OACAP: CAPTION-GUIDED ACTIVE OBSERVATION FOR DETAILED VIDEO CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlike question answering, where a query specifies a perceptual target, open-ended detailed captioning must determine what evidence to acquire and how to integrate it into a coherent global description. We propose **Observation-Adaptive Video Captioner (OACap)**, a caption-centered agent that turns unsupported or missing information into targeted observation requests, uses spatiotemporal tools to acquire evidence, and iteratively verifies and revises the caption. OACap captions consistently improve performance across general video-understanding benchmarks using a fixed text-only reader. To better diagnose the information they preserve, we introduce **Factual-Semantic Captioning Bench (FSCap-Bench)**. Its *Core–Support Factuality Score (CSFS)* combines claim precision with importance-aware fact coverage and responds as expected to four controlled perturbation types on 93.0–98.3% of valid caption pairs. Its *Cross-Modal Semantic Round-Trip Score (CSRS)* measures semantic preservation through a video–caption–video–caption round trip by matching atomic facts from source and recovered captions. Together, they connect downstream textual-memory utility with factual and semantic preservation.

1 INTRODUCTION

Detailed video captions must preserve entities, attributes, state changes, and temporal relations, yet a single global pass often misses small objects, brief actions, and distant transitions (Chen et al., 2024; Tang et al., 2025). Unlike VideoQA, open-ended captioning provides no query that specifies what evidence to seek. Text-only refinement cannot recover unseen facts, while existing tool-augmented video methods typically condition additional observation on an external query or reasoning state (Wang et al., 2024; Jeoung et al., 2024; Wang et al., 2025).

We propose **O**bservation-**A**daptive Video **Cap**tioner (OACap), illustrated on the left of fig. 1, which treats its evolving caption as a testable account of the video. Comparing that caption with provenance-aware evidence exposes omissions, unsupported claims, conflicts, and unresolved temporal relations. OACap distinguishes *revision gaps*, resolvable from existing evidence, from *observation gaps*, which trigger a targeted tool call. It may crop a small object, densely sample an unclear action, apply OCR to transient text, or retrieve external evidence for entity identification. The returned evidence supports the next whole-caption revision, which in turn determines what to inspect next (fig. 2).

We evaluate captions as question-agnostic textual memory: a fixed text-only reader answers public video-understanding questions using captions generated without access to those questions. To diagnose quality beyond task accuracy, we also introduce **FSCap-Bench**, summarized on the right of fig. 1. Its Core–Support Factuality Score (CSFS) combines claim precision with importance-aware coverage, while the Cross-Modal Semantic Round-Trip Score (CSRS) measures how much caption-expressed visual semantics remain recoverable through a video–caption–video–caption round trip. Controlled perturbations test sensitivity to correct additions, hallucinations, and omissions.

Our contributions are:

1. We formulate detailed video captioning as **caption-guided active observation** and introduce OACap, which separates revision gaps resolvable from existing evidence from observation gaps requiring targeted spatial–temporal inspection.

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.

2. We introduce **FSCap-Bench**: 498 videos from five public benchmarks, 996 verified observation challenges, and human-verified Core and Support facts. CSFS measures factual precision and importance-aware coverage; CSRS probes recoverability through a fixed cross-modal cycle.
3. On Video-MME, ActivityNet-QA, and VCapsBench, OACap consistently improves caption-based understanding across five matched backbones. Controlled additions, hallucinations, and omissions also show that CSFS responds correctly on at least 93% of examples in every condition.

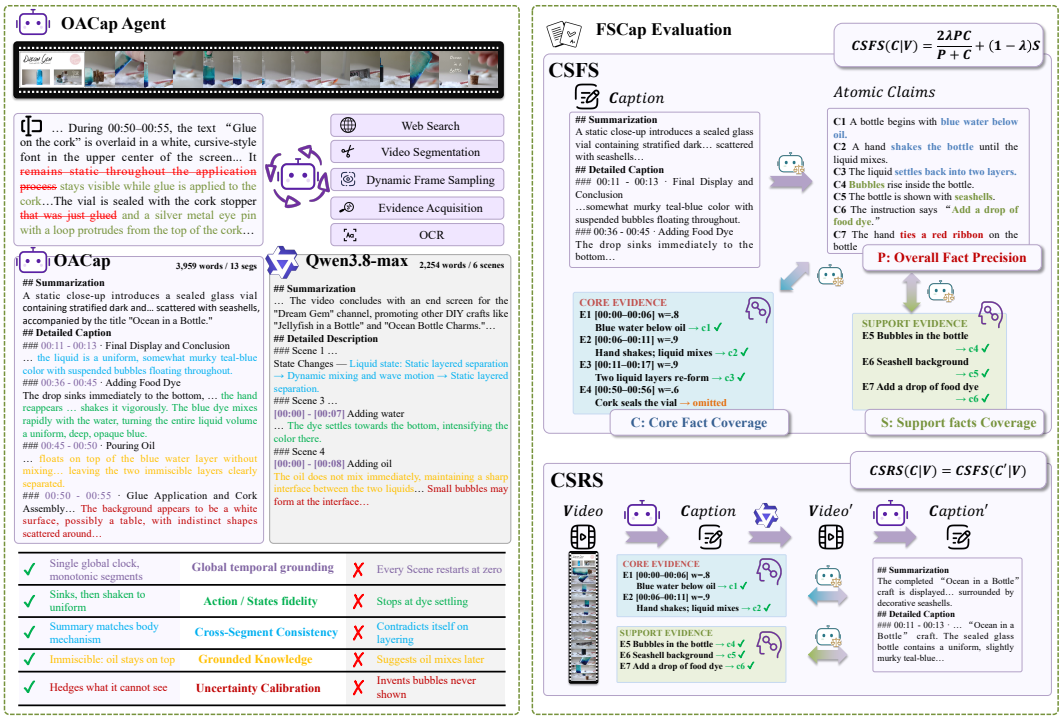


Figure 1: **OACap and FSCap-Bench at a glance.** *Left:* OACap is a caption-centric video captioner that invokes tools to continually self-verify and revise its evolving caption; fig. 2 presents the complete caption-guided observation–revision workflow. *Right:* FSCap-Bench evaluates atomic claims against human-verified Core and Support facts. CSFS measures claim precision and importance-aware fact coverage, while CSRS measures visual-semantic recoverability through a fixed video–caption–video–caption round trip.

2 Related Work

Detailed video captioning. Dense video captioning localizes and describes individual events, with models such as Vid2Seq jointly predicting event boundaries and captions from narrated-video pre-training (Krishna et al., 2017; Yang et al., 2023a). Recent work improves detailed captions through large-scale supervision, tool-assisted data generation, or model self-improvement: ShareGPT4Video uses differential captioning (Chen et al., 2024), Omni-Captioner trains on detailed captions produced by a tool-assisted pipeline (Ma et al., 2026), and VDC-Agent distills generation and self-reflection trajectories into preference data (Wang et al., 2026). OACap is complementary: at inference time, it uses gaps in the evolving caption to decide what additional evidence to acquire and how to revise the whole-video description.

Agentic visual understanding. Visual agents combine language models with vision experts, executable programs, or adaptive visual operations (Yang et al., 2023b; Surís et al., 2023; Su et al., 2025; Wang et al., 2024; Jeoung et al., 2024; Wang et al., 2025). VideoGAIA evaluates multi-turn assistants that iteratively perceive video, invoke tools, and integrate multimodal evidence (Zhang et al., 2026a). Such observation is typically controlled by an external query or task

state. OACap instead uses the evolving caption itself: its omissions, ambiguities, and unsupported claims determine what to inspect next.

Factual and semantic caption evaluation. Caption metrics range from reference matching to learned image–text similarity such as CLIPScore (Hessel et al., 2021), but do not directly verify every claim in a long description. CHAIR shows that conventional scores may not penalize object hallucination (Rohrbach et al., 2018). FactVC and ARGUS evaluate video-caption factual errors, with ARGUS also measuring omissions (Liu & Wan, 2023; Rawal et al., 2025); VDC, LongCaption-Bench, and VCapsBench assess detailed-caption quality or coverage (Chai et al., 2025; Wei et al., 2025; Zhang et al., 2026b). Correctness alone can favor short captions, whereas unweighted coverage can reward verbosity and equate central events with incidental details. CSFS instead combines claim precision with importance-aware Core and Support coverage. Cycle consistency has served as an information-preserving caption-training objective (Hagiwara et al., 2019), a reconstruction-based caption metric (Cui et al., 2025), and an image–text alignment signal (Bahng et al., 2025). CSRS extends this principle to detailed video captions by measuring fact recovery through a fixed video–caption–video–caption cycle.

3 OACap: Caption-Guided Active Observation

3.1 Agent Formulation

Given video $V = \{v_t\}_{t=0}^{T}$, OACap produces a faithful whole-video caption without an external query. Its evolving caption supplies the observation objective: missing, unsupported, or ambiguous content becomes an evidence request. With tool set $\mathcal{Z}$,

$$C = \mathrm{Compose}(\mathcal{E}_N), \quad a_n = \pi(S_n), \quad S_{n+1} = \mathrm{Update}(S_n, \mathrm{Observe}(a_n, V)), \tag{1}$$

where S_n is the state, a_n the selected action, and $\mathcal{E}_N$ the verified evidence at termination.

Unlike query-driven agents, the policy cannot optimize evidence for a known answer. It must balance breadth—covering entities, events, and relations across the full video—against verification of details already proposed by the draft. The caption provides this shared interface: its claims can be checked for support, while comparison with coarse visual context exposes content that has not yet been expressed. Consequently, observation and generation are interleaved rather than executed as independent stages.

Architecture overview. As shown in fig. 2, one iteration couples caption-guided acquisition with evidence-guided revision. Claim verification identifies statements that lack support, while coverage inspection uses the coarse video context to identify missing or underspecified content. The resulting gap is routed to a spatial, temporal, textual, or externally grounded observation. The returned evidence determines which claims to retain, remove, or rewrite, making the updated caption both the output of the current iteration and the control signal for the next one.

3.2 Evidence State and Gap Diagnosis

An inexpensive global pass yields coarse events, candidate temporal regions, and draft C_0. Each observation is normalized as

$$e_j = (f_j, \tau_j, m_j, p_j, r_j, s_j), \tag{2}$$

recording its fact, interval, modality, source, confidence, and verification status (candidate, supported, conflicting, or insufficient). The state

$$S_n = (\mathcal{E}_n, C_n, \mathcal{G}_n, \mathcal{H}_n, B_n) \tag{3}$$

contains evidence, the current draft, unresolved gaps, action history, and budget.

OACap treats C_n as a testable hypothesis and identifies

$$\mathcal{G}_n = \mathcal{G}_n^{\mathrm{miss}} \cup \mathcal{G}_n^{\mathrm{unsupported}} \cup \mathcal{G}_n^{\mathrm{conflict}} \cup \mathcal{G}_n^{\mathrm{temporal}}. \tag{4}$$

Coverage inspection compares the draft with coarse context and accumulated evidence, allowing events absent from the text to become gaps. Each gap is expressed as a concrete evidence question. If existing evidence already resolves it, OACap edits the caption; otherwise it requests a discriminating observation.

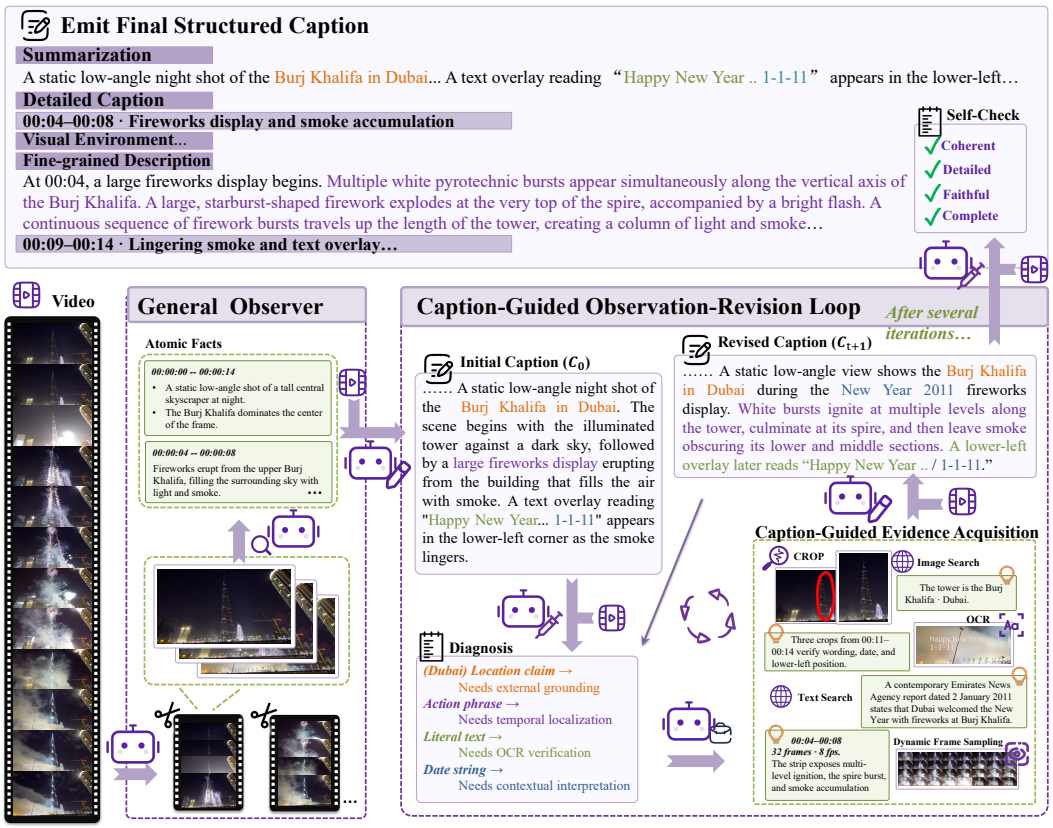


Figure 2: **Caption-guided observation and revision.** Caption gaps route targeted temporal sampling, cropping, OCR, or retrieval; returned evidence supports the next revision, which exposes the next gaps.

Revision or re-observation. This distinction separates information missing from the text from information missing from the evidence state. An omitted attribute already established by an observation calls only for revision. If the available evidence cannot distinguish plausible descriptions, the agent instead seeks an observation that resolves the uncertainty: a small-object attribute calls for spatial inspection, whereas uncertainty about whether an object was inserted or removed calls for an ordered temporal sequence. Thus every tool call is tied to a possible caption change rather than a generic request for more detail.

3.3 Observation-Adaptive Acquisition

For a prioritized gap, OACap selects

$$a_n = (z_n, \tau_n, \rho_n, q_n), \tag{5}$$

where z_n is the tool, τ_n the temporal range, ρ_n the spatial or temporal granularity, and q_n the evidence question. Keyframes and crops resolve static attributes and local layout; denser sampling or clip analysis resolves trajectories and event order; OCR grounds visible text. Image retrieval or search may ground an identity or context, but external evidence is stored separately and cannot establish video actions or override contradictory visual evidence. Benchmark references are never inference inputs.

Illustrative trace. In fig. 2, temporal inspection corrects the claim that the on-screen text remains static by showing that "Glue on the cork" appears during the application step. A crop reveals the looped metal pin on the cork, while OCR recovers the separate instruction "Add a drop of food dye." These observations resolve temporal, local-attribute, and textual gaps and support different edits. An inconclusive observation, by contrast, leaves its gap unresolved and cannot license a stronger caption claim.

3.4 EVIDENCE-GROUNDED REVISION

Returned observations are atomized, deduplicated, temporally aligned, and checked against $\mathcal{E}_n$. Agreement adds support; conflict can trigger another targeted action. OACap retains supported caption claims, corrects contradictions, and converts important unsupported claims into new gaps:

$$S_{n+1} = \mathrm{Update}(S_n, \mathrm{Normalize}(\mathrm{Observe}(a_n, V))) \,. \tag{6}$$

The loop stops when important content and temporal relations are resolved, the budget is exhausted, or expected gain is too small. The final caption is composed from verified evidence, retaining time ranges and tool traces as provenance for auditing; unresolved claims are hedged or omitted.

4 FSCAP-BENCH

FSCap-Bench evaluates factual comprehensiveness with CSFS and cross-modal recoverability with CSRS.

4.1 CONSTRUCTION

We curate videos from Video-MME, ActivityNet-QA, VCapsBench, Video-MMMU, and MVBench through four sequential filters: visual usability (F1), semantic density (F2), temporal coherence (F3), and caption-relevant evidence dependence (F4). F4 requires at least two verifiable challenges that benefit from targeted observation. Of $10{,}716$ candidates, 498 pass all filters, with two verified challenges retained per video (996 total). Figure 3 summarizes the filtering outcomes and final distribution; per-source retention, filter criteria, and audit details are provided in section A.

Annotators then construct time-aligned, importance-stratified Core and Support facts for each retained video. Automatic proposals are treated only as candidates: annotators verify them against the source video, add missing facts, and adjudicate ambiguous temporal boundaries and Core/Support assignments. Full annotation and quality-control procedures are described in section A.

4.2 CORE–SUPPORT FACTUALITY SCORE

The reference fact set is partitioned into weighted Core facts $\mathcal{E}_{\mathrm{core}}$ and complementary Support facts $\mathcal{E}_{\mathrm{sup}}$. A frozen judge atomizes caption C into claims $\mathcal{Q}(C)$. Let $s(q)$ indicate whether claim q is evidence-supported and $m(e, C)$ whether reference fact e is recoverable from C. Then

$$\begin{aligned} P &= \frac{\sum_{q \in \mathcal{Q}(C)} s(q)}{|\mathcal{Q}(C)|}, \\ R_{\mathrm{core}} &= \frac{\sum_{e \in \mathcal{E}_{\mathrm{core}}} w_e m(e, C)}{\sum_{e \in \mathcal{E}_{\mathrm{core}}} w_e}, \\ R_{\mathrm{sup}} &= \frac{\sum_{e \in \mathcal{E}_{\mathrm{sup}}} m(e, C)}{|\mathcal{E}_{\mathrm{sup}}|}. \end{aligned} \tag{7}$$

Core facts describe the entities, events, and state transitions needed to preserve the video's principal semantics, with importance weights assigned before candidate captions are evaluated. Support facts capture complementary attributes and contextual detail. "Support" here denotes a reference category; a supported caption claim may be justified by evidence from either category. Each evaluation case contains both types, so a caption cannot obtain the aggregate score through incidental detail alone.

CSFS combines precision and Core coverage harmonically, with bounded Support credit:

$$\mathrm{CSFS}(C) = \lambda \frac{2 P R_{\mathrm{core}}}{P + R_{\mathrm{core}}} + (1 - \lambda) R_{\mathrm{sup}}, \quad \lambda = 0.8. \tag{8}$$

The judge uses the same decomposition, prompts, and matching rules for every captioner. Contradicted and unsupported claims receive zero; semantically equivalent claims are deduplicated, each reference fact is counted at most once, and reverse matches cite the supporting caption span. We report all components and macro-average per-video scores. Unsupported claims reduce P; omissions

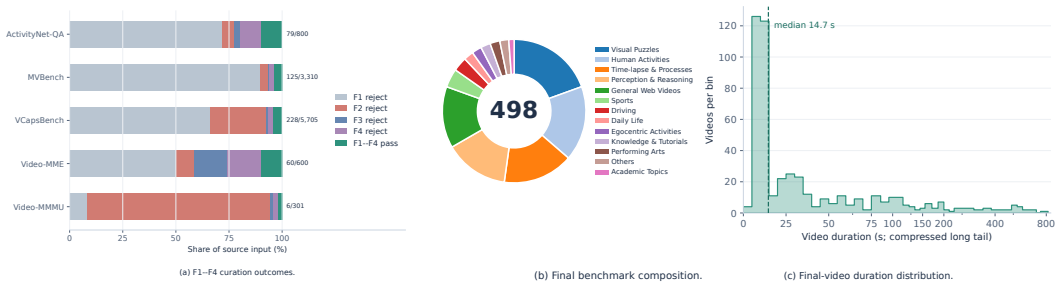


Figure 3: **FSCap-Bench curation and composition.** **(a)** Terminal filtering outcomes; **(b)** category composition of the final 498 videos; **(c)** duration distribution.

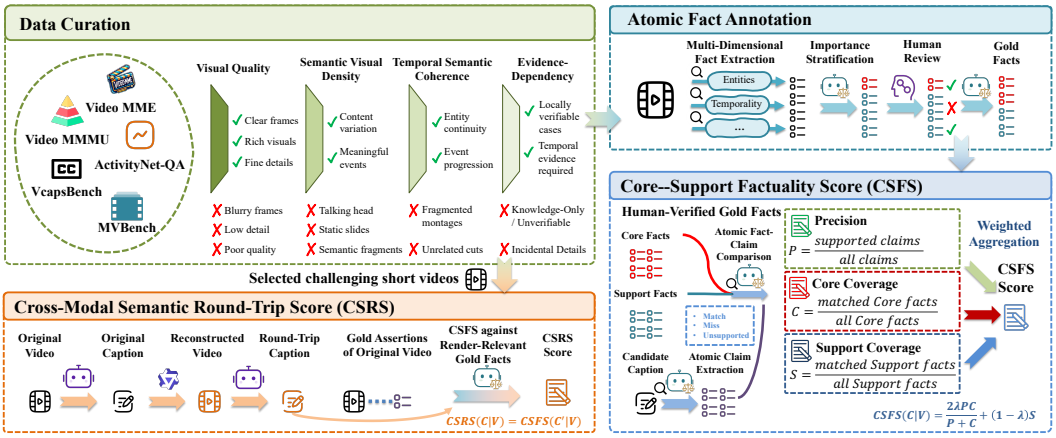


Figure 4: **FSCap-Bench.** Videos are curated and annotated as Core and Support facts. CSFS measures factual precision and importance-aware coverage; CSRS matches facts before and after $V \to C \to V' \to C'$.

reduce coverage, preventing either terse or unsupported captions from scoring highly. Contradiction, unsupported-claim, temporal-order, and global-consistency rates are retained as diagnostics but are not additional factors in CSFS.

4.3 Cross-Modal Semantic Round-Trip Score

Given V, a captioner produces C; a text-to-video model renders C as V', which the same captioner describes as C', forming $V \to C \to V' \to C'$. Using the same captioner at both ends avoids conflating reconstruction loss with a change in captioning model. A frozen atomizer exhaustively decomposes C and C' into self-contained, verifiable fact sets A and B. A second frozen LLM jointly inspects both lists and returns semantically equivalent (b, a) pairs, allowing paraphrases and differences in descriptive granularity while rejecting different objects, events, or attributes. We enforce one-to-one credit by retaining at most one returned pair for each fact in A and B; let M denote the resulting set. We also report counts $N_A = |A|$, $N_B = |B|$ and densities $D_A = 100|A|/W_C$, $D_B = 100|B|/W_{C'}$, where W denotes word count. We compute

$$R = \frac{|M|}{|A|}, \qquad P = \frac{|M|}{|B|}, \qquad \text{CSRS} = \frac{2RP}{R+P} = \frac{2|M|}{|A|+|B|}. \tag{9}$$

Here R is source-fact recovery, P is recovered-fact precision, Loss $= 1 - R$, and Drift $= 1 - P$ is the proportion of recovered facts absent from C. For empty sets, we use the explicit conventions $(R, P, \text{CSRS}) = (1, 1, 1)$ when both are empty, $(1, 0, 0)$ when only A is empty, and $(0, 1, 0)$ when only B is empty. We compute all scores per video before macro-averaging. CSRS measures C–C' agreement, not correctness against V: an unsupported fact repeated at both ends can still match, and Drift denotes cross-modal semantic change rather than visual hallucination. CSRS thus evaluates the complete channel and complements gold-grounded CSFS. Full generation, atomization, pairing, and matching configurations are given in section B.2.

5 EXPERIMENTS

We first test whether captions serve as reusable textual memory for downstream video understanding, then use FSCap-Bench to measure factual comprehensiveness and cross-modal recoverability. All caption-based comparisons use one whole-video caption, the same text-only reader, question prompt, output budget, and abstention policy. Observation cost is reported separately because equal caption length does not imply equal inference compute. Direct-video systems are different-input references rather than caption-only competitors or strict upper bounds.

5.1 CAPTION AS MEMORY

This experiment tests whether caption-guided observation preserves information that remains useful for questions unavailable during caption generation. Captioners do not see downstream questions. Every single-pass baseline receives the same whole-video captioning prompt, reproduced in section B.1. A fixed Qwen3.8-27B text-only reader receives only the caption and question, with no frames or tool traces. It must abstain when the caption is insufficient, and abstentions count as incorrect. This separation tests whether a description preserves reusable information rather than solving a known question during observation, although reader priors cannot be eliminated.

Table 1 compares seven single-pass models and five model-matched OACap variants. OACap improves every matched Video-MME split. With Qwen3.5-A3B, Short/Medium/Long accuracy rises from $70.22/62.00/57.33$ to $79.39/73.44/66.95$; with Qwen3.8-27B, it rises from $76.17/65.44/63.52$ to $79.84/73.36/66.56$. ActivityNet-QA gains are consistent ($+0.11$ to $+1.83$). On VCapsBench, every OACap variant improves AR and CR and reduces IR; OACap/Kimi-K2.6 achieves the best AR (71.37) and IR (19.57), while OACap/Qwen3.8-27B achieves the best CR (89.32).

The matched comparisons hold the reader and downstream protocol fixed, so the gains isolate the information made available by each captioning procedure rather than a stronger answer model. Improvements are largest on the longer Video-MME splits, where a single global pass is more likely to miss temporally separated evidence. The smaller but consistent ActivityNet-QA gains suggest that active observation still helps when questions emphasize dominant activities, while VCapsBench shows that the additional detail does not merely increase verbosity: answer and coverage rates rise as the irrelevant-information rate falls. These numbers follow our caption-as-memory protocol and are not interchangeable with official leaderboard scores that expose videos directly or use different prompting.

5.2 CSFS VALIDATION

This experiment tests whether CSFS rewards supported detail, penalizes unsupported detail, and distinguishes omissions of central facts from omissions of complementary facts. We perturb clean captions from 300 Video-MME short videos by adding approximately 20% correct detail, adding approximately 20% hallucinated content, deleting Core facts, or deleting Support facts, producing 1,500 captions. Three intended hallucination edits did not introduce unsupported content and are excluded from that paired condition ($n = 297$); all other conditions contain 300 pairs. The expected response is to reward verified information, penalize unsupported additions and omissions, and penalize Core omissions more strongly than Support omissions.

CSFS follows all intended directions (table 2): correct detail raises $0.9291 \rightarrow 0.9694$, hallucination lowers it to 0.8752, and Core/Support deletion lowers it to $0.8836/0.9039$. Its components respond selectively: correct additions raise Support coverage from 0.6677 to 0.8734; hallucinations reduce precision from 0.9987 to 0.8853; Core deletion reduces weighted Core coverage from 0.9919 to 0.8979; and Support deletion reduces Support coverage to 0.5384. At the video level, directional accuracy is 97.3%, 98.3%, 93.0%, and 93.3%, respectively—the only automatic method exceeding 93% in all four conditions.

Baselines fail for complementary reasons. The tested PAC-S, EMScore, and FactVC configurations truncate long captions to 77 tokens; similarity metrics weakly or inconsistently respond to fine-grained edits; and Qwen adaptations do not reliably distinguish correct additions from hallu-

Table 1: **Caption-as-memory results (%).** Video-MME and ActivityNet-QA report accuracy; VCapsBench reports answer recall (AR), inconsistency (IR), and coverage (CR). Direct-video overall scores use a different input condition. Bold marks the best comparable result.

Model / Method	Video-MME			ActivityNet-QA ↑	VCapsBench		
	Short ↑	Medium ↑	Long ↑		AR ↑	IR ↓	CR ↑
Qwen3.8-27B / Direct Video QA	86.72 (overall)			62.10	72.94 (overall)		
Qwen3.5-A3B / Single-pass Caption	70.22	62.00	57.33	49.04	61.63	22.52	79.55
Qwen3.5-27B / Single-pass Caption	78.33	69.89	65.11	49.22	66.16	20.84	83.57
Qwen3.5-122B / Single-pass Caption	76.03	67.89	64.99	49.07	64.93	21.48	82.69
Qwen3.8-27B / Single-pass Caption	76.17	65.44	63.52	48.86	69.21	21.51	88.19
Kimi-K2.6 / Single-pass Caption	78.01	71.67	62.33	49.65	70.35	19.79	87.71
Qwen3.5-A3B / OACap **(ours)**	79.39	73.44	66.95	49.15	69.78	19.79	87.00
Qwen3.5-27B / OACap **(ours)**	78.78	72.79	66.55	50.29	70.47	20.50	88.64
Qwen3.5-122B / OACap **(ours)**	78.33	72.74	**67.14**	50.11	70.13	19.93	87.59
Qwen3.8-27B / OACap **(ours)**	79.84	73.36	66.56	**50.69**	70.24	21.36	**89.32**
Kimi-K2.6 / OACap **(ours)**	**80.33**	**75.34**	66.67	50.68	**71.37**	**19.57**	88.73

cinations. These results validate directional sensitivity, but not human agreement or invariance to alternative atomic decompositions.

5.3 CSRS Validation

This experiment tests two complementary claims: whether OACap improves factual comprehensiveness under matched captioning backbones, and how much caption-expressed information survives a fixed cross-modal reconstruction channel. Table 3 compares direct captioning with OACap under matched backbones on FSCap-Bench, while table 4 evaluates recoverability through the fixed reconstruction channel. OACap improves CSFS for all three backbones, with gains of 1.94, 5.20, and 20.75 points for Qwen3.5-27B, Qwen3.5-122B, and Qwen3.8-27B, respectively.

For CSRS, we evaluate six configurations on the same 33 FSCap-Bench videos sourced from Video-MME, all shorter than two minutes. Qwen3.8-27B or Qwen3.8-Max produces C and is reused to caption V'. LTX-2.5-Diffusers renders both 10- and 60-second videos, while Wan2.2-T2V-A14B renders 10-second videos, giving 2×2 generator–captioner combinations at 10 seconds and two additional LTX configurations at 60 seconds. A frozen atomizer extracts source and recovered facts, and a separate frozen judge returns one-to-one semantic matches; section B.2 specifies the complete protocol.

At 10 seconds, LTX preserves more caption-expressed information than Wan for both captioners: CSRS improves from 0.206 to 0.245 with Qwen3.8-27B and from 0.200 to 0.231 with Qwen3.8-Max. Extending LTX to 60 seconds raises CSRS to 0.254 and 0.291, respectively. The captioners expose a recall–precision trade-off: Qwen3.8-Max generally recovers more source facts, whereas Qwen3.8-27B yields higher precision and lower drift. LTX-60s/Qwen3.8-Max achieves the best overall CSRS (0.291) and recovery (0.291); LTX-60s/Qwen3.8-27B achieves the best precision

Table 2: **Controlled caption perturbations** on 300 Video-MME short videos. Green arrows match the expected direction; red arrows do not.

Evaluator / configuration	Clean	Controlled perturbation			
		+Correct ↑	+Hallucination ↓	−Core ↓	−Support ↓
Exact-provenance D (Oracle)	*0.9485*	*0.9974* ↑	*0.8696* ↓	*0.8376* ↓	*0.9168* ↓
PAC-S	0.7744	0.7744	0.7748 ↑	0.7706 ↓	0.7756 ↑
EMScore	0.2736	0.2736	0.2736	0.2727 ↓	0.2743 ↑
FactVC	0.2261	0.2261	0.2261	0.2261	0.2266 ↑
BLEU-4	0.0268	0.0236 ↓	0.0226 ↓	0.0278 ↑	0.0283 ↑
CIDEr-D	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
BERTScore F1	0.7819	0.7816 ↓	0.7812 ↓	0.7816 ↓	0.7828 ↑
ARGUS-Qwen	0.5028	0.4978 ↓	0.5115 ↑	0.4607 ↓	0.4982 ↓
VDCScore-Qwen	4.9457	4.9668 ↑	4.9460 ↑	4.6820 ↓	4.7410 ↓
CapST-Qwen	0.7561	0.7398 ↓	0.7537 ↓	0.7386 ↓	0.7608 ↑
Direct Qwen Scene Judge	0.3539	0.3435 ↓	0.2788 ↓	0.3600 ↑	0.3583 ↑
CSFS (ours)	**0.9291**	**0.9694** ↑	**0.8752** ↓	**0.8836** ↓	**0.9039** ↓

Table 3: **Factual comprehensiveness on FSCap-Bench.** CSFS is reported on a 0–100 scale. Direct Caption uses a single whole-video caption generated by each backbone, while OACap Pipeline applies our caption-guided observation and revision framework with the same backbone. Δ is the absolute improvement in CSFS points and is computed from unrounded scores.

Backbone	Direct Caption	OACap Pipeline	$\Delta \uparrow$
Qwen3.5-27B	65.87	67.81	+1.94
Qwen3.5-122B	63.38	68.58	+5.20
Qwen3.8-27B	**66.50**	**87.24**	**+20.75**

Table 4: **Cross-modal round-trip results** on 33 FSCap-Bench videos sourced from Video-MME, all shorter than two minutes (per-video macro average). R: source-fact recovery; P: recovered-fact precision; CSRS: fact-level F_1; Loss$= 1 - R$; Drift$= 1 - P$. LTX is evaluated at 10 and 60 seconds; Wan is evaluated at 10 seconds.

T2V / Duration / Captioner	$R \uparrow$	$P \uparrow$	CSRS$\uparrow$	Loss$\downarrow$	Drift$\downarrow$
LTX-2.5 / 10s / Qwen3.8-27B	0.206	0.352	0.245	0.794	0.648
LTX-2.5 / 10s / Qwen3.8-Max	0.230	0.257	0.231	0.770	0.743
Wan2.2 / 10s / Qwen3.8-27B	0.168	0.296	0.206	0.832	0.704
Wan2.2 / 10s / Qwen3.8-Max	0.187	0.241	0.200	0.813	0.759
LTX-2.5 / 60s / Qwen3.8-27B	0.213	**0.362**	0.254	0.787	**0.638**
LTX-2.5 / 60s / Qwen3.8-Max	**0.291**	0.327	**0.291**	**0.709**	0.673

(0.362) and lowest drift (0.638). Absolute information loss remains high (70.9%–83.2%), reflecting the complete caption–generation–caption channel rather than the T2V model alone.

6 Limitations

OACap trades additional observation and model calls for caption quality. The present comparisons hold the downstream reader fixed but do not establish a quality–cost Pareto frontier against equal-compute uniform sampling or alternative multi-stage policies. The routing and stopping decisions also inherit the perceptual and calibration errors of the models that diagnose caption gaps and interpret tool outputs.

The evaluations have distinct scope limits. Caption-as-memory scores depend on reader priors and question coverage. CSFS depends on annotation completeness, importance assignments, atomic decomposition, and judge reliability; an omitted reference can make a correct claim appear unsupported. Because FSCap-Bench draws from datasets also used in the downstream evaluation, claims of independent generalization require a video-level overlap audit. CSRS combines errors from the generator, re-captioner, atomizer, and matcher, and the current 33-video subset covers only generation-compatible visual content. We therefore interpret downstream utility, CSFS, and CSRS jointly rather than as interchangeable estimates of caption quality.

7 Conclusion

OACap treats a detailed caption as both the output and the control state of video understanding. By separating gaps that can be repaired from existing evidence from gaps that require new observation, it turns caption deficiencies into targeted tool calls and evidence-grounded revisions. Across matched backbones, the resulting captions preserve more information for downstream question answering and achieve higher factual comprehensiveness on FSCap-Bench. CSFS and CSRS provide complementary views of this improvement: the former measures supported claims and importance-aware coverage against video-grounded facts, whereas the latter measures information recoverability through a fixed cross-modal channel. Together, the method and evaluation framework support a broader view of captioning as active evidence acquisition rather than one-pass description generation.

REFERENCES

Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle Consistency as Reward: Learning Image-Text Alignment without Human Preferences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22934–22946, 2025.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. OpenReview ID: tTDUrseRRU.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. *arXiv preprint arXiv:2406.04325*, 2024.

Tianyu Cui, Jinbin Bai, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ye Shi. Evaluating Image Caption via Cycle-consistent Text-to-Image Generation. *arXiv preprint arXiv:2501.03567*, 2025.

Keisuke Hagiwara, Yusuke Mukuta, and Tatsuya Harada. End-to-End Learning Using Cycle Consistency for Image-to-Caption Transformations. *arXiv preprint arXiv:1903.10118*, 2019.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive Video Understanding Agent: Enhancing Efficiency with Dynamic Frame Sampling and Feedback-Driven Reasoning. NeurIPS 2024 Workshop on Multimodal Algorithmic Reasoning (MAR), 2024. Oral presentation; non-archival workshop.

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)*, 2017.

Hui Liu and Xiaojun Wan. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11807–11823, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.723.

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. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. OpenReview ID: Z091XLyVkJ.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. ARGUS: Hallucination and Omission Evaluation in Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20280–20290. IEEE, October 2025. doi: 10.1109/ICCV51701.2025.01886.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023. doi: 10.1109/ICCV51070.2023.01092.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Captioning-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*, 2025.

Qiang Wang, Xinyuan Gao, Yuhang He, Jizhou Han, Jiangyang Li, Songlin Dong, Zhiheng Ma, and Yihong Gong. VDC-Agent: When Video Detailed Captioners Evolve Themselves via Agentic Self-Reflection. In *European Conference on Computer Vision (ECCV)*, pp. 515–535, 2026.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form Video Understanding with Large Language Model as Agent. *arXiv preprint arXiv:2403.10517*, 2024.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. *arXiv preprint arXiv:2512.05774*, 2025.

Hongchen Wei, Zhihong Tan, Yaosi Hu, Chang Wen Chen, and Zhenzhong Chen. LongCaptioning: Unlocking the Power of Long Video Caption Generation in Large Multimodal Models. *arXiv preprint arXiv:2502.15393*, 2025.

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, 2023a.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action. *arXiv preprint arXiv:2303.11381*, 2023b.

Fan Zhang, Guangming Yao, Jinyang Wu, Hao Wu, Zheng Lian, Xinyu Geng, Jingdong Chen, Yi Yuan, and Pheng-Ann Heng. VideoGAIA: A Benchmark for General AI Assistants on Agentic Video Understanding. *arXiv preprint arXiv:2608.14718*, 2026a.

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. VCaps-Bench: A Large-scale Fine-grained Benchmark for Video Caption Quality Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12726–12734, 2026b. doi: 10.1609/aaai.v40i15.38269.

A Curation Protocol and Coverage

A.1 Benchmark Composition

Data provenance. The curation manifest contains 10,716 unique video IDs with item-level media metadata, stage reachability, measured gate outputs, terminal-stage decisions, and F4 verification records. Durations span 2.08–2,029.63 seconds. All counts in fig. 3 are computed from this manifest rather than transcribed aggregates. Terminal stages are mutually exclusive, while diagnostic counterfactual readings from unreached stages are retained separately and never counted as funnel outcomes.

Overall, 4.6% of candidates survive all four filters. Retention is $60/600$ (10.0%) for Video-MME, $79/800$ (9.9%) for ActivityNet-QA, $228/5{,}705$ (4.0%) for VCapsBench, $125/3{,}310$ (3.8%) for MVBench, and $6/301$ (2.0%) for Video-MMMU. After merging source-specific labels, the largest groups are Visual Puzzles, Human Activities, Time-lapse & Processes, Perception & Reasoning, and General Web Videos. The set is intentionally short-video-skewed (median 14.7 seconds), but retains a long tail: $433/498$ videos are under two minutes. These statistics characterize the selected evaluation distribution rather than the prevalence of observation challenges in the source datasets.

A.2 Four-Stage Filtering

F1: visual usability. Twelve uniformly spaced midpoint frames provide grayscale sharpness, high-frequency detail, brightness, and invalid-frame statistics. Brightness-normalized Laplacian variance reduces dark-scene bias. F1 rejects a video if quality is below 0.60, blurred-frame fraction exceeds 0.30, minimum resolution is below 480p, black-frame fraction exceeds 0.15, or invalid-frame fraction exceeds 0.25. The quality score combines sharpness, detail, non-blurred fraction, bitrate per pixel per frame, and resolution. F1 is evaluated locally without VLM calls and rejects 7,645 videos, leaving 3,071.

F2: semantic visual content. Bounded local windows first remove hard static or image-like cases, rejecting 304 videos. For the remaining candidates, a VLM reads timestamped frames and must return keep$=1$, density≥ 20, events≥ 4, variety≥ 2, and no banned image-content type. The VLM gates reject another 1,668 videos, leaving 1,099. Less restrictive static, dialog, and content-type indicators remain diagnostic only. These exclusions define the benchmark scope and can remove visually informative instructional or screen content; low retention alone does not validate rejection precision.

F3: temporal coherence. A VLM reads timestamped frames and distinguishes trackable processes from unrelated montage. F3 requires event density≥ 45, montage noise≤ 0.5, causality≥ 3, and entity continuity≥ 2. Shifted-sampling probes are retained when present, and passing the measured funnel decision determines eligibility. F3 rejects 202 videos and sends 897 to F4. Strong-causality gating may exclude coherent events without an explicit causal chain, motivating separate rejection-set auditing.

F4 audit record. Each proposed challenge records a fact, evidence anchor, caption impact, required observation, and an optional plausible alternative. A separate verifier checks the indicated region or segment. Irrelevant decoration, unsupported intent, and external-knowledge-only difficulty do not qualify. The release gate requires at least two verified challenges per video and rejects knowledge-only cases; two verified challenges are retained for each passing video. Among the 897 candidates reaching F4, 399 are rejected and 498 pass, yielding 996 valid challenge records. State changes are the most frequent challenge type (364 videos), followed by local attributes (228) and on-screen text (163); types may overlap within a video. Processing failures remain unresolved rather than being counted as easy cases.

A.3 Fact Annotation and Quality Control

Annotators create time-aligned atomic facts spanning attributes, spatial relations, visible text, actions, transient motion, state changes and order, long-range relations, and audio-visual evidence. Automatic multi-tool proposals are treated only as candidates: annotators verify them against the

source video and add missing important facts. Facts are partitioned into importance-weighted Core facts and complementary Support facts. Ambiguous temporal boundaries and Core/Support assignments are adjudicated. Externally grounded knowledge requires a visible anchor and never substitutes for evidence that a video event occurred.

B Evaluation Prompts and Round-Trip Configuration

B.1 Single-Pass Direct-Caption Prompt

All single-pass captioners in table 1 receive the same prompt below. The prompt is question-agnostic and requests one temporally structured whole-video description.

```
Watch the entire supplied video and write a faithful, coherent, and
appropriately detailed English caption. The video is the only
observational source. Describe its beginning, intermediate phases,
and ending in chronological order.

Use this Markdown structure:

## Summary

A concise account of the main subjects, events, and outcome actually
visible.

## Detailed Caption

### MM:SS – MM:SS: Descriptive phase title

Connected prose describing this meaningful phase of the video.
Repeat phase blocks as needed to cover the full video.

Timestamp requirements:

– Use absolute video times rounded to whole seconds.
– Use HH:MM:SS for videos longer than one hour.
– Keep blocks chronological, within the actual video duration, and
  non–overlapping.
– Every block must have a positive duration: the end time must be
  later than the start time.
– Split phases when the setting, principal action, participants,
  topic, or object state changes meaningfully.
– Do not create a separate block for every camera cut when the same
  event continues.
– Brief events lasting less than one second may be described inside
  the surrounding phase. Do not create zero–duration blocks or
  falsely extend a brief event.
– Treat timestamps as useful approximate boundaries rather than
  fabricated frame–level precision.

Coverage and writing requirements:

– Cover the beginning, intermediate developments, and ending.
– Give greater detail to information–rich or rapidly changing
  portions.
– Preserve distinct stages even when they occur in the same setting.
– Mention repeated actions only when the repetition or count is
  meaningful.
– Avoid repeating the same description across adjacent blocks.
– Do not turn the caption into a shot–by–shot camera log unless
  camera movement affects understanding.
– Describe only what can be supported by the supplied video. Do not
```

infer unseen actions or guess unclear details.
– Do not include analysis, speculation, external knowledge, hidden reasoning, QA answers, or comments about the captioning process.
– Do not use tables, JSON, confidence scores, or additional top–level sections.
– Return only the requested Markdown caption.

B.2 CSRS Configuration and Fact Extraction

Configurations and pairing. All configurations share the same 33 FSCap-Bench videos sourced from Video-MME; every selected video is shorter than two minutes. At 10 seconds, we cross LTX-2.5-Diffusers and Wan2.2-T2V-A14B with Qwen3.8-27B and Qwen3.8-Max, producing four configurations. Because the evaluated Wan pipeline supports only the short setting, the 60-second evaluation contains the two LTX–captioner combinations, for six configurations and 198 source–recovered caption pairs in total. For each pair, a captioner produces source caption C from the original video, the T2V model renders C into V', and the *same* captioner produces C' from V'. The captioners use temperature 0.2, a 32,768-token output limit, disabled explicit thinking, and a common duration-aware frame prompt. Records are joined by video ID and verified through the stored SHA-256 hash of C, preventing accidental cross-captioner or cross-configuration pairing.

Atomic-fact extraction. Qwen3.8-Max serves as a fixed atomizer for both C and C', with temperature 0, a 16,384-token output limit, disabled explicit thinking, a 300-second timeout, and at most three attempts separated by two seconds. The prompt requests only a JSON object containing an exhaustive Chinese-language fact list. Every fact must be self-contained, concrete, observable, non-speculative, and limited to one assertion. Composite descriptions are split so that people and clothing, object attributes, colors, spatial relations, actions, temporal order, visible text, camera motion, lighting, environment, transitions, and audio cues can be matched independently. Quoted on-screen text is normalized to Chinese book-title brackets to reduce malformed JSON. We parse, in order, the complete response, a fenced JSON block, or the substring from the first opening to the last closing brace; a final regular-expression fallback recovers quoted items from the JSON fact array. Each distinct caption is processed once and cached by its SHA-256 hash.

Semantic matching. Kimi-K3-SPOT serves as the fixed matcher, with temperature 0, a 4,096-token output limit, disabled explicit thinking, a 180-second timeout, and the same three-attempt retry policy. It receives the complete indexed lists A and B and determines, for every recovered fact $b \in B$, whether a semantically equivalent source fact exists in A. Paraphrases, reordered wording, and a shared core claim at different descriptive granularity count as equivalent; statements about different objects, events, or attributes, and statements that are merely inferentially related, do not. The model returns aligned arrays of zero-based B and A indices. Each B fact may match at most one A fact. We additionally scan returned pairs in order and discard later pairs that reuse an already matched A fact, thereby enforcing one-to-one credit. Match outputs are cached using the first 16 hexadecimal characters of the source and recovered caption hashes.

Metrics and aggregation. For each video, $|M|$ is the number of retained pairs, $R = |M|/|A|$, $P = |M|/|B|$, and $\mathrm{CSRS} = 2RP/(R+P)$. We also report $\mathrm{Loss} = 1 - R$ and $\mathrm{Drift} = 1 - P$. If both lists are empty, $R = P = \mathrm{CSRS} = 1$; if only A is empty, $R = 1$, $P = 0$, and $\mathrm{CSRS} = 0$; if only B is empty, $R = 0$, $P = 1$, and $\mathrm{CSRS} = 0$. Fact density is the number of extracted facts per 100 words, counting each Chinese character as one word and whitespace-delimited English tokens as words. All reported configuration-level values are macro averages over the 33 per-video scores. The full pipeline yields 198 paired evaluations; fact extraction and semantic matching are cached separately to support incremental reruns without changing completed judgments.