SOUNDSENSE: EVALUATING VISUAL SOUND UNDERSTANDING IN COMICS AND WEBTOONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision and sound provide complementary evidence for understanding physical events. While generative models increasingly produce audiovisual content and vision–language models (VLMs) are widely evaluated on image captioning and visual question answering, their ability to infer auditory events from static imagery remains less explored. We introduce SoundSense, an evaluation framework and annotation resource for studying whether depicted moments warrant sound effects. Comics provide a useful testbed because artists explicitly represent selected sounds through onomatopoeia. On eight Manga109-s volumes, we use these annotations as a proxy for sound placement, remove the lettering through local inpainting, and analyze VLM predictions across $6{,}238$ panel-level decisions. Complementing this proxy, we provide 450 author-confirmed sound-effect regions across four vertical-scroll webtoons, including an audio-backed core of 359 moments with 421 reference clips. Among 444 regions with lettering metadata, 65.5% contain no sound-effect lettering. We benchmark five VLMs, including two hosted open-weight model families, on a frozen scope of 808 webtoon segments. We report localization, output-format coverage, cohort differences and clustered bootstrap intervals. On 750 shared parse-valid segments, localization F_1 ranges from 0.159 to 0.680. An output audit reveals that coordinate-format compliance and spatial matching materially affect these scores, highlighting the need to distinguish spatial output failures from sound judgments. SoundSense provides a reproducible basis for evaluating visual sound understanding, analyzing prediction errors, and developing sound-aware systems for visual narratives. We release the webtoon annotations, reference audio, and source links.

1 INTRODUCTION

Human perception integrates vision and hearing, as demonstrated by audiovisual speech illusions (McGurk & MacDonald, 1976). Machine systems have learned to predict impact sounds from silent videos (Owens et al., 2016) and to generate visual–audio content jointly (Xing et al., 2024). Vision–language models (VLMs) also follow instructions about visual content (Liu et al., 2023). Less established are benchmarks for deciding where to place sound effects in static comics, including whether to add a cue at all. We isolate a concrete operational task: given an image, propose regions that warrant a sound effect under an explicit curation protocol.

Comics are a convenient testbed, with one caveat we make explicit. Drawn onomatopoeia provides artist-selected cues to sounds or states of depicted objects (Baek et al., 2022), giving a scalable but incomplete annotation proxy. The caveat: this record is a *proxy*, not a definition of audibility, since a panel can imply sound (a shout, a footstep) without any drawn word, and ideophonic expressions can depict sensory imagery beyond sound (Dingemanse, 2012). We therefore use manga to *analyse* model behaviour against this proxy, and build a separate webtoon corpus, curated from depicted events, to measure the proxy's coverage of chosen sound placements. The inference is image-only: unlike video-to-audio there is no audio track and no motion, so a model must judge from a single static drawing. Unlettered panels are often visually rich: 24.7% of our requested manga panels are proxy-positive, while many negatives still contain sound-capable objects. This separates object presence from sound-placement decisions without equating unlettered imagery with physical silence. The task is two-sided: identify supported sound moments while avoiding unnecessary additions.

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.

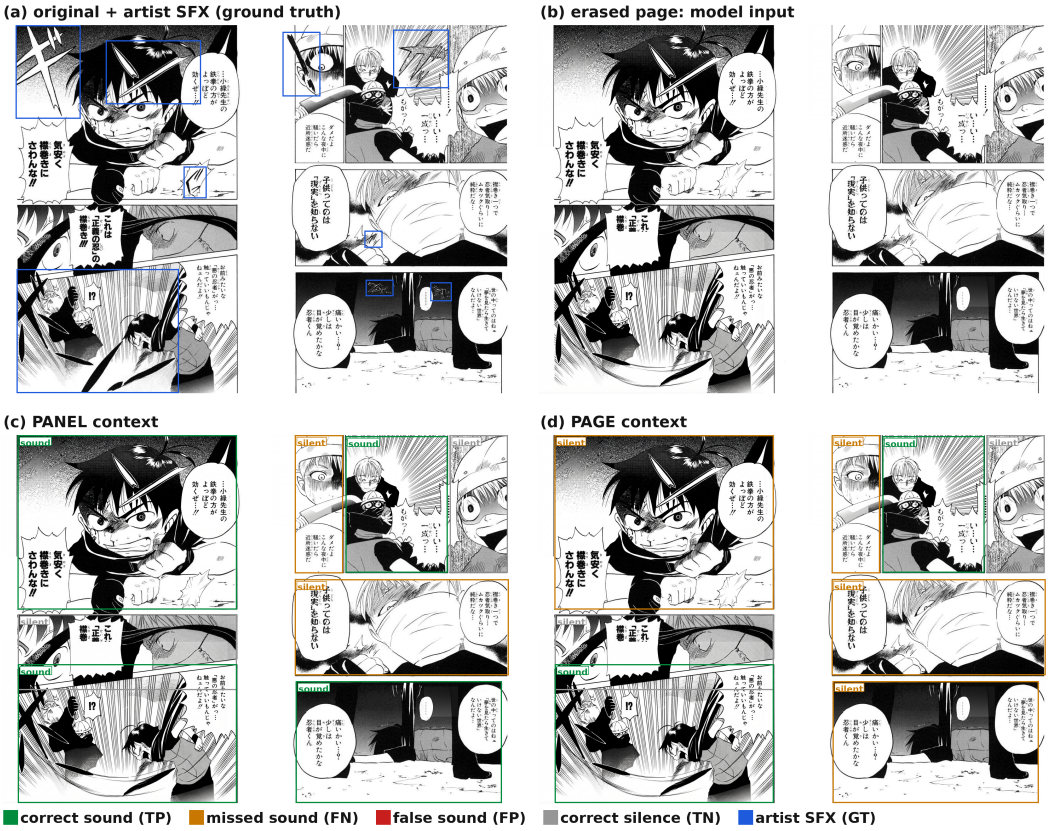


Figure 1: One page (*AkkeraKanjinchou* p.85; 8 panels, 9 annotated sound words, 6 proxy-positive panels), from Manga109-s. **(a)** original with ground-truth onomatopoeia boxes; **(b)** after local erasure — the model's input; **(c)** PANEL context recovers 4 of the 6 proxy-positive panels with no false alarms; **(d)** the PAGE regime retains only 2. The edit-detection audit in §4.3 measures a possible shortcut (AUC-ROC 0.714), rather than establishing leakage-free input. This page is an illustration — among pages where PANEL beats PAGE it was chosen for erasure quality by a pixel residue and seam measure; aggregates are in Table 4.

We contribute:

1. **An author-curated webtoon resource, with a five-model benchmark** for the sound-effect placement task (§3.2, §4.2). To extend beyond the page-based format of Manga109-s, we curate 450 regions across four works. The resource contains 359 audio-backed moments, six second-pass additions, and 85 regions from two newly reviewed chapters. Models attain localization F_1 from 0.159 to 0.680 under the requested coordinate format. We report output-format errors and matching sensitivity alongside these scores. We provide the webtoon annotations (bounding boxes, categories, trigger points and sound labels) together with the 421 reference audio clips and the manga analysis protocol; the data bundle excludes source images. The anonymous supplementary package contains the authoritative versioned annotations, audio, source-input manifests, raw responses and replay code for this submission.
2. **An analysis of VLM sound reasoning on manga**, using artist onomatopoeia as a cheap proxy (§4, §5). Classifying 16,221 model justifications shows that, among the *stated rationales*, models appeal to *visible* energy where they assert *audible* energy: optical-effect rationales are 2.7% of decisions but 12.9% of false alarms, while speech rationales accompany 42.3% of misses. We name *visual-effect-as-sound* as a recurring pattern of disagreement with the lettering proxy. In the hard-label comparison, the full-page regime attains lower

F_1 in all eight selected volumes, with resolution, batching and panel association remaining confounded (§6).

3. **A measurement of lettering coverage of curated sound placements** (§4.2). 65.5% of the 444 regions with lettering metadata carry no lettering. We report reference recall separately for lettered and unlettered regions. These observations describe this curated corpus rather than population-level audibility or a causal effect of text on model predictions.

4. **Auditable evaluation protocols** (§3.3, §4.3, App. C). We quantify presentation and output-format effects, audit the image-edit channel, and supply exact prompts, source manifests, raw predictions, matching sensitivity and paired uncertainty estimates. Historical confidence-elicitation results are distinguished from wording-controlled comparisons (App. A.2).

2 Related work

Comic understanding. Manga109 was introduced for sketch-based retrieval (Matsui et al., 2017) and extended with frame, text and character annotations (Aizawa et al., 2020). COO (Baek et al., 2022) adds onomatopoeia polygons, transcription and truncated-text links for text detection and recognition. Recent benchmarks evaluate multimodal models on comics more broadly—MangaUB (Ikuta et al., 2024) on scene and emotion, CoMix (Vivoli et al., 2024) across detection, linking and dialogue, PixelHumor on multimodal humor and panel sequencing (Ryan et al., 2025), and STRIPCIPHER on implicit narrative reasoning in silent comics (Wang et al., 2025). These evaluations emphasize visual entities, text and narrative understanding. AI4VA-FG combines region-aware zooming with reinforcement learning for fine-grained comic tasks (Chen et al., 2025), relevant to our PANEL-vs-PAGE confound (§6). We reuse Manga109-s and COO for their coordinates; our primary webtoon target is curated cue placement, including unlettered and zero-reference segments.

Sound for visual media. Diff-Foley synthesizes soundtracks from silent video (Luo et al., 2023); audio-visual event localization identifies events that are both audible and visible in audiovisual input (Zhou et al., 2025). Cross-modal diffusion also supports visual-steered audio and joint visual–audio generation (Xing et al., 2024). SonoWorld places language-guided sound anchors and renders spatial audio in a scene generated from one image (Jin et al., 2026). FlatSounds distinguishes perceptual plausibility from physical correctness using controlled video-to-audio tests (Li et al., 2026). Our focus is the agreement of static-comic cue proposals with curated placements, rather than evaluating generated waveforms. Onomatopoeia has been related to audio directly, e.g. cross-modal retrieval between sounds and onomatopoeic images (Imoto et al., 2026); we use the lettering not as a retrieval target but as a positional label in the manga protocol, and then remove it. Webtoon inputs retain their original lettering.

Hallucination, grounding and abstention. CHAIR evaluates object hallucination in captions (Rohrbach et al., 2018), while POPE uses object-presence queries for VLMs (Li et al., 2023). Both target unsupported visual object assertions. Our negatives are different in kind: an unannotated panel can contain sound-capable objects; its negative label is specific to our placement protocol, not a verified physical absence of sound. That the justification a model gives may not reflect the computation behind its answer (Turpin et al., 2023) is why we treat the reasons we collect as descriptive evidence for a taxonomy rather than as explanations. We state the reference and scoring rules explicitly for each evaluation.

Uncertainty without token probabilities. Self-evaluation can signal correctness on studied language tasks (Kadavath et al., 2022). Verbalized confidence was often better calibrated than conditional probabilities for studied RLHF models on three QA benchmarks (Tian et al., 2023); this does not establish calibration for our VLM task. Repeated reasoning paths support self-consistency voting (Wang et al., 2023), while semantic entropy groups sampled answers by meaning to detect confabulations (Farquhar et al., 2024). Our historical experiment averages verbalized decision confidences as a ranking score, not as either of those algorithms (App. A.2). Ranking performance is distinct from probability calibration, and that experiment has a wording confound.

3 METHOD: TWO REFERENCE PROTOCOLS FOR SOUND PLACEMENT

Both protocols are stated before any results. The manga protocol buys scale and a low-collateral local edit, at the cost of defining sound as whatever the artist lettered; the webtoon protocol drops that assumption, at the cost of scale; its labels are curated by one author with model-assisted candidate marking and prompt drafting.

3.1 MANGA: ONOMATOPOEIA AS A PROXY FOR SOUND PLACEMENT

Labels. We use Manga109-s (Aizawa et al., 2020; Matsui et al., 2017) with COO onomatopoeia polygons (Baek et al., 2022). A panel is defined by a Manga109 frame box; its label is positive iff at least one onomatopoeia centroid falls inside it. Centroid assignment covers 98.2% of sound words (an area-overlap rule covers 98.7%). High assignment coverage does not resolve the semantic mismatch between lettering and audible events.

Erasure. To limit possible collateral changes to dialogue and dense line art, we erase *locally*: for each onomatopoeia box we crop a padded patch, ask an image editor to remove only that lettering, resize the returned patch to the crop, and paste it back. All pixels outside the patches are byte-identical to the original.

Quality control. Each attempt is scored by (i) mean absolute pixel change inside the sound-word box (did the lettering go?) and (ii) whole-page mean change *outside* all sound-word boxes (was unrelated art damaged?). Up to three attempts are made per region; a page failing either criterion is marked `not_clean` and **the entire page is excluded**, since residual text or collateral edits could contaminate both its positive and negative panels. Of 746 prepared page images, 639 contribute panel decisions. Rechecking the 107 others distinguishes 71 failing the pixel-based QC rule from 36 QC-passing pages without frame annotations. Thus the 14.3% attrition is not solely erasure failure. These exclusions may depend on style, typography and annotation coverage (§6). Implementation parameters and the page inventory are provided in Appendix D.

Task. Given a panel (with the sound word erased), the model outputs a binary decision, an open-vocabulary sound description, and a one-sentence justification. Every prediction retains its justification and the raw API response, enabling failure analysis.

3.2 WEBTOONS: AUTHOR-CURATED SOUND-EFFECT REGIONS

Labels defined by drawn onomatopoeia can only mark moments the artist chose to *letter*. Whether that is a small or a fatal restriction cannot be answered inside the proxy, because a sonic-but-unlettered moment is indistinguishable there from a silent one. We therefore built a second reference set from *depicted events*, without requiring lettering, extending the study to vertical-scroll webtoons.

The original three Chinese-language webtoons contribute 10 curated installments (including one preview) and 571 strip segments. One author curated the references segment by segment, reviewing every strip segment in full rather than a pre-filtered subset. To accelerate the work, a VLM first marks candidate sound moments from imagery (explicitly including moments with no drawn sound word); the annotator then examines each segment, correcting the box and trigger point, adjusting or rejecting candidates, adding missed events when identified, and—crucially—**attaching a reference audio clip** (uploaded or generated) to every confirmed moment, so "this should sound" is committed to a concrete sound rather than left as a label. A moment is one confirmed region and may carry more than one clip, giving 359 curated moments with 421 clips. The annotator reports adjusting regions, trigger points, prompts and audio during curation; the final files do not retain a complete edit history. The annotation records a candidate's initial *source* (model or hand-added), not the extent of subsequent editing, so those counts do not measure the manual effort. Curation was performed by a single annotator, so we report no inter-annotator agreement. A later source-blinded review of 20 selected unmatched candidates accepted six additions, identified three as existing events, and rejected eleven. One accepted box was redrawn by the author. That revision has 365 regions; these six additions have no verified audio, trigger, category or lettering metadata. Audio statistics concern the original 359 moments; lettering denominators are stated separately for each version. Appendix C.2 documents the selection and before/after results. We then froze two completed chapters of a fourth work, *GM*

Re: Start: the author retained $52/79$ and $33/50$ proposals, adding 85 confirmed regions. These proposals used earlier confirmed labels/prompts as examples, so this is also AI-assisted curation, not an independent human-first test. They have no reference audio. The final scope is 12 installments, 817 source slices and 450 reference regions. Unreviewed expansion drafts are excluded. Lettering metadata is available for 444 regions.

The target is sound-effect placement, not exhaustive dialogue dubbing or waveform reconstruction. Each original audio-backed region records one of six categories: impact, motion, voice, animal, ambient or music. Reference clips specify a chosen sound-design interpretation; they are not recordings of the depicted event. A region or segment without a curated cue is a negative reference for this placement task, not a claim that the corresponding physical scene must be acoustically silent.

3.3 Two evaluation artifacts that must be controlled

Prediction granularity. Whole-page and per-panel free-box outputs confound presentation with query granularity and output opportunity. Our manga comparison therefore uses a **fixed decision set**: each condition is evaluated on the same panels. This equalizes evaluated decisions, but not target resolution or multi-output burden. Webtoon localization instead permits free proposals and reports proposal counts, coordinate compliance and a content-free fixed-centre baseline.

Artifact 2: prompt and output specification. Historical prompt implementations differed in confidence requests, location hints, justification length and the silence instruction. Those runs are not an isolated single-clause ablation. We disclose exact templates and separate annotation, historical calibration and evaluation responses by provenance rather than attributing differences solely to silence priming (App. B). Prompt-format sensitivity has also been demonstrated in few-shot language-model evaluation (Sclar et al., 2024); our diagnostics address the particular contracts used here.

4 Experiments

4.1 Manga presentation regimes

With the decision set fixed, we compare panel crops with numbered full-page inputs and a preceding-page variant. The corrected hard-label PANEL/PAGE prompts share the silence instruction and parked-car example. Full-page input produces fewer positives and lower F_1 in all eight selected volumes. This is a presentation-regime comparison, not an isolated context intervention: target resolution, panel-number association and response batching change together. Appendix A gives the full results and prompt provenance. The separate confidence experiment retained older panel wording and is reported as historical evidence only; it does not independently establish a controlled context effect (App. A.2).

4.2 What the onomatopoeia proxy misses

We turn the webtoon corpus (§3.2) on the assumption behind the manga labels.

Lettering covers only part of the curated placements. $\mathbf{65.5}\%$ of the 444 regions with known lettering metadata carry *no* sound-effect lettering (Fig. 2a); the original audio-backed core alone gives 61.3%. Coverage varies across the selected works. These are descriptions of curated placements, not estimates of population audibility or a causal effect of art style.

Evaluation scope and matching. The frozen manifest includes 12 completed installments: 817 source slices minus 9 shorter than 200 px gives 808 requested segments. Geometry restores four original annotations lacking legacy segment IDs. The two cross-boundary references each count once, with predictions eligible from either visible segment. The primary comparison uses the 750 segments with parse-valid outputs from every model and 418 visible references; per-model format coverage is reported separately (App. C). Of these segments, 394 have no reference region. They are included when counting false positives, but absence of a reference is not independent proof of physical silence.

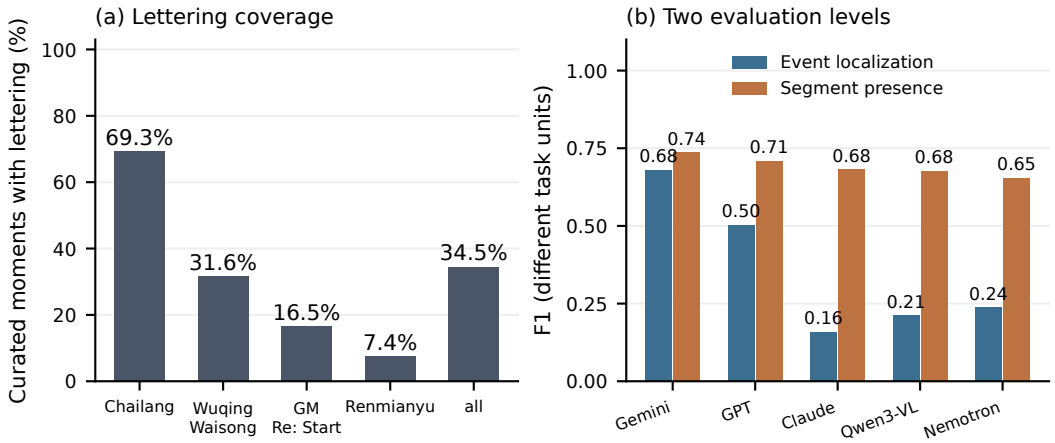


Figure 2: The webtoon result. **(a)** Share of regions carrying sound-effect lettering, by work; the denominator is 444, excluding six additions with unknown lettering metadata. These proportions describe this corpus, not a population estimate. **(b)** End-to-end event localization and segment-presence scores on 750 common parse-valid segments (418 references), derived from the same model responses. The task units differ: this is not a before/after improvement comparison. Separating the two levels helps distinguish proposing sound from placing it; uncertainty and baselines are reported in Tables 1 and 2.

Table 1: Localization on the common parse-valid scope: 750 segments and 418 references, using the requested normalized coordinates. Invalid coordinates count as unmatched output-format failures, not verified semantic hallucinations. Brackets are 95% episode-clustered percentile bootstrap intervals (10,000 shared resamples of 12 installments; Field & Welsh, 2007); segment-level and paired-difference intervals are supplied with the results. The fixed-centre baseline emits one full-segment box per input without reading image content.

Model	Precision	Recall	F_1
Gemini	0.537 [0.45,0.63]	0.926 [0.89,0.95]	0.680 [0.60,0.75]
GPT	0.381 [0.31,0.45]	0.739 [0.68,0.79]	0.502 [0.43,0.57]
Claude	0.098 [0.08,0.12]	0.421 [0.33,0.51]	0.159 [0.13,0.19]
Qwen3-VL	0.134 [0.10,0.17]	0.502 [0.45,0.56]	0.212 [0.17,0.26]
Nemotron	0.153 [0.12,0.19]	0.531 [0.50,0.56]	0.237 [0.20,0.28]
Fixed centre	0.232	0.416	0.298

All five models receive the same prompt and image bytes, without reference labels or annotation examples. New requests use settings-keyed evaluation caches; valid original evaluation responses are reused. The primary localization metric uses maximum-cardinality one-to-one centre-in-box matching, restricted to references visible in the input segment. It evaluates location, not correctness of the generated sound label. We additionally report legacy output-order greedy matching and intersection-over-union (IoU), a standard box-overlap measure in object detection (Everingham et al., 2010), at thresholds 0.1, 0.3, and 0.5. Reference regions may cover a depicted beat rather than an isolated object, making these different localization criteria rather than interchangeable scores (App. C).

Separating output compliance from sound judgment. Table 1 compares Gemini, GPT, Claude, Qwen3-VL and Nemotron. Across parse-valid calls, $945/1912$ Claude candidates violate the requested 0–1000 coordinate range, versus $0/772$ for Gemini and $1/867$ for GPT. Treating every model's coordinates instead as native image pixels gives Claude $F_1 = 0.311$, Gemini 0.179 and GPT 0.221. This alternative is a diagnostic, not evidence that pixel interpretation is correct. The large effect prevents interpreting Claude's low localization score entirely as poor auditory reasoning. Qwen3-VL is range-compliant but also convention-sensitive: uniformly swapping normalized axis order gives it $F_1 = 0.366$ versus 0.212 under the requested order. The content-free fixed-centre

Table 2: Segment-level presence derived from candidate lists, independent of coordinate validity. The reference set has 356 positive and 394 zero-reference segments. Model F_1 intervals use episode bootstrap; the always-positive row is a descriptive prevalence baseline.

Model	Precision	Recall	F_1 [95% CI]	Neg. FPR
Gemini	0.587	0.989	0.736 [0.67,0.81]	0.629
GPT	0.580	0.910	0.708 [0.64,0.78]	0.596
Claude	0.517	0.997	0.681 [0.60,0.77]	0.840
Qwen3-VL	0.515	0.989	0.678 [0.60,0.77]	0.840
Nemotron	0.511	0.910	0.655 [0.58,0.74]	0.787
Always propose	0.475	1.000	0.644	1.000

baseline attains $F_1 = 0.298$, further motivating an output-compliance audit before interpreting the ranking semantically. JSON wrappers and nested arrays are normalized by a documented adapter applied to every model; missing API responses are reported separately rather than counted as silence.

Presence versus localization. To separate the decision to propose sound from coordinate accuracy, we derive a binary segment prediction: positive iff the model emitted at least one candidate, even if its coordinates are invalid. Reference presence means at least one curated region is visible. This is a diagnostic of the same responses, not a separately prompted classification experiment. The model gap narrows substantially (Table 2), while false-alarm rates on the 394 zero-reference segments remain high. An always-positive baseline already attains $F_1 = 0.644$ at this prevalence, showing why high recall alone is insufficient evidence of accurate sound placement.

Coverage and cohort diagnostics. We report reference recall by lettering status and separate the original three-work cohort from the two GM chapters in Appendix C.1. This tests whether aggregate scores conceal cohort differences, but does not identify the effect of text: an unlettered region may share a segment with dialogue or other lettering. The GM extension has only two installments and was also curated from AI proposals; its scores are descriptive rather than independent validation.

4.3 EDIT-DETECTION AUDIT

Proxy-positive panels are inpainted and controls from unlettered pages are not, so a model could in principle score "sound" by spotting repairs rather than reasoning about the depicted event. This is a possible shortcut-learning channel (Geirhos et al., 2020). We probe it rather than assuming that edited inputs are leakage-free.

The probe distinguishes the selected groups. We show panels to the same VLM and ask *only* whether the image has been digitally retouched, never mentioning sound (150 inpainted positives vs. 150 never-touched panels from unlettered pages). The probe reaches **AUC-ROC** 0.714 (Average Precision 0.742, with equal scores grouped at each threshold): it flags 44% of edited panels but only 4% of untouched ones. We therefore *cannot* claim the shortcut is unavailable. These groups may also differ in content and style; the probe is not a causal estimate of how much sound-task performance comes from editing artifacts.

Detector misses and interpretation limits. Across 16,221 justifications, the rule-based classifier flags 2 text-related rationales (including 1 containing CJK characters) and 5 artifact-related rationales. The rarity of these mentions does not show that edit cues had no influence on predictions. 56% of edited panels go undetected, and the score distributions overlap heavily (median $P(\text{edited})$ 0.32 for edited vs. 0.28 for untouched); only the upper quartile separates. Detection rates differ across sound-word coverage bins (53.3% detected when it covered 1–5% of the panel vs. 41.7% at $\geq$5%); these rates do not isolate effects of glyph size or background complexity.

A negative probe result does not establish absence of leakage. The detector can miss edits, and sub-threshold artifacts may still carry label information. Equivalent sham edits on matched negative panels would be needed to isolate this channel. We have not completed that control, so absolute

Table 3: What the model appeals to, and what it gets wrong. 16,221 justifications from the eight-volume run, classified by stated basis and crossed with the outcome (four of six content categories shown; *event* groups physical actions). $P(\text{sound})$ is how often that basis asserts sound; the last two columns give each category's share of all 990 false alarms and 2,418 misses. These are associations with the lettering proxy, not mechanism estimates.

Stated basis	% decisions	$P(\text{sound})$	Precision	% FP	% FN
visual effect	2.7	0.521	0.436	12.9	1.0
event	26.2	0.200	0.497	43.1	24.4
static	9.6	0.005	0.571	0.3	11.5
speech	42.9	0.002	0.733	0.4	42.3

manga metrics may reflect editing cues as well as event understanding. The image-only webtoon evaluation avoids this particular erasure channel.

5 Discussion

What the model says it is reasoning about. Valid responses retain a one-sentence justification. We classify each by its stated basis with a transparent rule-based scheme: keyword patterns for six mutually-exclusive categories (visual effect, event, object, speech, static, other), applied in a fixed priority order so the first match wins, with any justification containing CJK characters flagged separately as a leakage indicator. This is a coarse automatic pass over the model's own words, not a validated human coding, and—because a justification need not reflect the computation behind the answer (§2)—the rates below describe *stated rationales*, not internal mechanism. Of the 16,239 requested decisions, 18 lack valid outputs and are excluded, leaving 16,221 rationales reconstructed from runner responses. Two categories exhibit contrasting patterns. Justifications appealing to a *visual effect* — "a bright sphere of energy glowing near the hand" $\rightarrow$ *flash of light* — are only 2.7% of decisions but 12.9% of false alarms, a $4.8\times$ over-representation, with the highest willingness to assert sound ($P(\text{sound}) = 0.521$) and the lowest precision (0.436) among content categories: optical language accompanies a sound assertion. This is a textual association, not proof of physical silence or of an internal cross-modal confusion. Conversely, *speech* justifications are 42.9% of decisions and almost never assert sound (0.002), yet accompany 42.3% of all misses — dialogue-heavy panels are treated as quiet even when the artist lettered a sound in them. Two further recurring patterns are ambient over-scoring ("wind blowing across the face" $\rightarrow$ *whoosh*, where nothing was drawn) and object-triggered assertion, naming a sound-capable object rather than an action.

Complementary evidence, distinct failure sources. Manga rationales mentioning optical effects are enriched among proxy false positives, while the webtoon results distinguish sound proposals from localization compliance. These observations motivate studying how models infer sound, but do not identify a shared internal mechanism. The webtoon coordinate audit adds another failure source: schema non-compliance can look like poor event localization. We therefore separate reference coverage, coordinate compliance, spatial matching and stated rationales rather than interpreting every unmatched prediction as an auditory hallucination.

Why this matters for evaluation. Different prompt/output specifications and prediction budgets changed our historical measurements; the webtoon audit additionally exposed coordinate-compliance effects. Over-prediction rates for sound-placement systems should therefore be interpreted with the full input, output, parser and matching protocol disclosed, rather than attributed directly to an internal reasoning failure.

6 Limitations

Labels and scope. Manga labels measure lettering placement, including mimetics, and absence of lettering does not imply physical silence. The webtoon references are one annotator's curated sound-design decisions; alternative valid choices and missed events remain possible. The possibility of multiple reasonable interpretations is also recognized in semantic annotation (Aroyo & Welty,

2015); our single-curator corpus does not measure that variation. Fresh model inference prevents reuse of annotation-time responses but does not remove possible effects of AI-assisted curation. Second-pass candidates were selected from the evaluated outputs, so the revised labels are not an independent test set; the small selected sample does not estimate corpus-wide omissions. We release the versioned references and report no inter-annotator agreement. Visual narrative meaning involves relations across images (Cohn, 2013). The author reads continuous chapters, while inference receives individual segments; some placement choices may depend on narrative context unavailable in a model's input. **Controls.** Inpainting is detectable; no matched sham-edit control has been completed. Panel/page input also changes resolution and response burden, while the historical confidence experiment has a wording mismatch. Neither establishes a causal context effect. **Metrics and generalization.** Localization does not validate sound identity or audio quality. Results depend on coordinate compliance, parsing coverage and matching policy. Hosted preprocessing, model size and language coverage also differ. Episode bootstrap accounts for within-episode dependence but has only twelve clusters and four works; neither it nor the density-selected eight-volume comparison supports generalization to all visual narratives.

7 Conclusion

SoundSense pairs a manga lettering proxy with 450 author-confirmed webtoon regions, including an audio-backed core. 65.5% of regions with lettering metadata are unlettered, motivating evaluation beyond visible sound words. Our five-model benchmark covers both annotated and zero-reference segments, and reports coordinate-format compliance, spatial matching and segment-level proposals separately. The resulting localization scores and sensitivity analyses provide a reproducible starting point for visual sound understanding, rather than a claim that audibility or sound-design preference is uniquely determined by a drawing. We release the webtoon resource and the accompanying evaluation artifacts.

AI use statement

Generative AI tools assisted with research ideation and experimental design, implementation and debugging of data-processing and evaluation code, literature discovery, interpretation of results, and drafting, restructuring, and editing the manuscript. VLMs proposed candidate annotations and audio prompts; the author reviewed the webtoon segments and curated the final annotations. Generative models were also used for local lettering removal and the generation of reference sound effects. Evaluated predictions also supplied candidates for the disclosed same-author second-pass review. The author is responsible for the final claims, annotations, code, and citations. No AI-assisted mathematical proofs are presented.

Reproducibility statement

The supplementary replay package contains the revised geometry-based reference manifest, raw webtoon model responses, deterministic JSON adapters, matching code, paired bootstrap analysis, and classifier-order sensitivity code. Main localization and sensitivity tables can be recomputed without source comic images or model API access. Exact prompt templates and regexes are included in Appendix B and as machine-readable files. The supplementary package preserves the original annotation/audio release and specifies the revised evaluation mapping, exclusions, second-pass receipts and localization-only additions. The reference versions and their results are supplied. Source hashes distinguish this evaluation from historical workshop runs. We document collection, intended use and reuse constraints following dataset-documentation recommendations (Gebru et al., 2021).

Ethics statement

The supplementary data bundle contains derived annotations and reference audio, not source comic image files. Creator/platform terms and audio-generation terms remain distinct from the annotation license; provenance and source links are supplied. Labels represent one author's sound-design choices, including fictional and stylized effects, rather than objective claims of physical audibility. No additional human annotators or external study participants were recruited.

REFERENCES

Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18, 2020. doi: 10.1109/MMUL.2020.2987895.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015. doi: 10.1609/aimag.v36i1.2564.

Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. COO: Comic onomatopoeia dataset for recognizing arbitrary or truncated texts. In *European Conference on Computer Vision (ECCV)*, pp. 267–283, 2022. doi: 10.1007/978-3-031-19815-1_16.

Yule Chen, Yufan Ren, and Sabine Süsstrunk. Zooming into comics: Region-aware RL improves fine-grained comic understanding in vision-language models. *arXiv preprint arXiv:2511.06490*, 2025. URL https://arxiv.org/abs/2511.06490.

Neil Cohn. Visual narrative structure. *Cognitive Science*, 37(3):413–452, 2013. doi: 10.1111/cogs.12016.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874.

Mark Dingemanse. Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10):654–672, 2012. doi: 10.1002/lnc3.361.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Hikaru Ikuta, Leslie Wöhler, and Kiyoharu Aizawa. MangaUB: A manga understanding benchmark for large multimodal models. *arXiv preprint arXiv:2407.19034*, 2024. URL https://arxiv.org/abs/2407.19034.

Keisuke Imoto, Yamato Kojima, and Takao Tsuchiya. Audio-image cross-modal retrieval with onomatopoeic images. *arXiv preprint arXiv:2605.17509*, 2026. URL https://arxiv.org/abs/2605.17509.

Derong Jin, Xiyi Chen, Ming C. Lin, and Ruohan Gao. SonoWorld: From one image to a 3d audio-visual scene. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30194–30204, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jin_SonoWorld_From_One_Image_to_a_3D_Audio-Visual_Scene_CVPR_2026_paper.html.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Tingle Li, Siddharth Gururani, Kevin J. Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking single-factor physical video-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1939–1949, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Benchmarking_Single-Factor_Physical_Video-to-Audio_Generation_CVPR_2026_paper.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 48855–48876, 2023. doi: 10.52202/075280-2121.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. doi: 10.1007/s11042-016-4020-z.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976. doi: 10.1038/264746a0.

NVIDIA. NVIDIA Nemotron Nano 12B v2 VL: Model card. https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-BF16, 2025. Model card, accessed September 25, 2026.

Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2405–2413, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Owens_Visually_Indicated_Sounds_CVPR_2016_paper.html.

Qwen Team. Qwen3-VL-235B-A22B-Instruct: Model card. https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct, 2025. Model card, accessed September 25, 2026.

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, 2018. doi: 10.18653/v1/D18-1437.

Yuriel Ryan, Rui Yang Tan, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Humor in pixels: Benchmarking large multimodal models understanding of online comics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14024–14050. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.755.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://iclr.cc/virtual/2024/poster/18650.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275.

Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. CoMix: A comprehensive benchmark for multi-task comic understanding. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, pp. 140828–140846, 2024. doi: 10.52202/079017-4470.

Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Qingxiu Dong, Rui Li, Yixin Yang, Yifan Pu, Weiyao Luo, Yiru Wang, Xiangdi Meng, Wenjie Li, and Zhifang Sui. Beyond single frames: Can LMMs comprehend implicit narratives in comic strip? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6436–6452. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.342.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://iclr.cc/virtual/2023/poster/11718.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7151–7161, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xing_Seeing_and_Hearing_Open-domain_Visual-Audio_Generation_with_Diffusion_Latent_Aligners_CVPR_2024_paper.html.

Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards open-vocabulary audio-visual event localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_Towards_Open-Vocabulary_Audio-Visual_Event_Localization_CVPR_2025_paper.html.

A NARRATIVE-CONTEXT EXPERIMENT (FULL)

This appendix gives the full context experiment summarised in §4.1: the fixed decision set on eight Manga109-s volumes, the added CHAPTER condition on five, and the threshold-free calibration on two. Caveats—the PANEL-vs-PAGE confound (resolution, panel-number association, batched output) and the clustered, density-selected nature of the volumes—are in §6.

A.1 CONDITIONS AND HARD-LABEL RESULTS

Conditions. With the decision set fixed, we vary the presentation regime: PANEL (the panel crop), PAGE (the full page with every panel outlined and numbered, answering for all panels in one call), and the historically named CHAPTER condition (all preceding QC-usable pages of the volume, then the numbered target page). This condition uses a book prefix, not independently annotated chapter boundaries.

Presentation-level differences. Across eight volumes, Table 4 shows the false-positive rate against lettering-proxy negatives falling from 0.096 to 0.066 between panel and page input. The chapter comparison uses a different five-volume subset and should not be inserted into that sequence.

Panel-only wins in $\mathbf{8/8}$ volumes (macro F_1 0.435 vs. pooled 0.455). The volumes were density-selected rather than randomly sampled, so this consistency is descriptive (§6). Prompt-keyed cache auditing matched $6{,}235/6{,}238$ panel responses only to the corrected prompt, with three identical under old and new prompts. Reconstruction also found 522 stale rationales in the old merged results, including 22 changed predictions; all table entries now use runner responses. Four page outputs and fourteen chapter outputs lack valid decisions and are treated as coverage failures, not silent predictions. The hard-label comparison still does not isolate context from resolution and batching.

Changes in predicted positives. Moving PANEL→PAGE changes true-positive counts by -39.8% and false-positive counts by -31.3%, while precision *falls* ($0.564 \rightarrow 0.532$). Predicted positives drop from $1{,}039$ to 664. This is an observed shift in the full-page regime, not a causal attribution to narrative context.

Density does not rescue context. Our first five volumes sat in the bottom $\sim$5% of the corpus by sound-word density, so we added three at the corpus median (6.05) and maximum (17.12). The ordering is unchanged in every case across these selected density levels; this does not isolate narrative or visual complexity.

A.2 Historical confidence experiment: wording not controlled

This archived experiment elicited confidence scores on a two-volume subset. An offline audit found all $2{,}673$ panel samples per model matched only the *legacy* panel prompt, which lacks "do not add sound just because something is present." Page prompts include that instruction. Table 5 reports this provenance audit instead of treating the archived AUC values as wording-controlled confirmation of the hard-label results.

Confidence elicitation. The archived protocol requested verbalized confidence to use a common scoring interface across the tested APIs, rather than relying on token log probabilities. We make no general claim about logprob support in other models or deployments.

Score. We average verbalized decision confidences across repeated samples: each returns a decision plus a confidence in *that* decision, converted to $P(\text{sound}) = c$ if the answer is "sound" and $1-c$ otherwise, averaged over up to $k = 3$ available samples, with confidence rescaled to $[0, 1]$. Model identifiers and known API settings are documented in Appendix D; missing historical settings are not reconstructed by assumption.

The three vendors share the same panel subset and are not independent datasets. Moreover, 70 expected GPT page-sample runner files were absent in the provenance audit. Decoding settings and provider snapshots were not included in the legacy cache key. These limitations prevent interpreting the archived AUC difference as a causal presentation or context effect.

Table 4: Fixed decision set, `gemini-3.7-flash`. The corrected PANEL/PAGE prompts share the silence instruction and parked-car example; CHAPTER omits the example and is exploratory (App. B). Upper block: eight volumes, $6{,}238$ requested decisions per condition, scored on $6{,}234$ paired valid outputs. Lower block: five volumes with $3{,}747$ outputs valid under all three conditions. Panel-only attains the best F_1 in $8/8$ volumes on the paired valid subset.

Regime	Precision	Recall	F_1	Neg. FP rate	Pred. positives
Eight volumes: 6,234 paired valid outputs					
PANEL	0.564	0.381	0.455	0.096	1,039
PAGE	0.532	0.230	0.321	0.066	664
Five volumes: 3,747 valid outputs in all three regimes					
PANEL	0.366	0.459	0.407	0.096	505
CHAPTER	0.352	0.303	0.325	0.067	347
PAGE	0.345	0.283	0.311	0.065	330

Table 5: Historical confidence-call provenance on the nominal 891-panel subset. Each model has 2,673 panel samples matching only the pre-fix prompt. Page prompts were unchanged; missing runner files and the wording mismatch prevent a controlled panel/page interpretation.

Model	Legacy-only panel	Matched page	Missing page
Gemini	2,673	405	0
Claude	2,673	405	0
GPT	2,673	335	70

B EXACT PROMPT TEMPLATES AND JUSTIFICATION RULES

The following UTF-8 files are exported directly from versioned source constants, not paraphrased. Page templates substitute `{n}` and `{last}` and unescape doubled JSON braces using Python's `format`; the files retain the source templates. The historical calibration panel template differs from the corrected hard-label panel template, as documented in App. A.2. All files accompany the source release with the provenance audit.

Corrected manga panel prompt.

```
You are looking at ONE panel from a comic. The onomatopoeia /
↪  sound-effect
lettering has been removed from the artwork.

Decide whether THIS panel should have a sound effect, based on the EVENT
↪  depicted (an action
happening), not merely on the presence of an object that could make
↪  noise. A parked car is
silent; screeching tires are not. Many panels are intentionally silent;
↪  do not add sound just
because something is present.

Output STRICT JSON:
{"needs_sound": true/false, "sound": "short phrase or null", "reason":
↪  "one short sentence"}
```

Earlier sound-reasoning prompt (different output schema). This historical template differs from the later panel prompt in more than a silence clause. Its reported false-positive-rate comparison is not a controlled single-sentence ablation.

```
You are looking at a comic/webtoon panel. The stylized sound-effect
lettering (onomatopoeia) has been removed from the artwork. Decide,
↪  using ONLY what you can
see in the picture (and the previous panels if provided), whether this
↪  moment should have a
sound effect, and if so, what it should sound like.

Reason about the EVENT that is happening (what action is occurring), not
↪  merely which
objects are present. A parked car is silent; screeching tires are not. A
↪  holstered gun is
silent; a fired gun is a gunshot.

Output strict JSON with keys:
  "needs_sound": true/false,
  "confidence": 0.0-1.0,
  "sound_description": short open-vocabulary phrase (e.g. "metallic
  ↪  impact", "explosion",
                       "wind gust") or null if needs_sound is false,
  "location_hint": which object/area in the panel emits it, or null,
  "reasoning": one or two sentences explaining WHY (name the event you
  ↪  inferred).
```

Manga page prompt.

```
You are looking at a comic page with the onomatopoeia / sound-effect
lettering removed. Each panel is outlined and labelled with a red number.

For EVERY numbered panel, decide whether that panel should have a sound
↪  effect, based on the
EVENT depicted (an action happening), not merely on the presence of an
↪  object that could make
noise. A parked car is silent; screeching tires are not. Many panels are
↪  intentionally
silent; do not add sound just because something is present.

Answer for ALL {n} panels (numbers 0..{last}). Output STRICT JSON, a
↪  list of:
{{"panel": <int>, "needs_sound": true/false, "sound": "short phrase or
↪  null",
  "reason": "one short sentence"}}
```

Manga chapter prompt (exploratory).

```
The first images are the preceding pages of this comic chapter, in
reading order, provided as narrative context. The LAST image is the
↪  TARGET page: its panels
are outlined and labelled with a red number.

Using the story so far, decide for EVERY numbered panel of the TARGET
↪  page whether it should
have a sound effect, based on the EVENT depicted (an action happening),
↪  not merely on the
presence of an object that could make noise. Many panels are
↪  intentionally silent; do not add
sound just because something is present.

Answer for ALL {n} panels (numbers 0..{last}) of the TARGET page. Output
↪  STRICT JSON, a list
of:
{{"panel": <int>, "needs_sound": true/false, "sound": "short phrase or
↪  null",
  "reason": "one short sentence"}}
```

Historical calibration panel prompt.

```
You are looking at ONE panel from a comic. The onomatopoeia /
sound-effect lettering has been removed from the artwork.

Decide whether THIS panel should have a sound effect, based on the EVENT
↪  depicted (an action
happening), not merely on the presence of an object that could make
↪  noise. A parked car is
silent; screeching tires are not. Many panels are intentionally silent.

Output STRICT JSON:
{"needs_sound": true/false, "confidence": <integer 50-100, how sure you
↪  are of THIS answer>,
 "sound": "short phrase or null", "reason": "one short sentence"}
```

Historical calibration page prompt.

```
You are looking at a comic page with the onomatopoeia / sound-effect
lettering removed. Each panel is outlined and labelled with a red number.

For EVERY numbered panel, decide whether that panel should have a sound
↪  effect, based on the
```

```
EVENT depicted (an action happening), not merely on the presence of an
↪  object that could make
noise. A parked car is silent; screeching tires are not. Many panels are
↪  intentionally
silent; do not add sound just because something is present.

Answer for ALL {n} panels (numbers 0..{last}). Output STRICT JSON, a
↪  list of:
{{"panel": <int>, "needs_sound": true/false,
  "confidence": <integer 50-100, how sure you are of THIS panel's
  ↪  answer>,
  "sound": "short phrase or null", "reason": "one short sentence"}}
```

Webtoon candidate prompt.

```
You are helping a sound designer score a vertical-scroll webtoon. This
↪  image is
ONE segment of the strip, read top to bottom.

Find the moments in this segment that should carry a SOUND EFFECT. Judge
↪  from what is DEPICTED
-- an impact, a movement, a voice, an environment -- not from any
↪  written text. Webtoon artists
often convey sound through imagery alone, so include moments with no
↪  drawn sound word. Most
quiet dialogue or still scenery needs NO sound: return few, well-chosen
↪  moments rather than
many.

For each moment return:
  "box_2d": [ymin, xmin, ymax, xmax]   integers normalized 0-1000 within
  ↪  THIS segment
  "category": one of impact | motion | voice | animal | ambient | music
  "sound": a short human label, e.g. "metal clash"
  "audio_prompt": a vivid one-sentence prompt for a text-to-audio model,
  ↪  describing timbre,
      material and dynamics (e.g. "a single sharp metallic clash of two
      ↪  sword blades, short
      bright attack with a fast metallic decay, dry close-miked")
  "trigger_hint": 0.0-1.0, where within the box the sound should land
  ↪  while scrolling
      (0 = top edge, 1 = bottom edge; usually the instant of impact or
      ↪  full reveal)
  "has_sfx_text": true only if the artist actually drew sound-effect
  ↪  lettering here
  "sfx_text": that lettering transcribed, or ""
  "confidence": 0-100

Output STRICT JSON: a list of such objects. Return [] if this segment
↪  needs no sound.
```

Edit-detection probe.

```
You are inspecting a scanned comic panel for signs of digital
↪  manipulation.

Has any region of this image been digitally retouched, inpainted, or had
↪  content removed and
the background reconstructed? Look for locally blurred or smoothed
↪  areas, discontinuous
screentone or hatching, inconsistent line weight, or patches that do not
↪  match the surrounding
drawing style.
```

```
Output STRICT JSON:
{"edited": true/false, "confidence": <integer 50-100, how sure you are
↪  of THIS answer>,
 "evidence": "one short sentence"}
```

Exact justification classifier. The classifier first checks Unicode character names for CJK, HIRAGANA or KATAKANA, assigning `leak_text` if present. It then applies the case-insensitive patterns below in order, returning the first match; no match returns `other`. Negations and multiple concepts are not resolved semantically. These are automatic labels of stated rationales, not human coding.

```
{
  "order": [
    "leak_text",
    "leak_artifact",
    "event",
    "visual_effect",
    "object",
    "speech",
    "static"
  ],
  "flags": "re.IGNORECASE",
  "cjk_precheck": "Unicode name includes CJK, HIRAGANA or KATAKANA ->
  ↪  leak_text",
  "patterns": {
    "leak_text":
    ↪  "\\b(?:visible|see|shows?|reads?|displays?|contains?|there
    ↪  is)\\b[^.]{0,40}\\b(?:onomatopoe\\w+|sound[- ]effect
    ↪  (?:text|word|letter\\w*)|sfx
    ↪  (?:text|word)|lettering)\\b|\\b(?:katakana|kanji|hiragana)\\b",
    "leak_artifact": "\\b(?:smudg\\w+|inpaint\\w*|retouch\\w*|erased|rem
    ↪  oved|artifact\\w*|blurred (?:patch|area|region)|white
    ↪  patch)\\b",
    "event": "\\b(?:strik\\w+|hit\\w*|hitting|punch\\w*|kick\\w*|slam\\w
    ↪  *|smash\\w*|crash\\w*|collid\\w*|explod\\w*|explosion|burst\\w*|
    ↪  shatter\\w*|break\\w*|breaking|snap\\w*|fall\\w*|falling|drop\\w
    ↪  *|throw\\w*|threw|swing\\w*|slash\\w*|stab\\w*|run\\w*|running|j
    ↪  ump\\w*|leap\\w*|dash\\w*|charg\\w*|open\\w*|opening|clos\\w*|cl
    ↪  osing|knock\\w*|slid\\w*|shoot\\w*|shot|fir\\w*ing|firing|blast\
    ↪  \w*|grab\\w*|push\\w*|pull\\w*|shov\\w*|slap\\w*|scream\\w*|shou
    ↪  t\\w*|yell\\w*|cry\\w*|gasp\\w*|laugh\\w*|sob\\w*|splash\\w*|pou
    ↪  r\\w*|spill\\w*|rush\\w*|gust\\w*|blow\\w*|blowing|ring\\w*|buzz
    ↪  \\w*|beep\\w*|creak\\w*|rattl\\w*|clatter\\w*)\\b",
    "visual_effect": "\\b(?:glow\\w*|glowing|flash\\w*|shining|shine|spa
    ↪  rkl\\w*|radiat\\w*|beam of light|light
    ↪  (?:burst|effect|ray)|speed lines?|motion lines?|focus
    ↪  lines?|aura|halo|shimmer\\w*)\\b",
    "object": "\\b(?:car|vehicle|truck|motorcycle|bike|train|engine|mach
    ↪  ine\\w*|machinery|gun|pistol|rifle|weapon|sword|blade|door|windo
    ↪  w|gate|phone|telephone|bell|clock|alarm|radio|television|tv|inst
    ↪  rument|guitar|piano|drum\\w*|glass|bottle|cup|dish\\w*)\\b",
    "speech": "\\b(?:speak\\w*|speaking|talk\\w*|talking|conversation|di
    ↪  alogue|dialog|says?|saying|said|discuss\\w*|convers\\w*|monologu
    ↪  e|narrat\\w*)\\b",
    "static": "\\b(?:no
    ↪  (?:action|movement|motion|sound|event|activity)|nothing (?:is
    ↪  )?happening|static|still|quiet|calm|silent|blank|empty|motionles
    ↪  s|without any (?:action|movement|physical)|does not
    ↪  (?:require|warrant))\\b"
  }
}
```

Priority sensitivity with fixed leakage rules. We preserve the two production leakage rules and permute the five content categories over all 120 orders, recalculating each category's denominator for

every order. Visual-effect rationales occupy 1.47–3.76% of decisions and 10.91–21.62% of proxy false positives; their false-positive enrichment remains 4.47–$8.08\times$. Their sound rate is highest among content categories in $116/120$ orders, not all orders. Speech accounts for 29.03–55.29% of misses and is the largest category in $95/120$ orders. Thus the visual-effect enrichment persists, while exact proportions and category rankings depend on rule priority.

C LOCALIZATION SENSITIVITY AND UNCERTAINTY

All localization variants use the same common parse-valid images, model responses, and 418 episode-level reference IDs. A prediction may match only a reference with positive-area overlap in its input segment. Centre-in-box uses inclusive reference edges, retaining the legacy matching convention. Maximum-cardinality bipartite matching removes dependence of TP counts on prediction ordering. The two references crossing segment boundaries are never duplicated. IoU is calculated against the original reference box in strip coordinates. The fixed-centre baseline emits one full-segment rectangle for every eligible input; it uses only image dimensions, not pixels, annotations or per-segment reference counts.

Table 6: Localization F_1 sensitivity. Pixels interprets coordinates as native image pixels; axis swap interprets the four normalized values as $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$. Each alternative is applied uniformly to every model, not as a per-model selected correction.

Model	Greedy	Maximum	IoU .1	IoU .3	IoU .5	Pixels	Axis swap
Gemini	0.680	0.680	0.687	0.658	0.622	0.179	0.316
GPT	0.502	0.502	0.504	0.415	0.351	0.221	0.294
Claude	0.159	0.159	0.206	0.125	0.054	0.311	0.129
Qwen3-VL	0.212	0.212	0.321	0.119	0.034	0.156	0.366
Nemotron	0.237	0.237	0.145	0.040	0.006	0.155	0.224

The JSON adapter strips fences or surrounding prose, accepts list wrappers named `moments`, `regions`, `sounds`, `results`, or `result`, and flattens nested lists in depth-first order. It preserves every candidate. A candidate with invalid coordinates is unmatchable and counts as a format failure in end-to-end localization, rather than being silently dropped or assigned an inferred coordinate system. No confidence cutoff or sound-label semantic matcher is used. The original response and normalized candidate are retained for inspection. For compatibility with the frozen earlier evaluation, finite numeric endpoints are sorted, rounded to input pixels and given a four-pixel minimum width/height; this can extend a degenerate edge box beyond the image. We disclose this adapter rather than silently repairing individual models. The reported out-of-range count is specifically a violation of the numeric range, not a claim that every other aspect of the requested schema was satisfied. JSON decoding uses the standard last-value convention for repeated object keys; raw responses remain available to inspect such ambiguities. Bare candidate objects, empty objects and prose without a candidate list are not silently interpreted as the requested list format.

Table 7: Response coverage on the 808 requested inputs per model, and coordinate-range violations among candidates in that model's parse-valid responses.

Model	Parsed calls	Out-of-range candidates
Gemini	808/808	0/772
GPT	808/808	1/867
Claude	804/808	945/1912
Qwen3-VL	798/808	0/1655
Nemotron	764/808	11/1471

The first successfully returned response is retained even if malformed; we retry only API/transport failures. A missing or malformed response is not interpreted as silence. If any response is unparseable, localization and presence comparisons use the same intersection of parse-valid segments for every model, with reference visibility recomputed on that subset. The supplied report lists every excluded segment and retains all requested responses. Thus these scores condition on parseability; output coverage is a separate part of the evaluation.

We use bootstrap resampling (Efron, 1979), with entire episodes as sampling units to preserve within-episode dependence (Field & Welsh, 2007). Uncertainty uses $10{,}000$ percentile resamples, with identical sampled clusters across models (seed 20270919). We report both episode and segment clusters, plus paired model differences. For segment aggregation, a matched reference contributes its TP to the segment with greatest visible reference area (reading-order tie break); unmatched predictions contribute FP to their input segment. Episode resampling also covers dependence across the cross-boundary reference. These intervals quantify sampling variability conditional on the references and scoring rules, not uncertainty in the annotator's judgments or the causal source of model errors.

Presence diagnostic. An additional score ignores candidate coordinates and predicts a positive segment whenever the response contains any candidate. Reference presence is defined by positive-area overlap with any curated region. This separates willingness to propose a sound from localization compliance but does not validate which event or sound was identified. It is derived from the same responses, not a new binary-prompt experiment. The primary benchmark remains event localization.

C.1 Cohort and lettering diagnostics

The original cohort contains the first three works and the six second-pass additions. The GM extension consists of two new, author-completed chapters, with 85 regions and no reference audio. All are AI-assisted labels. The author reviewed GM before these evaluation calls; subsequent automatic pruning of other pending chapters does not enter the benchmark. Those unreviewed proposals are excluded rather than treated as new ground truth.

Table 8: Common parse-valid localization by cohort. The GM columns are descriptive results on two installments, not an independent human-first validation or a population-level estimate.

Model	Original F_1	GM F_1	GM precision	GM recall
Gemini	0.701	0.604	0.438	0.975
GPT	0.508	0.483	0.344	0.810
Claude	0.149	0.191	0.113	0.608
Qwen3-VL	0.216	0.199	0.120	0.582
Nemotron	0.250	0.196	0.119	0.544

Table 9: Reference recall by lettering metadata: matched/total (recall). Unknown metadata concerns the six earlier localization-only additions. This does not assign unmatched predictions to a lettering class, so it is not a stratified precision or F_1 comparison.

Model	Lettered recall	Unlettered recall	Unknown recall
Gemini	137/146 (0.938)	248/266 (0.932)	2/6 (0.333)
GPT	123/146 (0.842)	181/266 (0.680)	5/6 (0.833)
Claude	52/146 (0.356)	124/266 (0.466)	0/6 (0.000)
Qwen3-VL	74/146 (0.507)	134/266 (0.504)	2/6 (0.333)
Nemotron	77/146 (0.527)	143/266 (0.538)	2/6 (0.333)

Lettering flags describe the reference region, not absence of all text elsewhere in the input. Neither this stratification nor between-work differences establish a causal effect of text.

Proposal-budget diagnostic. As a post-hoc diagnostic, we retain at most one candidate per segment using the highest finite reported confidence (ties retain output order; missing confidence ranks below finite values). We apply this rule to every model without filtering invalid boxes or using reference labels. This is not the primary metric: a segment can legitimately contain multiple reference events. It helps distinguish proposal quantity from localization under the fixed-centre baseline's one-proposal budget.

Reduced-output diagnostic. We additionally select 80 distinct input images, ten per work/reference-presence stratum (seed 20260925), without conditioning on model outputs. Keeping

Table 10: Post-hoc top-one proposal diagnostic on the same common parse-valid scope.

Model	Proposals	F_1	Top-one proposals	Top-one F_1
Gemini	721	0.680	600	0.640
GPT	812	0.502	559	0.512
Claude	1798	0.159	686	0.208
Qwen3-VL	1567	0.212	683	0.282
Nemotron	1452	0.237	634	0.272

the original sound-placement instructions and image bytes, a second prompt requests only `box_2d` and `sound`. It jointly removes category, audio-prompt, trigger, lettering and confidence requests. Across both contracts and all models, 68 inputs and 46 references are parse-valid. Results are mixed rather than uniformly improved. This post-hoc comparison reuses the full-schema responses; some original-model calls lack complete historical decoding settings. Combined with single stochastic samples, this prevents isolating a causal effect of the schema or any one field.

Table 11: Reduced-output diagnostic. Parse coverage uses all 80 sampled inputs; F_1 uses the same 68 parse-valid inputs and 46 references for both contracts.

Model	Full parsed	Minimal parsed	Full F_1	Minimal F_1
Gemini	80/80	80/80	0.750	0.699
GPT	80/80	80/80	0.481	0.559
Claude	80/80	78/80	0.204	0.269
Qwen3-VL	79/80	80/80	0.236	0.267
Nemotron	74/80	77/80	0.254	0.207

C.2 Same-author second-pass review

After the original 359-reference evaluation, we assembled unmatched candidates from all three models. Invalid-coordinate outputs and candidates overlapping any reference at IoU ≥ 0.1 were excluded from this semantic review queue. The remaining 1,059 candidates were shuffled with seed 20270919 and greedily deduplicated at same-episode IoU ≥ 0.5, leaving 904. Round-robin sampling by work and presence of a visible reference selected 20 cases. The author saw full segments, candidate boxes and sound descriptions, with model identity and confidence hidden. This is the same annotator's second pass, not independent annotation or IAA.

The author accepted six new events, identified three as existing events or different localizations, and rejected eleven for sound placement. One accepted box was redrawn to restrict it to a phone and metal surface within one panel. We add only the six accepted regions, preserving the original references and all predictions. Existing-event decisions remain recorded without inferred reference-ID links or semantic matching overrides. The six additions retain accepted candidate descriptions; category, lettering, reading trigger and audio are not independently verified and are left unset. That localization revision has 365 references. The later GM extension brings the current total to 450; audio statistics describe the original 359, and current lettering statistics use 444 records with available metadata.

Table 12: Earlier reference revision only: 359 original versus 365 post-review regions, scored with identical model responses and maximum-cardinality centre matching.

Model	TP before	TP after	F_1 before	F_1 after
Gemini	333	335	0.708	0.707
GPT	259	264	0.505	0.512
Claude	142	142	0.156	0.156

Human acceptance of a sound event does not force a geometric match: an accepted candidate's original predicted centre may still fall outside the corrected box. All models are rescored against the same

revised references. Raw responses, both manifests, decision receipts, corrected coordinates and deterministic revision code are supplied. Selection from evaluated predictions can favor their proposed events; neither $6/20$ nor $11/20$ estimates a corpus-wide omission or false-alarm rate. Confidence intervals condition on the selected reference version and do not quantify annotation uncertainty.

D IMPLEMENTATION AND PREPROCESSING PROVENANCE

Inpainting implementation. The recorded editor model is `gemini-3.1-flash-lite-image`. Each COO bounding box is padded independently in width and height by 30% (at least 4 pixels per side), clipped to the page boundary, and cropped from the original image. The optional `word_hint` inserts the annotated sound text as "(such as ...)" in the following template:

```
This is a small crop from a comic panel containing onomatopoeia /
sound-effect lettering{word_hint}. Remove ONLY that SFX lettering and
↪ cleanly reconstruct
the artwork behind it so it looks natural. Do NOT change anything else,
↪ do NOT add text,
keep the same framing and aspect ratio. Return the edited crop.
```

Returned patches are converted to RGB, resized to the original crop size using Lanczos if needed, and pasted into the page. The implementation attempts up to three edits per region. Candidate selection prefers an edit passing both removal and patch-collateral thresholds, then one passing removal, then lower collateral mean difference, lower changed fraction, and larger removal difference. The exact current constants and historical QC-version counts are distributed as `inpainting_parameters.json`; some records originate from earlier QC versions, so these constants should not be mistaken for a complete per-request generation log.

QC thresholds and page inventory. Removal uses mean absolute RGB change within an SFX box, with threshold 15 on the 0–255 scale. Patch-collateral selection thresholds are mean change 12, changed-pixel fraction 0.10, and per-pixel change 20. Final page eligibility is recomputed on the composed image: every SFX box must reach removal change 15, and the mean change outside the union of SFX boxes (with a 2-pixel mask margin) must be at most 6. This pixel test is not semantic verification of complete lettering removal. Of 107 prepared images absent from the decision set, 71 fail this rule and 36 pass but contain no frame annotations; Table 13 separates them.

Table 13: Prepared page inventory and fixed panel decision set. The full volume names, page IDs, QC versions, exclusions and record hashes are supplied in the preprocessing manifest.

Volume	Prepared	Scored	QC fail	No frames	Panels	Positive
AkkeraKanjinchou	91	56	30	5	605	256
EverydayOsakanaChan	86	71	10	5	786	240
GinNoChimera	103	89	9	5	685	113
HanzaiKousyounin...	100	89	9	2	1084	637
HealingPlanet	95	88	4	3	755	93
HinagikuKenzan	89	83	3	3	791	82
MagicianLoad	88	76	4	8	716	57
YasasiiAkuma	94	87	2	5	816	60

Image consistency audit. All $6{,}238$ archived panel crops match the corresponding regions of the current clean-page images pixel-for-pixel. Of 639 numbered page files, 638 match the same underlying artwork outside a mask covering numbering badges and frame borders (with a 4-pixel margin). One numbered page file is unreadable; its four missing page decisions are excluded from the paired-valid comparison. The displayed Figure 1 page passes both checks. These checks verify input consistency, not successful semantic removal of all lettering or absence of edit leakage.

Model access and settings. Webtoon evaluation uses `gemini-3.7-flash` through Google GenAI, `global.anthropic.claude-sonnet-5` and

`global.openai.gpt-5.6-terra` through Bedrock Converse. Gemini receives image bytes and JSON response mode at temperature 1.0; Bedrock Claude uses temperature 1.0 and a 4,096 token limit. Bedrock GPT rejects the temperature field, which is omitted; its token limit is 4,096. Thus identical task prompts do not imply identical vendor decoding implementations. The new missing-sample requests record these settings and timestamps. Historical caches retain raw responses keyed by model ID, prompt, image hash and call tag, but not all decoding settings or immutable provider snapshots; we cannot retrospectively certify those missing fields. Each newly needed request runs in a subprocess with a 180-second parent timeout and is checkpointed separately. Existing matching responses are reused.

We additionally evaluate hosted open-weight model families Qwen3-VL-235B-A22B (Qwen Team, 2025) and Nemotron Nano 12B v2 VL (NVIDIA, 2025), using Bedrock IDs `qwen.qwen3-vl-235b-a22b` and `nvidia.nemotron-nano-12b-v2` in `us-east-2`, temperature 1.0 and a 4,096-token output limit. New requests retain input, prompt, model, settings and namespace in their cache identity, plus response hashes, timestamps, stop reasons and token usage. Identical image bytes do not imply identical vendor preprocessing. The Nemotron model card lists English-only language support; comic lettering is Chinese and our task prompt is English. These deployments differ in size, language coverage and preprocessing, so the comparison does not isolate any one of those factors or certify equivalence to a particular self-hosted snapshot.

Source input reconstruction. The supplement provides exact chapter URLs and ordered source-image manifests for all 12 installments, including byte hashes, dimensions and strip offsets. A public-chapter acquisition command verifies hashes and stops if publisher content differs or access is locked. Representative first-image downloads from all 12 chapters matched the frozen bytes; all locally used inputs were checked for dimensions and decodability before new inference. Publisher availability may change. Source comics and source-site music are not redistributed or used as reference audio. The GM site offers assigned music, but our evaluation inputs contain only the comic images.

Annotation and audio provenance. An author performs curation with AI-assisted candidate proposals; each of the original 359 audio-backed references includes box location, category, sound description, reading trigger and selected audio. The initial source field is not an edit log, and non-default trigger values do not establish a count of manual edits. The release preserves source labels/prompts with a known consistency note in `dataset/QUALITY_NOTES.md`; semantic agreement between these fields is not part of the location-only scorer and has not been independently validated. An earlier cache audit reported 82.7%/55.2% proximity correspondence for lettered/unlettered candidates using a 24-pixel top-left rule. These are not logged human acceptance rates: edits can break proximity and several candidates can match one annotation. They are not used as acceptance or causal evidence in the revised evaluation. The original release contains 420 clips generated through the recorded Fal/ElevenLabs endpoint and one uploaded clip that the author reports also self-generating; the latter's generation service and settings were not retained. Tool-generated clip records retain prompts and content identifiers; these do not guarantee deterministic audio regeneration. The supplementary snapshot preserves exact files and SHA-256 checksums without redistributing source comic images. Creator/platform and audio-source terms are distinct from the annotation license; the package does not assert a new blanket license over all media.

Edit-probe metric correction and replay. ROC and precision–recall summaries are not interchangeable (Davis & Goadrich, 2006). Our non-interpolated Average Precision is $\sum_t (R_t - R_{t-1}) P_t$, where P_t and R_t are precision and recall at distinct descending score thresholds, $R_0 = 0$, and equal scores enter together. The supplement retains all 300 probe responses, image/response hashes and prompt-keyed cache matches. Re-parsing the raw responses reproduces the archived per-case scores. Average Precision is computed by including all examples with the same score at one threshold, yielding 0.742; ROC-AUC is 0.714 with half-credit for ties. The historical AP estimator processed tied examples one at a time, with positives preceding negatives, producing 0.835; reversing input order alone changes that estimator's value to 0.719. We retain the historical summary for provenance and use the tie-aware result in the paper. Tests cover constant-score and mixed-tie permutations; this experiment is now included in the offline replay checks.