Auditing the Auditor: A Reproducibility Study of Grounded Hallucination Auditing in Video Summarization

No Author Given

No Institute Given

Abstract. GroundXum audits hallucination in generative video summarization by extracting the entities a summary names and testing whether each is visually observable in the source video, using two image-text scorers and a vision-language model (VLM) as judge. Rather than proposing a new method, this paper asks whether the audit's conclusions are reproducible across four axes: run-to-run determinism, sensitivity to the grounding decision threshold, sensitivity to the number of sampled keyframes, and robustness of the VLM judge to prompt phrasing. All four studies run from the cached intermediate outputs of the original pipeline, requiring no new inference except the judge-prompt sweep. We find that the audit is fully deterministic on re-execution ($\kappa = 0.99$), that the relative ranking of the two audited summarizers is stable across the decision threshold (sign preserved across 51 of 60 tested values) and across all keyframe counts down to four frames, and that the judge is robust to prompt rewording and option reordering ($\kappa = 0.83$ to 0.85) but is materially altered by chain-of-thought prompting ($\kappa = 0.59$). We also deepen the original paper's observation that its single human annotator is permissive: a per-type breakdown across 800 annotated entities shows human-judge agreement of 61% to 92% on visually concrete entity types versus 24% to 54% on non-visual ones, localizing the divergence to an inference-based annotation criterion rather than to judge error. Every table and figure in this paper can be regenerated on a single workstation; the complete pipeline, cached intermediate outputs, and reproduction scripts are provided as anonymized supplementary material for review[1], and will be made publicly available upon acceptance.

Keywords: Reproducibility · Video summarization · Hallucination auditing · Visual grounding

1 Introduction

Generative video summarization compresses a video together with its transcript into a concise textual account of the events it depicts. Recent multimodal systems generate increasingly fluent summaries by fusing visual and linguistic signals through cross-modal attention and instruction-tuned large language models

[1] Code: anonymous.4open.science/r/GroundXum-065A. Large files (frame grids, checkpoints): Zenodo record (anonymized).

[12,2,7]. Fluency, however, does not entail faithfulness: a model may name entities that never appear in the video, or emit corrupted non-lexical tokens such as *verdalago* or *garala*, while still scoring well under the reference-based overlap metrics that dominate video summarization evaluation, ROUGE [9] and BLEU [13], because such metrics compare text against text and never consult the source video.

GroundXum [1] renders this failure mode auditable. Given a generated summary and its source video, GroundXum extracts the entities named in the summary, scores each one against sampled video frames using two pretrained vision-language scorers, and adjudicates visual presence with a VLM judge. Each entity receives a verdict of present, absent, or ambiguous, yielding in aggregate a video-grounded factuality score for each summarizer. Across two benchmarks and 53,036 entities, the original study reports that a multimodal summarizer grounds 8.5 percentage points more of its entities than a text-only baseline in the fine-grained cooking domain, and that fabricated non-lexical tokens ground more than 30 points less often than genuine entities, a direct hallucination signal that overlap metrics cannot detect.

A pipeline that chains an LLM extractor, two neural scorers, a VLM judge, and a calibrated decision threshold carries many implicit degrees of freedom, and the extent to which these affect the headline conclusions is precisely what a reproducibility study should establish. We pose four questions. Does re-executing the pipeline reproduce the same verdicts? Do the headline comparisons survive changes to the decision threshold and the keyframe count? Does the VLM verdict change when the judge prompt is reworded? And when the human cross-check disagrees with the judge, what accounts for the divergence?

Our contribution is a reproducibility analysis of GroundXum. We characterize the original study's conclusions along three axes and add one practical recommendation:

- **Robust conclusions.** We identify which findings reproduce unchanged on re-execution and survive perturbations to the decision threshold and keyframe count.
- **Relative versus absolute claims.** We distinguish conclusions that hold only as relative comparisons between summarizers from those that hold in absolute terms.
- **Annotation-dependent claims.** We isolate the findings that depend on the human annotation criterion, traceable through the cases where the human cross-check diverges from the VLM judge.
- **A caching design for re-checkable audits.** Because the original pipeline caches its per-frame grounding scores, three of our four studies require no GPU and execute in seconds as re-reductions of already-computed data; only the judge-prompt sweep invokes the VLM. We recommend this design to future audit-pipeline authors, as it renders the analysis re-checkable by a reviewer who lacks access to the source videos.

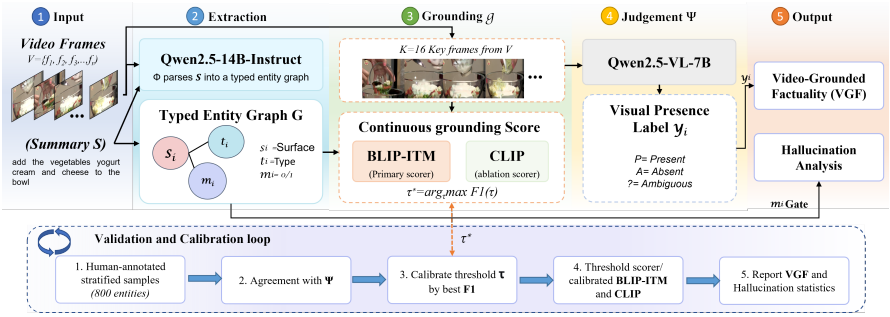

Fig. 1: The GroundXum audit pipeline, reproduced from [1]. The extraction operator Φ parses a generated summary S into a typed entity set $E(S)$, preserving malformed surface forms verbatim with flag m_i. The grounding operator g scores each entity against $K{=}16$ keyframes from V using BLIP-ITM (primary scorer) and CLIP (comparison). The judgment operator Ψ emits a visual-presence label $y_i \in \{\mathsf{P}, \mathsf{A}, ?\}$ over the keyframe composite. Video-grounded factuality $\mathrm{VGF}(S)$ is the fraction labeled present; m_i gates the hallucination audit. A human-annotated sample of 800 entities validates Ψ and calibrates the scorer threshold τ.

2 Related Work

Faithfulness analyses in text summarization establish that overlap metrics correlate poorly with factual correctness [11,5], and the visual domain exhibits an analogous failure mode: object hallucination was identified as systematic in image captioning [15], and multimodal LLMs hallucinate at decoding time [6,16]. Prior work audits against a reference text or a static image, whereas GroundXum audits against the source video itself, targeting the generative video summarization setting in which no visual ground truth exists for the generated text. Its grounding scorers build on BLIP [8] and CLIP [14], both known to exhibit bag-of-words behavior and to struggle with compositionality and counting [17], a limitation that bears directly on our finding that grounding reliability depends on entity type. The audit further relies on a VLM judge: such judges are now widely used [19], yet their reproducibility remains underexplored, with known failure modes including sensitivity to option ordering, prompt wording, and reasoning format [18], sensitivities our judge-prompt sweep measures directly for the judgment task in GroundXum.

3 The GroundXum Audit Pipeline

We summarize the audited pipeline only as far as our analysis requires, and refer the reader to [1] for the complete formalization. The pipeline composes three operators, illustrated in Fig. 1.

The **extraction operator** Φ maps a summary S to a typed entity set $E(S)$, in which each entity $e_i = (s_i, t_i, m_i)$ comprises a surface form s_i, a domain-specific type $t_i \in \mathcal{T}$, and a malformedness flag $m_i \in \{0, 1\}$ that marks corrupted non-lexical tokens without normalizing them. The type inventory is domain-specific: cooking video uses $\mathcal{T}_{\text{cook}} = \{\text{ingredient}, \text{tool}, \text{action}, \text{attribute}\}$, and open-domain video uses $\mathcal{T}_{\text{open}} = \{\text{person}, \text{object}, \text{setting}, \text{action}, \text{attribute}\}$.

The **grounding operator** g samples K=16 keyframes uniformly in time and scores each entity against them with a pretrained scorer ϕ, reducing the per-frame scores by a maximum:

$$g(e_i, V) = \max_{k \in \{1,\dots,K\}} \phi\big(e_i, f_{\pi(k)}\big), \quad \pi(k) = \left\lfloor \frac{(k-1)T}{K} \right\rfloor + 1. \tag{1}$$

Two scorers instantiate ϕ: BLIP-ITM, $\phi_{\text{ITM}}(e, f) = \sigma(h_{\text{ITM}}(e, f)) \in [0, 1]$, and CLIP, $\phi_{\text{CLIP}}(e, f) = \psi_{\text{txt}}(e)^\top \psi_{\text{img}}(f) / (\|\psi_{\text{txt}}(e)\| \|\psi_{\text{img}}(f)\|) \in [-1, 1]$. The maximum reduction reflects that an entity need appear in only a single frame to be visually grounded.

The **judgment operator** Ψ tiles the K keyframes into a single 4×4 composite and, conditioned on the entity surface form, emits a categorical label $y_i = \Psi(e_i, \hat{F}(V)) \in \{\mathsf{P}, \mathsf{A}, ?\}$. Video-grounded factuality aggregates these labels into a per-summarizer score,

$$\text{VGF}(S) = \frac{1}{|E(S)|} \sum_{i=1}^{|E(S)|} \mathbf{1}[y_i = \mathsf{P}], \tag{2}$$

while the continuous scorer output is reduced to a binary decision through a threshold τ, with $\hat{y}_i(\tau) = \mathsf{P}$ when $g(e_i, V) > \tau$ and A otherwise. The threshold is calibrated as $\tau^\star = \arg\max_\tau F_1(\tau)$ against the judge labels on a held-out set.

We instantiate the audit with Qwen2.5-14B-Instruct for Φ (JSON-constrained decoding, temperature 0), BLIP-ITM (blip-itm-base-coco) and CLIP (ViT-B/32) for ϕ, and Qwen2.5-VL-7B for Ψ (temperature 0, single-token verdict). Two summarizers are audited: MM, a multimodal generative summarizer that is the subject of separate work under review and is treated here as a black box, and TXT, a text-only baseline sharing the same encoder-decoder backbone and training data. Two datasets disentangle framework properties from domain properties: YouCook2 [20] (fine-grained cooking, 457 test videos) and VideoXum [10] (open-domain activities, 4,000 test videos), together spanning 53,036 extracted entities across 8,914 summaries. A stratified sample of 800 entities, 200 per dataset-summarizer cell, balanced across score buckets, furnishes the human cross-check used to calibrate $\tau^\star$ and validate Ψ.

4 Reproducibility Setup and Stability Analysis

Table 1 gives the complete environment specification required to reproduce the audit. The pipeline admits only a small set of potential nondeterminism sources,

Table 1: Environment and software dependencies for reproducing the GroundXum audit.

Component	Specification	Role
Hardware	1× NVIDIA RTX 4090 (24 GB VRAM)	Scorer and judge inference
Φ (extraction)	Qwen2.5-14B-Instruct via Ollama, T=0	Entity extraction
ϕ scorer 1	BLIP-ITM (`blip-itm-base-coco`, HuggingFace)	Primary grounding scorer
ϕ scorer 2	CLIP (`clip-vit-base-patch32`, HuggingFace)	Comparison scorer
Ψ (judge)	Qwen2.5-VL-7B via Ollama, T=0	Visual presence judgment
Keyframes K	16, evenly spaced; tiled into 4×4 grid	Input to Ψ
Score aggregation	Max over frames	Eq. 1
Threshold calibration	$\tau^\star = \arg\max_\tau F_1(\tau)$	Best-F1 against Ψ labels
Datasets	YouCook2 (457 videos), VideoXum (4,000 videos)	Benchmarks
Human annotation	800 entities, 200 per cell, stratified by bucket	Validation of Ψ

each eliminated by construction. Keyframe selection uses fixed, evenly spaced indices (Eq. 1), so a given video length always yields the same frames. The extraction LLM Φ and the judge VLM Ψ both decode greedily at temperature 0, and the BLIP-ITM and CLIP scorers run under `torch.inference_mode()` with no stochastic sampling. The pipeline is therefore deterministic by construction, which we verify empirically below before turning to its sensitivity to the pipeline's free parameters.

Throughout, we report raw agreement $A(x, y) = \frac{1}{|E|} \sum_e \mathbf{1}[x_e = y_e]$ alongside Cohen's kappa [4], $\kappa = (p_o - p_e)/(1 - p_e)$, with $p_o = A(x, y)$ and $p_e = \sum_\ell \hat{p}_x(\ell)\hat{p}_y(\ell)$ the agreement expected from the marginal label frequencies. Kappa is indispensable here because, as Section 5 shows, one label source is strongly skewed toward P; raw agreement alone would overstate concordance by crediting chance agreement on the majority label.

Determinism verification. To confirm that the cached intermediates faithfully represent a fresh run, we re-executed the judge Ψ from scratch on the YouCook2/MM validation set (n=200) and compared its verdicts against the cached verdicts on which the original paper's tables were built. The two runs agreed on 199 of the 200 entities (99.5%, $\kappa = 0.99$), the lone disagreement attributable to last-digit floating-point noise from nondeterministic cuDNN kernel ordering rather than to any modeling difference. The audit thus reproduces itself up to numerical tie-breaking at the decision boundary. We accordingly treat the cached intermediate outputs as a faithful stand-in for re-execution: this is precisely what renders the threshold and frame-count studies below, together with the human-comparison study of Section 5, inexpensive to reproduce, since each is a deterministic re-reduction of data already on disk.

Decision threshold. The continuous scorer ϕ becomes a present/absent decision through the threshold τ. The calibrated value $\tau^\star$ differs by scorer and domain (BLIP-ITM: 0.10 for YouCook2, 0.15 for VideoXum; CLIP: 0.20 for both with MM), so a natural reproducibility question is how sensitive the headline comparisons are to the choice of τ. We swept τ over 60 values spanning the full observed CLIP score range $[0.16, 0.35]$ on YouCook2 and recomputed the grounding rate $\mathrm{GR}(S) = \mathrm{VGF}(S)$ for each summarizer at every threshold. Ta-

Table 2: Decision-threshold stability. Gap is GR(MM) − GR(TXT) over a 60-point sweep of the CLIP threshold τ on YouCook2. The relative ranking is stable throughout the discriminative operating region.

Quantity	Value
Scorer / Dataset	CLIP / YouCook2
τ range swept	$[0.16, 0.35]$
Number of thresholds tested	60
Thresholds where sign is preserved	51 of 60
Gap at calibrated $\tau^\star$=0.20	+0.003
Maximum gap (mid-range, $\tau\approx$0.26)	+0.070
Gap range over full sweep	$[-0.001, +0.070]$
Sign non-positive only at	saturated ends

ble 2 summarizes the result. The gap GR(MM)−GR(TXT) keeps its positive sign across 51 of the 60 tested thresholds and ranges from −0.001 to +0.070 over the sweep. At the calibrated $\tau^\star = 0.20$ the gap is +0.003; it widens to its maximum of +0.070 in the discriminative mid-range near $\tau \approx 0.26$. The nine non-positive thresholds all sit where both grounding rates are saturated and the comparison degenerates: seven are negligible negatives (within −0.001) at low thresholds where both rates are near 1.0, and two are at the strict high end where nearly all entities are labeled absent and both rates approach zero. Within the discriminative operating region the ordering of MM over TXT is threshold-robust, while the absolute grounding rates change substantially across the sweep, which is why we treat relative comparisons as the reproducible quantity.

Score aggregation. The grounding operator reduces the K per-frame scores to one entity score via a maximum, encoding a presence semantics: an entity is grounded if it appears in any sampled frame. We tested three alternatives from the cached per-frame scores: mean, top-3 mean, and top-5 mean. Table 3 reports best-F_1 against the VLM judge and the grounding rate at the best-F_1 threshold for each. The max aggregation achieves the best or near-best F_1 in every setting (0.635 for MM, 0.549 for TXT); top-3 mean nominally edges it (0.637 for MM, 0.553 for TXT) within run-to-run noise, while mean aggregation under-performs by 1 to 2 F_1 points. The grounding rates in Table 3 are each taken at that system's own best-F_1 threshold and so are not directly comparable across systems; the F_1 values are. The MM > TXT F_1 ranking holds under every aggregation, confirming that the headline result does not depend on the max-pooling design.

Number of sampled keyframes. The grounding stage depends on how many frames are sampled. We approximate k-frame sampling using the cached 16-frame score arrays: we select k of the 16 cached frames at even stride (the same rule as Eq. 1), re-aggregate by max, and re-apply the decision threshold. This is a CPU-only approximation of re-running the scorer; because the 16 cached frames are themselves evenly spaced, the k selected here land near a true uniform-k

Table 3: Score aggregation sensitivity (CLIP, YouCook2). Best F_1 against the VLM judge and resulting grounding rate (GR) at the best-F_1 threshold for each aggregation. The MM > TXT F_1 ordering holds under all four.

Aggregation	Summarizer	Best F_1	Threshold $\tau^\star$	GR at $\tau^\star$
max (default)	MM	0.635	0.238	0.701
	TXT	0.549	0.234	0.709
mean	MM	0.619	0.197	0.841
	TXT	0.530	0.208	0.618
top-3 mean	MM	0.637	0.227	0.748
	TXT	0.553	0.226	0.699
top-5 mean	MM	0.633	0.223	0.726
	TXT	0.542	0.222	0.694

sample's positions, and we treat the agreement with a true re-sample as an upper bound on stability.

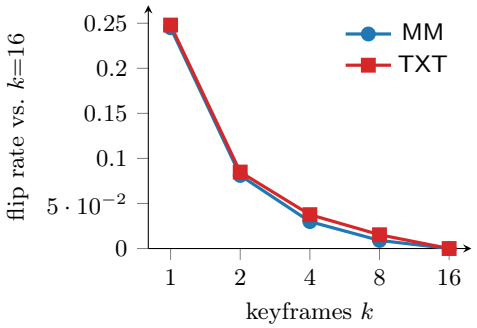

Fig. 2: Keyframe-count sensitivity (CLIP, YouCook2, τ=0.20). Fraction of per-entity verdicts that flip relative to k=16. Verdicts are stable down to four frames; only single-frame sampling perturbs them materially.

k	Flip rate		GR	
	MM	TXT	MM	TXT
1	0.245	0.248	0.740	0.735
2	0.081	0.085	0.904	0.898
4	0.030	0.038	0.955	0.945
8	0.009	0.015	0.976	0.968
16	0.000	0.000	0.985	0.983

Table 4: Per-k verdict flip rate and grounding rate (GR) at τ=0.20. GR is strongly k-dependent; the MM > TXT gap stays positive at every k.

Figure 2 and Table 4 show the results. The verdict flip rate relative to the full 16-frame reference falls below 2% at k=8 and below 4% at k=4, rising sharply only at k=1 (about 25% for both summarizers). The absolute grounding rate, however, is strongly frame-dependent: MM rises from 0.740 at k=1 to 0.985 at k=16. The lesson is the same as for the threshold: absolute grounding rates must not be compared across runs that used different keyframe counts, whereas the relative ranking of MM over TXT is stable at every tested k.

Judge-prompt robustness. The judgment operator Ψ is the component most exposed to subjective design choices: the prompt phrasing, the set of answer options, and the reasoning format can all vary without changing the task. We

Table 5: Judge-prompt pairwise Cohen's kappa on the YouCook2/MM validation set (n=200 for all pairs except v2 vs v3 where n=196). Variants: v0 original, v1 reword, v2 chain-of-thought, v3 forced binary, v4 option-order swap. The CoT variant (v2) is the outlier; all other pairs are stable.

	v0	v1	v2	v3	v4
v0	1.000	0.848	0.591	0.841	0.827
v1		1.000	0.554	0.733	0.793
v2			1.000	0.610	0.577
v3				1.000	0.677
v4					1.000

designed five prompt variants (reproduced verbatim in Appendix A) and re-ran Ψ on the same 200 grid images from the YouCook2/MM validation set. The variants are v0 (the original prompt), v1 (a reworded paraphrase with identical options), v2 (chain-of-thought reasoning before a final letter), v3 (a forced binary that removes the ambiguous option), and v4 (an option-order swap). Table 5 reports pairwise Cohen's kappa between all variant pairs. The judge is highly stable under rewording (v0 vs v1: $\kappa = 0.848$), forced binary (v0 vs v3: $\kappa = 0.841$), and option reordering (v0 vs v4: $\kappa = 0.827$). The only variant that substantially moves the judge is chain-of-thought: v0 vs v2 yields $\kappa = 0.591$, and v2 against every other variant sits in the range 0.554 to 0.610. We conclude that the audit is robust to incidental prompt wording and option presentation, but chain-of-thought prompting constitutes a meaningfully different reasoning mode that produces different verdicts. We flag the reasoning format as a design parameter that must be fixed and reported explicitly rather than left to the prompt author.

5 Human Cross-Check and Construct Validity

The original study validates Ψ against a single human annotator over 200 entities per cell (800 in total), reporting raw agreement of 49.0% and 46.0% on YouCook2 for MM and TXT, and 63.0% and 64.5% on VideoXum, with strongly asymmetric disagreements: the human labels P where the judge labels A, but almost never the reverse. On this basis the original paper adopts Ψ as the primary annotator. We reproduce these aggregate figures exactly and then trace their origin through a per-type breakdown, which reveals that the low agreement reflects a divergence in annotation criteria rather than judge error.

The permissiveness asymmetry. Table 6 pairs the aggregate figures with the human P rate in each cell. The annotator labels P between 86% and 95% of the time, so a trivial always-present predictor would agree with the human more often than the visually strict judge does. This is precisely why chance-corrected agreement is near zero overall ($\kappa = 0.06$ for YouCook2/MM): raw agreement above 50% is largely an artifact of the judge matching the human's high P base rate, not evidence that the two sources resolve individual cases the same

Table 6: Human-judge aggregate agreement reproduced from [1], augmented with the human P rate (H_P%) and Cohen's kappa, which clarify why raw agreement is misleading.

Dataset	Sys.	n	Agree%	H_P%	Both P	Both A
YouCook2	MM	200	49.0	95.4	90	8
	TXT	200	46.0	91.8	76	16
VideoXum	MM	200	63.0	88.4	104	22
	TXT	200	64.5	86.0	102	27

Table 7: Per-type human-judge agreement. Left: VideoXum, where both visual and non-visual types are represented. Right: YouCook2, whose schema is almost entirely non-visual ($n \leq 5$ for visual types, omitted). $H_P V_A$%: human-present but judge-absent.

(a) VideoXum.

Sys.	Type	Group	n	Agree%	κ
MM	person	visual	49	75.5	0.373
	object	visual	40	72.5	0.313
	setting	visual	23	82.6	0.000†
	action	non-visual	81	44.4	0.118
TXT	person	visual	49	91.8	0.626
	object	visual	36	61.1	0.248
	setting	visual	34	67.6	0.375
	action	non-visual	81	48.1	0.146

(b) YouCook2 (non-visual types).

Sys.	Type	n	Agree%	κ
MM	ingredient	159	53.5	0.085
	action	34	23.5	0.000
TXT	ingredient	139	52.5	0.145
	action	51	31.4	0.042

† No label variance: the human labeled all 23 setting entities P.

way. The asymmetry alone already indicates that the disagreement is structured rather than random.

Per-type breakdown. The aggregate masks a sharp and interpretable signal. Table 7 disaggregates the validation set by entity type. On visually concrete types, agreement is consistently high, person (75.5% to 91.8%), object (61.1% to 72.5%), and setting (67.6% to 82.6%). On the non-visual action type it falls to 44.4% to 48.1%, with κ between 0.118 and 0.146. YouCook2, whose schema is dominated by ingredient and action, exhibits the same pattern more starkly: agreement on ingredient is 52.5% to 53.5%, and on action it collapses to 23.5% to 31.4%. Grouping the types into visual and non-visual makes the dichotomy explicit (Table 8): on VideoXum, agreement is 75.6% to 76.1% for visual types against 46.0% to 48.1% for non-visual, and the share of entities the human labels P while the judge labels A is 23.0% to 23.5% for visual types but 51.9% to 54.0% for non-visual. The disagreement is thus concentrated entirely on entities that lack a direct visual referent.

Construct divergence. The pattern across Tables 7 and 8 is the signature of an inference-based annotation criterion. The human annotator judged whether

Table 8: Visual versus non-visual grouped agreement. $H_P V_A$%: human-present but judge-absent. YouCook2 visual cell ($n \leq 5$) omitted.

Dataset	Sys.	Group	n	Agree%	κ	$H_P V_A$%
VideoXum	MM	visual	113	76.1	0.344	23.0
	MM	non-visual	87	46.0	0.135	54.0
VideoXum	TXT	visual	119	75.6	0.413	23.5
	TXT	non-visual	81	48.1	0.146	51.9
YouCook2	MM	non-visual	195	47.7	0.055	51.8
	TXT	non-visual	196	45.4	0.100	54.6

an entity was plausibly part of the depicted activity, rather than literally visible in the keyframes: a named sauce was marked P once its constituent ingredients appeared on screen, and an action was marked P when implied by the recipe context rather than directly observed. The judgment operator Ψ, in contrast, is instructed to decide on visible evidence alone and applies that criterion uniformly to every entity. The two label sources therefore coincide exactly where their two notions of presence coincide, on physically salient on-screen objects, and part where the notions part, on seasonings, fine-grained actions, and abstract attributes. The divergence is systematic in entity type, which is the hallmark of differing constructs rather than of noise or judge failure.

We accordingly read the human pass as a characterization of annotator permissiveness rather than as visual ground truth, and we retain Ψ as the consistent reference for the audit. This interpretation is consistent with, and strengthens, the reading advanced in the original paper [1]. We do not claim Ψ is infallible: Section 6 notes that VLM judges inherit compositionality and counting weaknesses from pretraining. We claim the narrower and more defensible point that Ψ is the reproducible, internally consistent label source, and that its divergence from the human labels is explained by construct rather than by error.

6 Lessons Learned and Limitations

Artifact availability. Every numeric result in this paper is regenerated by a single script, mapped in Table 9. Because the pipeline caches its per-frame grounding scores, judge verdicts, and human labels, the threshold, aggregation, keyframe, and human cross-check analyses are pure CPU re-reductions of stored data and need neither the source videos nor a GPU; only the judge-prompt sweep and the determinism check invoke Qwen2.5-VL-7B [3]. The full pipeline, cached outputs, and scripts will be released on acceptance.

Four lessons follow for anyone reproducing or extending a grounded video-summary audit.

Cache per-frame scores. The most reproducibility-friendly design decision in GroundXum is to store the full per-frame score array for every entity rather than only the aggregated score. This is what makes the threshold, aggregation,

Table 9: Artifact-to-script mapping. Every numeric result is produced by one of these scripts from the repository root. CPU scripts complete in under 30 seconds; GPU scripts invoke Qwen2.5-VL-7B via Ollama and require the keyframe grids.

Result	Script (in `sweeps/` unless noted)	Compute
Table 2 (threshold stability)	`sweep_threshold.py -scorer clip -grid 60`	CPU
Table 3 (aggregation)	`sweep_aggregation.py -grid 60`	CPU
Fig. 2, Table 4 (keyframes)	`sweep_frames_cached.py -scorer clip -thresh 0.20 -ks 1,2,4,8,16`	CPU
Table 5 (judge-prompt)	`sweep_judge_prompt.py -variants v0,v1,v2,v3,v4`	GPU
Tables 6 to 8 (human cross-check)	`sweep_per_type.py -datasets youcook2,videoxum`	CPU
All aggregate numbers	`master_analysis.py`	CPU
Determinism check (Sec. 4)	`sweep_judge_prompt.py -variants v0`, compare to cached	GPU

and keyframe studies free to re-run, and what renders the audit re-checkable by a reviewer with neither the source videos nor a GPU.

Report relative comparisons, not absolute rates. Absolute grounding rates shift substantially with both the decision threshold and the keyframe count, whereas the relative ranking of summarizers is stable across both. Comparisons of absolute grounding rates drawn from runs with different settings are therefore not meaningful.

Treat the judge's reasoning mode as a fixed hyperparameter. Prompt wording and option ordering are incidental, with κ remaining above 0.82 across those variants; chain-of-thought reasoning is not, moving κ to 0.59, and must be fixed and reported explicitly.

Compare label distributions before concluding a judge is wrong. A permissive annotator who defaults to P can drive raw agreement below chance while the judge remains internally consistent. The per-type breakdown, not the aggregate figure, exposes the cause and localizes the divergence.

Three limitations qualify these conclusions. First, the keyframe study is a cached approximation of true k-frame sampling: because the 16 cached frames are already evenly spaced, the subsampled positions fall near, but not exactly at, those a fresh k-frame sampler would produce, so the reported flip rates are upper bounds on stability. Second, the human cross-check rests on a single annotator per cell; a multi-annotator study would place the construct-validity analysis on firmer ground. Third, Ψ inherits the compositionality and counting weaknesses of its pretraining [17], and may correctly identify an isolated entity yet misjudge a compound description or a repeated one. We regard these as motivations for extending the audit to relational and temporal claims, which we leave to future work.

7 Conclusion

Examining GroundXum as an artifact rather than a method, we find an audit that reproduces itself deterministically, a summarizer ranking that is stable to both the grounding decision threshold and the keyframe count even as the absolute rates are not, a VLM judge robust to prompt wording and option ordering

but not to chain-of-thought reasoning, and an apparent disagreement with human labels that resolves into a difference in what *presence* means rather than a judge error. The per-type breakdown localizes that divergence to non-visual entity types, concentrated in fine-grained cooking ingredients and abstract actions, where human labeling is driven by inference rather than direct observation. We release the complete pipeline, cached intermediate outputs, and reproduction scripts, so that every table and figure can be regenerated on a single workstation. Taken together, these results delineate which of GroundXum's conclusions transfer with the artifact and which are contingent on its configuration, the distinction a downstream user most needs in order to deploy or extend the audit responsibly.

A Judge Prompt Variants

We reproduce verbatim the five prompts used in Section 4; the entity surface form is substituted at `{surface}`, and v0 is the prompt used in the original paper. The variants differ only in the dimensions named in their labels, wording (v1), reasoning format (v2), option set (v3), and option ordering (v4), so that any change in the judge's verdicts can be attributed to a single controlled factor.

v0 (original).

```
This image is a 4x4 grid of 16 frames sampled uniformly from a short video.
Question: Is "{surface}" visibly present somewhere in any of these 16 frames?
Decide based ONLY on what you can see in the frames, not on what is plausible.
Answer with ONE letter:
 P : Present. You can clearly see the entity in at least one frame.
 A : Absent. The entity is not visible in any frame.
 ? : Ambiguous. Occluded, out-of-focus, ambiguous reference, or abstract.
Your answer (one letter only):
```

v1 (reword).

```
Look at this 4x4 montage of 16 frames taken from one short video.
Can you actually SEE "{surface}" in any of the 16 frames? Judge only from
the pixels, not from what you would expect to be there.
Reply with a single letter:
 P = yes, clearly visible in at least one frame
 A = no, not visible in any frame
 ? = cannot tell (occluded, blurry, abstract, or unclear what is referred to)
Letter:
```

v2 (chain-of-thought).

```
This image is a 4x4 grid of 16 frames from a short video.
Question: Is "{surface}" visibly present in any of the 16 frames?
Judge only from what is visible.
First, in one short sentence, note what you see relevant to "{surface}".
Then give your verdict on a new line as exactly "Answer: X" where X is one of:
 P (clearly visible), A (not visible), ? (occluded/blurry/abstract/unclear).
```

v3 (forced binary).

```
This image is a 4x4 grid of 16 frames sampled uniformly from a short video.
Question: Is "{surface}" visibly present somewhere in any of these 16 frames?
Decide based ONLY on what you can see in the frames.
You must answer with ONE letter:
 P : Present. Visible in at least one frame.
 A : Absent. Not visible in any frame.
Your answer (one letter only):
```

v4 (option-order swap).

```
This image is a 4x4 grid of 16 frames sampled uniformly from a short video.
Question: Is "{surface}" visibly present somewhere in any of these 16 frames?
Decide based ONLY on what you can see in the frames, not on what is plausible.
Answer with ONE letter:
 A : Absent. The entity is not visible in any frame.
 ? : Ambiguous. Occluded, out-of-focus, ambiguous reference, or abstract.
 P : Present. You can clearly see the entity in at least one frame.
Your answer (one letter only):
```

References

1. Anonymous:: Groundxum: Auditing hallucination in generative video summarization via visual grounding. GenAAI Workshop ICPR Under Review (2026)
2. Argaw, D.M., Yoon, S., Heilbron, F.C., Deilamsalehy, H., Bui, T., Wang, Z., Dernoncourt, F., Chung, J.S.: Scaling up video summarization pretraining with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8332–8341 (2024)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv e-prints pp. arXiv–2502 (2025)
4. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
5. Durmus, E., He, H., Diab, M.: Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 5055–5070 (2020)
6. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
7. Lee, M.J., Gong, D., Cho, M.: Video summarization with large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18981–18991 (2025)
8. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
9. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
10. Lin, J., Hua, H., Chen, M., Li, Y., Hsiao, J., Ho, C., Luo, J.: Videoxum: Cross-modal visual and textural summarization of videos. IEEE Transactions on Multimedia **26**, 5548–5560 (2023)
11. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 1906–1919 (2020)
12. Nazir, M., Aqeel, M., Zhang, R., Setti, F.: Multimodal abstractive summarization of instructional videos with vision-language models. International Conference on Pattern Recognition (2026)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
15. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045 (2018)
16. Tang, B., Huang, Z., Liu, C., Sun, Q., Yang, H., Lim, S.N.: Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In: International Conference on Learning Representations. vol. 2025, pp. 27745–27776 (2025)

17. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936 (2022)
18. Zheng, C., Zhou, H., Meng, F., Zhou, J., Huang, M.: Large language models are not robust multiple choice selectors. In: International Conference on Learning Representations. vol. 2024, pp. 19426–19454 (2024)
19. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
20. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)