsupported annotation values, including empty reference sets, do not enter this denominator. A field with no supported matched-parent instance is N/A. Unmatched parents are penalized by record-level correspondence and are not counted again in matched-field agreement. This is not an end-to-end field micro-F1.

Closed-set values are consistent only after exact normalized equality. Semi-open values use a fixed Gemini-3.8-Flash judge to decide semantic consistency for each candidate pair; these are judge-mediated comparisons, not independent human-validated correctness. For entity- and evidence-reference fields, predicted identifiers are first transported through the established annotation correspondences and then compared with the entities or evidence records referenced by the annotation. These choices separate annotation correspondence from the fidelity of attributes and references.

Temporal fields are reported as video-macro tIoU over aligned records. Descriptive text fields use video-macro BERTScore-F1 with `bert-base-chinese`; verbatim fields such as speech, screen text, and raw claim text use video-macro exact-match accuracy. As in Table 3, text metrics average over reference-supported records, with missing predictions assigned zero before video-macro aggregation; they are not restricted to matched parents. Unsupported annotation fields are N/A rather than zero. In this test set, `claims.normalized_value` and screen-text bounding boxes lack annotation support and are therefore not reported. We do not average the heterogeneous field families into an overall score.

In Table 3, *Form* denotes the product entity's `form_factor`, *Predicate* denotes an action's `predicate`, and *Claim Type* denotes `claim_type`. Their matched-field agreement values are reproduced in Tables 12 and 13.

Tables 11–15 retain all reported field families, including low audio, text, and evidence scores. These conditional or reference-supported metrics must be read alongside record correspondence; none is a direct video-truth assessment. In Table 14, the 100% subject-entity reference row is diagnostic because subject matching inherits entity correspondences, not independent evidence of perfect identity linking.

Table 11: Temporal localization: video-macro tIoU over aligned action and event records (%, higher is better). N/A denotes a track not generated by a vision-only model.

Method	Action	Speech	Screen text	Overlay	Audio
Qwen3-Omni	2.7	2.3	26.2	8.1	43.6
Keye-VL-2.0	3.2	N/A	20.6	3.3	N/A
Qwen2.5-Omni	1.7	0.3	7.0	6.5	0.2
Qwen3-VL	2.1	N/A	15.6	2.0	N/A
Qwen3.5	4.3	N/A	22.7	0.6	N/A
InternVL3.5	0.0	N/A	0.0	0.0	N/A
VideoLLaMA3	0.0	N/A	8.4	0.2	N/A
Valley3	1.4	3.5	2.6	2.2	**47.7**
Qwen-SAVER (ours)	**43.0**	**22.6**	**30.2**	**18.8**	22.1

Table 12: Closed-set matched-field agreement (%, higher is better), using Equation 8. Q2.5-O: Qwen2.5-Omni; Q3-O: Qwen3-Omni; IVL: InternVL3.5; VL3: VideoLLaMA3. N/A denotes an audio field excluded for a vision-only model, or no supported matched-parent instance because no parents match or their annotated field values are null.

Field	Ours	Q2.5-O	Q3-VL	Q3.5	Q3-O	Keye	IVL	VL3	Valley
Global language	98.6	58.4	91.2	55.8	92.6	93.0	77.0	58.8	88.2
Person gender	93.9	97.1	95.6	84.9	96.2	96.7	94.4	96.9	95.4
Product form factor	69.0	45.5	58.1	52.5	65.4	63.0	44.1	47.2	47.1
Product primary color	64.6	68.0	69.3	56.7	76.3	68.8	61.4	44.6	73.1
Shot mixed-media type	26.2	0.0	2.2	20.2	26.0	25.5	N/A	0.0	3.0
Shot scale	70.1	34.0	53.2	58.7	59.9	56.3	N/A	37.8	45.5
Shot camera motion	66.6	63.5	63.9	64.4	64.6	64.4	N/A	37.8	63.9
Shot viewpoint	81.9	67.4	68.6	79.1	78.1	74.2	N/A	29.7	73.3
Shot time of day	92.6	29.6	68.3	59.1	72.9	93.1	N/A	8.3	86.5
Shot lighting	84.1	27.0	41.2	59.3	68.7	69.1	N/A	8.1	45.0
Shot visual density	78.3	62.9	70.7	74.0	72.8	73.5	N/A	5.4	71.2
Subject kind	95.2	N/A	83.3	44.8	70.0	100.0	N/A	N/A	33.3
Subject prominence	79.7	87.7	76.6	77.4	83.9	80.6	N/A	100.0	73.0
Speech voice type	58.4	35.3	N/A	N/A	62.5	N/A	N/A	N/A	38.6
Audio type	96.5	100.0	N/A	N/A	99.7	N/A	N/A	N/A	91.6
Audio tempo	63.7	50.0	N/A	N/A	63.5	N/A	N/A	N/A	37.0

Table 13: Semi-open matched-field agreement (%, higher is better), using Equation 8. A field value receives credit only when the semantic-consistency rule accepts it. Set-valued fields report the mean fraction of annotation values recovered within supported matched parents. Abbreviations and N/A conventions follow Table 12.

Field	Ours	Q2.5-O	Q3-VL	Q3.5	Q3-O	Keye	IVL	VL3	Valley
Global ad format	79.6	2.0	3.0	6.4	4.2	8.4	0.8	0.4	5.2
Global narrative structure	65.8	1.0	6.6	8.4	17.8	28.1	8.8	1.2	4.4
Global visual style	79.6	2.5	11.6	30.9	27.2	20.0	5.5	1.7	10.4
Global audio style	67.9	1.0	N/A	N/A	5.0	N/A	N/A	N/A	2.9
Product role	98.4	8.0	61.6	74.6	47.0	89.5	20.5	6.0	6.8
Product category	68.1	18.7	15.3	15.7	20.9	12.2	23.6	19.0	22.0
Shot location	63.0	16.3	28.0	19.6	21.8	27.5	N/A	2.7	23.6
Shot stage	68.2	10.1	9.8	4.1	11.8	29.5	N/A	0.0	5.9
Shot functions	60.4	12.1	18.9	17.0	20.0	16.9	N/A	0.0	13.8
Action predicate	40.0	11.8	8.9	9.8	9.7	15.2	N/A	N/A	8.3
Speech tone	60.0	11.8	N/A	N/A	18.2	N/A	N/A	N/A	22.3
Overlay type	41.0	0.8	16.2	0.0	11.0	29.0	N/A	0.0	20.0
Screen-text role	73.8	57.0	32.9	9.1	34.2	53.3	N/A	32.4	19.0
CTA type	64.2	26.8	27.6	8.0	48.4	0.0	2.3	3.3	43.4
Claim type	61.6	10.5	24.2	36.2	25.5	N/A	10.1	3.3	20.2

Table 14: Entity- and evidence-reference matched-field agreement (%, higher is better), using Equation 8. This is a content agreement measure within matched parents, distinct from schema conformance and reference-graph validity. Predicted identifiers are scored after correspondence projection; reference sets use the fraction of annotation links recovered within each supported matched parent. N/A follows Table 12; empty annotated reference sets are unsupported even when parents match, whereas missing predicted links to supported references score zero. † is diagnostic only because subject matching inherits entity correspondences.

Field	Ours	Q2.5-O	Q3-VL	Q3.5	Q3-O	Keye	IVL	VL3	Valley
Action subject ref.	67.6	4.1	45.9	51.3	45.6	3.2	N/A	N/A	46.7
Action object ref.	56.4	3.2	32.0	27.8	17.1	2.2	N/A	N/A	36.4
Action instrument ref.	33.6	0.0	20.0	0.0	0.0	0.0	N/A	N/A	0.0
Subject entity ref.†	100.0	100.0	100.0	100.0	100.0	100.0	N/A	100.0	100.0
Speech speaker ref.	61.6	0.0	N/A	N/A	25.0	N/A	N/A	N/A	36.3
Speech entity mentions	49.1	0.0	N/A	N/A	19.6	N/A	N/A	N/A	21.3
Screen-text entity ref.	33.5	4.3	19.9	12.5	8.5	0.1	N/A	3.9	6.6
Audio source entity ref.	31.6	N/A	N/A	N/A	60.0	N/A	N/A	N/A	33.3
Global promoted refs.	89.8	13.8	29.4	22.4	28.3	53.1	21.9	0.9	30.9
CTA promoted refs.	80.6	8.9	22.1	15.7	24.0	0.0	1.1	2.3	23.0
Claim entity ref.	88.7	9.6	46.7	32.6	26.8	N/A	37.7	13.3	37.8
Claim evidence refs.	15.3	0.7	7.1	14.8	8.4	N/A	0.0	4.4	1.7
CTA evidence refs.	17.8	0.6	3.1	5.3	7.2	0.0	0.0	0.1	0.9
Shot evidence refs.	27.6	1.2	4.7	7.7	9.4	1.1	N/A	0.0	2.1

Table 15: Selected text-field fidelity over reference-supported records (%, higher is better). Descriptive columns report video-macro BERTScore-F1; the last three columns report video-macro exact-match accuracy for verbatim text. Missing predictions receive zero. Audio-dependent results are N/A for vision-only models.

	Descriptive text: BERTScore-F1				
Method	CTA	Target	Pain	Solution	Benefit
Qwen3-Omni	85.9	61.1	67.4	66.7	60.0
Keye-VL-2.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-Omni	68.4	47.0	52.5	54.7	49.7
Qwen3-VL	56.2	38.3	41.9	42.4	34.7
Qwen3.5	34.2	30.0	21.2	21.2	20.5
InternVL3.5	3.7	3.0	3.1	3.3	1.6
VideoLLaMA3	13.7	1.2	0.5	0.5	0.3
Valley3	70.5	49.7	54.5	55.5	48.7
Qwen-SAVER (ours)	82.3	**66.1**	**68.6**	**69.0**	**61.5**

	Descriptive text: BERTScore-F1			Verbatim text: exact match		
Method	Product name	Scene	Action	Speech	Screen	Claim
Qwen3-Omni	34.5	52.0	1.9	0.3	12.6	16.9
Keye-VL-2.0	68.2	43.0	2.4	N/A	14.2	0.0
Qwen2.5-Omni	20.8	51.9	1.1	0.1	2.7	11.0
Qwen3-VL	49.6	44.0	1.5	N/A	9.8	14.0
Qwen3.5	28.1	51.1	3.2	N/A	1.1	0.5
InternVL3.5	32.7	0.0	0.0	N/A	0.0	0.2
VideoLLaMA3	19.0	0.1	0.0	N/A	2.8	4.4
Valley3	44.4	60.4	0.8	0.5	1.7	13.2
Qwen-SAVER (ours)	**75.4**	**72.7**	**32.9**	**6.9**	**18.2**	**28.0**

H VAO MEASUREMENT DETAILS

VAO measures corpus-level multimodal information density; it is neither a model-quality metric nor a validation of StructuredScript annotations. We reimplement the E-VAds procedure (Liu et al., 2026) and apply one shared configuration to SAVER and E-VAds. Table 1 reports the E-VAds values obtained with this reimplementation, providing a controlled cross-corpus comparison rather than a claim to reproduce unpublished original measurement code exactly.

H.1 METRIC DEFINITIONS AND IMPLEMENTATION

We sample one RGB frame per second and compute V_{den} as 100 times the mean cosine dissimilarity between each normalized DINOv3-Base CLS feature and its neighboring features. Neighbors within radius $d = 10$ are weighted by $\exp(-|i-j|/(2d))$; sensitivity settings use $d \in \{5, 15, 20, 25, 30\}$. Videos with fewer than two frames have undefined visual density and are excluded from visual aggregates. For duration T, $A_{\text{den}} = W(\text{ASR})/T$ and $O_{\text{den}} = \sum_i W(\text{OCR}(i))/T$. Audio is decoded at 16 kHz mono and transcribed with Whisper-large-v3; OCR transcribes legible text in each sampled frame. W uses `jieba` segmentation, counting Chinese segments and contiguous Latin/alphanumeric strings as words. Repeated text contributes at every visible frame: OCR density measures text exposure rather than unique phrases. We use `facebook/dinov3-vitb16-pretrain-lvd1689m`, `openai/whisper-large-v3`, and temperature-zero OCR requests.