AVE-COMPASS: TOWARDS HOLISTIC EVALUATION FOR AUDIO-VIDEO EDITING ABILITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

While instruction-based video editing has advanced rapidly, real-world videos contain tightly coupled audio and visual signals, and editing one modality often requires coordinated changes in the other. Existing benchmarks primarily evaluate visual transformations on silent clips or isolated audio editing, leaving complex audio-visual editing and cross-modal consistency underexplored. We introduce **AVE-Compass**, a comprehensive benchmark with 145 curated source videos, 196 human-verified editing instructions, and 2,688 fine-grained checklist items. It evaluates Instruction Following, Fidelity Preservation, Realism, and Editing Intent through checklist-based MLLM judging and a dedicated realism rubric, complemented by automated cross-modal, video, and audio metrics. Evaluation of six systems shows persistent failures in cross-modal execution and non-target preservation. We further propose **AVE-Agent**, which decomposes instructions into dependent subtasks and improves edits through an evaluator-driven reflection loop. AVE-Agent improves instruction execution, Fidelity Preservation, and audio-visual alignment in joint editing while maintaining competitive perceptual quality.

1 INTRODUCTION

Recently, instruction-guided video editing models have made significant progress, enabling users to modify subjects, actions, styles, and scenes in videos through natural language instructions (Qi et al., 2023; Geyer et al., 2023; Kara et al., 2024; Ku et al., 2024; Team Wan et al., 2025; Jiang et al., 2025). However, real-world videos are not silent visual sequences, as their semantics emerge from visual dynamics together with foreground sounds, background ambience, and speech content (Chen et al., 2020; Cheng et al., 2025; Cao et al., 2026). Realistic editing requests often require synchronized modifications across audio and vision. For example, a model may need to change the corresponding sound when a visible action is modified, remove the acoustic trace when a visual target is removed, or preserve lip motion and background ambience when speech-related content is edited. Handling such instructions requires models to understand implicit cross-modal requirements, maintain audio-visual synchronization, and avoid unintended modifications to non-target visual regions and non-target sounds (Ishii et al., 2025; Xu et al., 2025; Lin et al., 2026).

However, existing benchmarks are insufficient for evaluating this capability in both task design and evaluation protocol. Most video editing benchmarks focus on visual transformations of silent videos, assessing whether models can modify visual attributes, objects, actions, or styles (Sun et al., 2025; Chen et al., 2025b;a; Li et al., 2025b; Wang et al., 2025; Gao et al., 2026; Wu et al., 2025a). Audio-related evaluations, in contrast, usually focus on isolated audio generation, audio replacement, or single synchronization scenarios (Zheng et al., 2025; Feng et al., 2025). More importantly, existing metrics are too coarse-grained to diagnose the failures that truly matter in audio-visual editing (Xu et al., 2021; Elizalde et al., 2023; Huang et al., 2024; Goncalves et al., 2024). Therefore, a benchmark for realistic audio-visual editing should evaluate not only whether the specified edit is completed, but also whether unedited content is faithfully preserved and whether the edited audio and visual signals remain physically and temporally consistent.

To address these problems, we introduce **AVE-Compass**, a comprehensive benchmark for free-form audio-visual editing. AVE-Compass comprises 145 carefully curated source videos and 196 human-verified editing instructions, together with 2,688 fine-grained checklist items. These instructions cover diverse real-world editing scenarios, ranging from local visual or acoustic modifications to joint audio-

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.

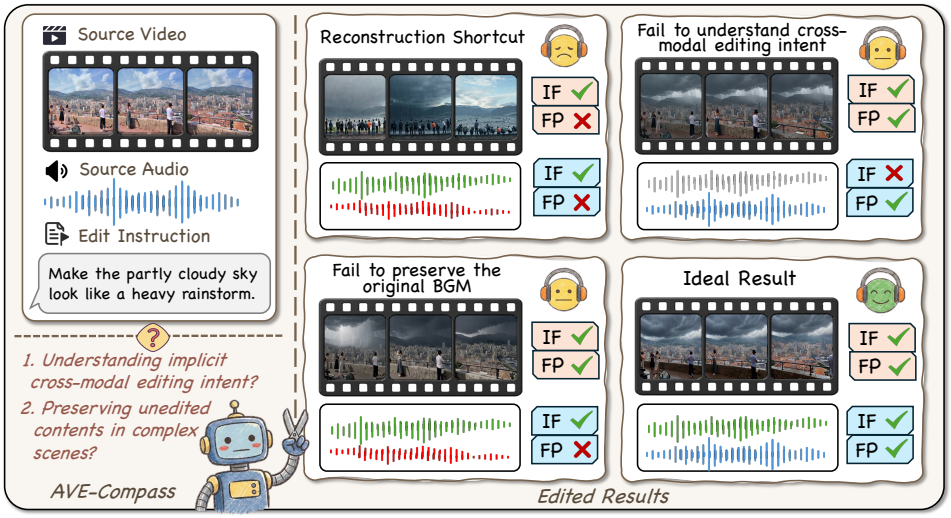


Figure 1: **Illustration of AVE-Compass.** Given the same editing instruction, the four panels illustrate characteristic ways current audio-visual editing models drift, with IF (Instruction Following) and FP (Fidelity Preservation) marks indicating where each output breaks.

visual reconstruction, with a particular focus on models' ability to handle cross-modal dependencies and implicit synchronization constraints. To enable diagnostic evaluation, AVE-Compass decouples model performance into Instruction Following, Fidelity Preservation, Realism, and Editing Intent. Specifically, MLLM-based checklists judge edit execution and non-target content preservation, while a separate MLLM rubric evaluates perceptual realism (Liu et al., 2023b; Zheng et al., 2023). Automated metrics further measure cross-modal, video, and audio quality signals.

Systematic evaluations with AVE-Compass show that current open-source and closed-source systems are still far from reliable audio-visual editing. The most prominent failures occur on the audio side. For example, models tend to incorrectly remove or distort background sounds, or generate audio that is temporally inconsistent with the edited visual events. We argue that existing models lack an explicit mechanism for planning editing operations across different modalities while jointly enforcing fidelity preservation and synchronization constraints. We therefore propose **AVE-Agent**, a planning-based framework that decomposes complex free-form instructions into task topology graphs and refines edits through an evaluator-driven reflection loop. This design improves edit execution, fidelity preservation, and audio-visual synchronization.

The main contributions of this work are summarized as follows:

- We propose **AVE-Compass**, a comprehensive benchmark for free-form audio-visual editing. It contains 145 source videos, 196 human-verified editing instructions, and 2,688 fine-grained checklist items, covering coordinated edits and strict preservation of non-target content.
- We design a decoupled evaluation protocol covering Instruction Following, Fidelity Preservation, Realism, and Editing Intent. With these signals, AVE-Compass can diagnose incomplete edit execution, non-target content drift, low perceptual realism, and audio-visual misalignment.
- We evaluate representative open-source and closed-source editing systems, revealing persistent failures under cross-modal instructions. We further propose **AVE-Agent**, a planning-based audio-visual editing agent framework. Through task topology decomposition and an evaluator-driven reflection loop, AVE-Agent improves editing quality in complex joint editing tasks.

2 Related Work

Audio-Visual Editing Methods. Video editing has evolved from diffusion-based zero-shot methods to instruction-guided and unified generation and editing frameworks. Early methods adapt image generators using attention injection, feature propagation, or temporal consistency constraints (Qi et al., 2023; Ceylan et al., 2023; Geyer et al., 2023; Kara et al., 2024). Instruction-guided methods

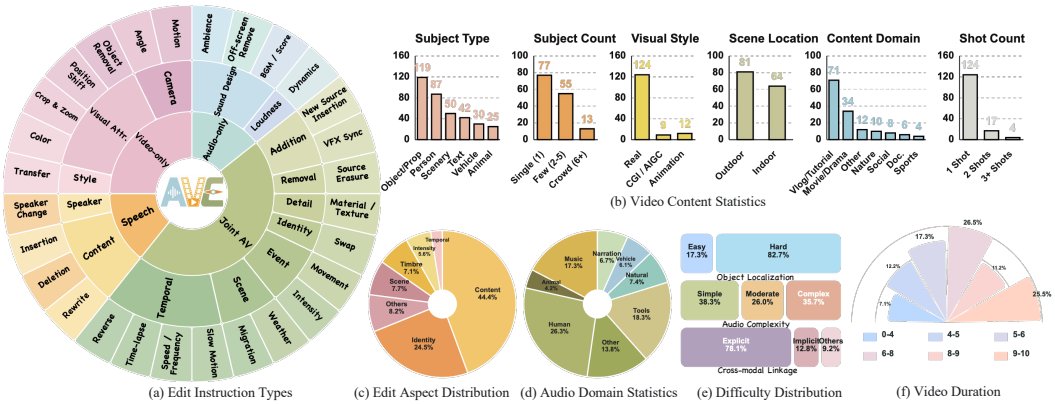


Figure 2: Dataset statistics of AVE-Compass: (a) task taxonomy, (b) source-video diversity, (c,d) edit-aspect and audio-domain distributions, and (e,f) difficulty and duration distributions.

support natural-language edits (Ku et al., 2024; Wu et al., 2025b; Jiang et al., 2025). Audio methods cover text-to-audio generation, video-conditioned generation, and instruction-based editing (Liu et al., 2023a; Cheng et al., 2025; Wang et al., 2023; Lan et al., 2026). These methods largely treat modalities separately, overlooking dependencies between visual events, timing, and sound. Recent audio-visual methods improve cross-modal consistency through sounding-object alignment, instance-level editing, or synchronization (Lin et al., 2026; Fu et al., 2025; Xu et al., 2025; Zheng et al., 2025), but focus on specific objects or sound effects. AVE-Agent supports free-form audio-visual and speech editing through cross-modal planning and synchronization checks.

Audio-Visual Editing Benchmarks. Visual editing benchmarks evaluate changes to subjects, attributes, actions, styles, and visual effects in silent videos, with fine-grained object-level control and multi-dimensional quality assessment (Sun et al., 2025; Chen et al., 2025b;a; Li et al., 2025b; Wang et al., 2025; Gao et al., 2026; Wu et al., 2025a). Audio evaluation focuses on instruction execution, perceptual quality, and preservation of audio characteristics (Jia et al., 2025). Both usually assess modalities separately, limiting evaluation of coordinated editing. Global similarity, quality, and synchronization metrics capture generation quality and audio-visual correspondence (Xu et al., 2021; Elizalde et al., 2023; Huang et al., 2024; Feng et al., 2025), but struggle to assess implicit cross-modal requirements. For example, they cannot reliably determine whether action edits update corresponding sounds, visual-target removal preserves background audio, or speech edits maintain subject and scene consistency. AVE-Compass evaluates edit completion, preservation of non-target visual and auditory content, and audio-visual consistency through fine-grained checklists, realism judgments, and complementary automated metrics.

3 AVE-Compass: Benchmark and Evaluation

Fig. 2 summarizes the statistics of AVE-Compass. The source videos and instructions cover diverse audio-visual content and editing targets, with varied difficulty levels and video durations. Together, these distributions support diagnostic evaluation of free-form audio-visual editing across different source conditions and editing demands.

3.1 Data Construction

Source Video Collection. To construct a diverse source-video pool, we collect candidate samples from existing video datasets (OmniVideoBench (Li et al., 2025a) and UltraVideo (Xue et al., 2025)), in-the-wild web videos, and AIGC-generated videos from contemporary systems. We then remove videos with invalid audio tracks or audio-visual misalignment. We also discard samples with ambiguous subjects, ambiguous sound sources, or unclear editable targets. For the retained samples, we annotate the visual content and shot structure, together with subject information, visual style, and audio sources. After filtering and annotation, AVE-Compass contains **145** source videos.

Table 1: Comparison of AVE-Compass with existing editing benchmarks. **#** denotes counts. **Cross-Modal Evaluation** indicates whether the benchmark evaluates audio–video coupled editing, ranging from object-level correspondence to global-scene dependencies with diagnostic checks.

Benchmark	Target Modality	Task Labels	# Metrics	# Task Categories	Cross-Modal Evaluation	Multi-Shot	Speech Edit	Difficulty Grading
Video-Only Editing Benchmarks								
VEBench Sun et al. (2025)	Video	Object, Scene	3	N/A	✗	✗	✗	✗
IVEBench Chen et al. (2025a)	Video	Object, Scene	12	35	✗	✗	✗	✗
FiVE Li et al. (2025b)	Video	Object	15	6	✗	✗	✗	✗
UniVBench Wei et al. (2026)	Video	Scene	21	6	✗	✓	✗	✗
CoVEBench Wu et al. (2026)	Video	Object, Scene	11	19	✗	✗	✗	✗
Audio-Visual Editing Benchmarks								
SAVEBench Xu et al. (2025)	Video Audio	Object	12	1	Object-level	✗	✗	✗
AVED-Bench Lin et al. (2026)	Video Audio	Object	5	3	Object-level	✗	✗	✗
AVI-Edit Zheng et al. (2025)	Video Audio	Audio Event	7	3	Object-level	✗	✗	✗
AVE-Compass (Ours)	Video Audio	**Object, Scene, Audio Event, Speech**	**19**	**28**	**Global Scene**	✓	✓	✓

Edit Instruction Generation. To ensure broad task coverage, AVE-Compass organizes these instructions into **28** fine-grained operation types across four branches: joint audio-visual, speech, video-only, and audio-only editing (Fig. 2(a)). The video-only and audio-only branches test non-target modality preservation, while the joint audio-visual and speech branches emphasize cross-modal synchronization and speech-related editing. Beyond operation types and audio edit types, each instruction carries three separate difficulty-related labels: object-localization hardness, audio-source complexity, and cross-modal linkage. These labels support stratified analysis across different editing demands, with definitions provided in Appendix A.1.

Covering realistic editing demands, we generate instructions from structured source-video descriptions through a rigorous human-in-the-loop pipeline. First, modality-specific LLM generators propose candidate instructions based on the visual content, audio information, and candidate edit types. A critic model then filters out instructions that are unnatural, infeasible, or insufficiently grounded in the source video. Human annotators perform a final review, and only candidates that pass both stages are retained. This strict selection process yields **196** human-verified audio-visual editing instructions.

Checklist Generation. For fine-grained diagnostic evaluation, we convert each verified instruction into atomic checklist items. Given the source-video description and the editing instruction, checklist generators produce verifiable Yes/No questions that assess both edit execution and the preservation of non-target content. Human annotators then review every checklist item, remove redundant or weakly grounded checks, and revise unclear questions to ensure that each item is specific, verifiable, and aligned with the intended evidence condition. After this careful refinement, the benchmark contains **2,688** checklist items for MLLM-based evaluation. Checklist prompts are in Appendix F.2.

As summarized in Table 1, prior benchmarks largely assess individual modalities or object-level audio-visual operations. AVE-Compass evaluates cross-modal dependencies across diverse edit types, with multi-shot clips, speech editing, and three-axis difficulty stratification.

3.2 Evaluation Methodology

As summarized in Tables 2 and 3, AVE-Compass evaluates edited clips with two complementary components. The MLLM-as-Judge evaluation focuses on semantic edit correctness, source preservation, and perceptual plausibility. The automated metrics provide diagnostic measurements of cross-modal synchronization and visual quality. They also cover temporal consistency, audio quality, and speech quality. We report the two components separately so that instruction-level behavior and low-level audio-visual quality are not conflated.

3.2.1 MLLM-as-Judge Evaluation

We report four dimensions: **Editing Intent (EI)**, **Instruction Following (IF)**, **Fidelity Preservation (FP)**, and **Realism (REAL)**. Each has overall, video, and audio scores (Table 2).

Editing Intent (EI). EI is computed as IF$\times$FP for each case and macro-averaged where both factors are defined. This proxy for joint execution and preservation penalizes non-response, incorrect edits, and non-target changes. It differs from multiplying dataset-level means (Appendix A.3).

Instruction Following (IF). We evaluate IF using atomic Yes/No questions derived from each editing instruction. These questions assess whether the requested modification is correctly applied to the intended target and satisfies the instruction's semantic and spatiotemporal requirements. For example, they check whether a target object is modified at the correct location or a requested sound change occurs at the intended time. For joint audio-visual edits, IF additionally verifies that both modalities are edited consistently with the coupled relationship expressed in the instruction.

Fidelity Preservation (FP). We evaluate FP using checklist questions that compare the source and edited videos from two perspectives. First, preservation questions examine whether non-edited visual and auditory content remains consistent with the source. Second, unintended-content questions assess whether the result introduces unrequested content, such as subtitles or background music. The evaluator ignores trivial differences and direct consequences of the requested edit, but penalizes noticeable changes to retained content and clearly unrequested additions.

Realism (REAL). We evaluate the naturalness and coherence of each edited clip using a separate MLLM rubric. Unlike edit-execution checklists, REAL assesses whether the result appears plausible on its own. On the video side, it examines physical consistency and whether edited elements blend naturally into the scene. On the audio side, it examines signal quality and whether acoustic properties match the visual content and scene context. Checklist and judging details are in Appendix A.3.

3.2.2 Automated Evaluation

Complementing MLLM-as-Judge evaluation, seven automated metrics provide diagnostic scalar measurements across Cross-Modal, Video, and Audio groups. Scores are normalized so larger values indicate better performance. Appendix A.2 provides category-specific activation and formulas.

Cross-Modal Metrics. We assess temporal alignment between the edited audio and visual streams using two metrics. For speech-bearing edits, **Lip Sync** measures whether the edited speech remains synchronized with visible mouth motion using SyncNet confidence. For edits in which sound events must correspond temporally to visual events, **AV Sync (AVS)** uses Synchformer to estimate the offset between the two modalities and converts it into a normalized score under a fixed tolerance window.

Video Metrics. We evaluate the edited video in terms of visual quality, appearance consistency, and temporal coherence. **Video Aesthetic (V-AES)** averages aesthetic scores over sampled frames to measure the overall visual appeal of the clip. **Subject Consistency (SC)** measures adjacent-frame DINOv2 feature similarity within the edited video, rather than preservation relative to the source. **Motion Smoothness (MSM)** evaluates whether motion evolves smoothly over time based on frame-interpolation consistency between adjacent frames.

Audio Metrics. We evaluate the perceptual quality of the edited audio and its speech content. **Audio Aesthetic (A-AES)** measures the naturalness and production quality of the complete audio track. For speech-bearing edits, **Speech Quality (SF)** further evaluates speech intelligibility and naturalness. It predicts a MOS score for the edited speech signal and normalizes the result to $[0, 1]$.

4 Audio-Visual Editing Agent

We build **AVE-Agent**, a modular pipeline that compiles a free-form natural-language instruction into an edited audio-visual clip. Its design makes cross-modal dependencies explicit and uses local and global feedback to correct errors (Fig. 3). Implementation details are deferred to Appendix B.

Planner agent. The planner agent converts the instruction into an executable and verifiable editing plan grounded in the source clip. It first builds a structured profile of the visual and audio content, then identifies both explicit requirements and their unstated cross-modal consequences. For example,

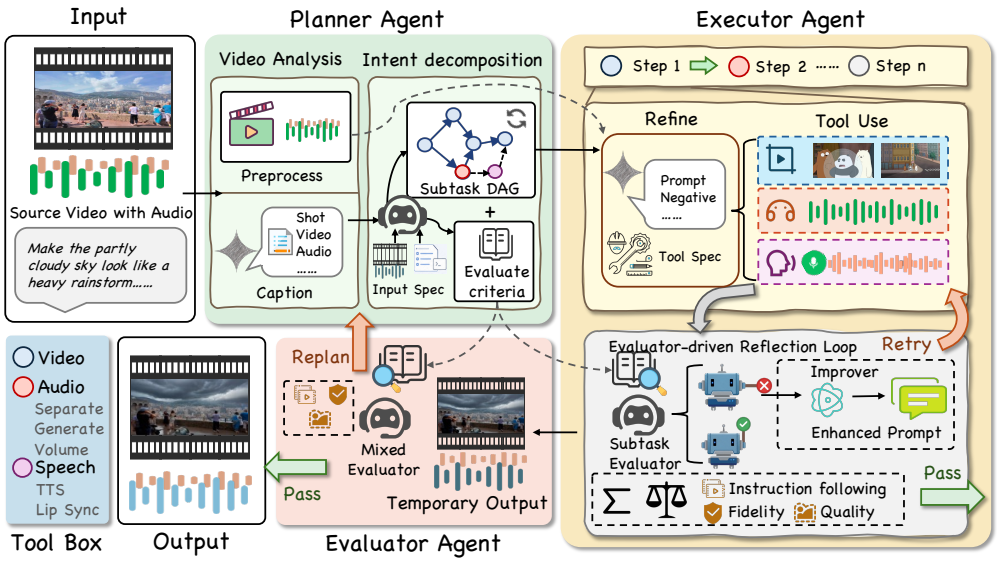


Figure 3: **AVE-Agent pipeline.** Given a source clip and an editing instruction, the planner constructs a subtask graph with cross-modal dependencies and evaluation criteria. The executor coordinates modality-specific tools through an evaluator-driven reflection loop. The mixed evaluator checks the assembled clip and returns either the final output or feedback for correction or replanning.

modifying a visible action may also require updating its associated sound. The planner decomposes these requirements into modality-specific subtasks and connects them with dependency edges that encode execution order and audio-visual relations, forming a directed acyclic graph (DAG). Each subtask is paired with evaluation criteria that guide subsequent execution and verification. Before execution, the plan is checked for completeness, executability, and cross-modal consistency. Missing dependencies or structural conflicts trigger replanning.

Executor agent with an evaluator-driven reflection loop. The executor follows the DAG, routes each subtask to the appropriate video, audio, or speech tools, and translates its intent into tool-specific guidance. A subtask evaluator checks the result against the planner's criteria (Madaan et al., 2023). The feedback specifies what is missing, what was added unintentionally, and which source content was altered, giving the improver a concrete basis for correction. When a check fails, an improver uses the identified errors to revise the guidance for the next attempt while preserving the original intent and constraints. The loop alternates evaluation and targeted correction within its retry budget, retaining each subtask's highest-scoring candidate.

Mixed evaluator agent. Locally valid edits can still conflict after assembly. For example, a generated sound effect may obscure dialogue. The mixed evaluator therefore checks the complete clip for instruction following, fidelity preservation, and overall quality. Its feedback determines the correction: mixing errors trigger remixing, faulty audio content triggers subtask regeneration, and inconsistent cross-modal dependencies trigger replanning with the feedback as context. The pipeline retains the highest-scoring complete output across corrections.

5 Experiments

5.1 Main Results

We evaluate AVE-Agent (Wan) and five baselines—Wan2.7 (Alibaba Cloud, 2026a), HappyHorse (Alibaba Cloud, 2026b), Gemini-Omni (Google DeepMind, 2026), Seedance (Team Seedance et al., 2026), and LTX2.3-Retake (LTX2) (Lightricks, 2026)—on AVE-Compass. AVE-Compass evaluation samples and checklists were not used to develop or tune AVE-Agent. Tables 2 and 3 report the MLLM-as-Judge and automated metric results, respectively. The benchmark's 196 instructions cover

Table 2: **MLLM-as-Judge results** on AVE-Compass. Editing Intent, Instruction Following, Fidelity Preservation, and Realism are reported as Overall / Video / Audio scores on a 0–100 scale. Models are ranked by Overall Editing Intent, which jointly assesses instruction execution and non-target preservation. Higher is better. Best in **bold**. Signed superscripts denote differences from Wan2.7. *Gemini misses 16 speech edits due to content moderation.

Model	Editing Intent			Instruction Following			Fidelity Preservation			Realism		
	Overall	Video	Audio	Overall	Video	Audio	Overall	Video	Audio	Overall	Video	Audio
AVE-Agent (Wan)	**59.8**$^{+17.4}$	**66.7**$^{+6.6}$	**50.2**$^{+25.4}$	**77.3**$^{+8.0}$	**84.8**$^{+6.5}$	**69.4**$^{+9.1}$	77.6$^{+13.4}$	80.1$^{+1.6}$	74.8$^{+26.7}$	62.1$^{+1.7}$	45.3$^{+1.8}$	78.9$^{+1.5}$
Wan2.7	42.4	60.1	24.8	69.3	78.3	60.3	64.2	78.5	48.1	60.4	43.5	77.4
HappyHorse	41.3	56.7	18.8	66.9	75.5	54.5	63.9	74.5	53.3	63.0	49.9	76.0
Gemini-Omni*	38.0	56.1	10.0	44.9	74.0	10.8	**84.9**	76.8	**96.1**	66.9	49.2	84.5
Seedance	26.6	36.1	13.5	37.4	50.5	24.4	81.7	**83.0**	80.4	**69.2**	**54.0**	84.4
LTX2	15.2	10.7	26.4	70.1	72.8	66.2	30.6	24.5	42.3	64.1	42.0	**86.2**

Table 3: **Automated metric results** on AVE-Compass (higher is better). Best in **bold**. Signed superscripts denote differences from Wan2.7. *Gemini misses 16 speech edits due to moderation. †Speech Quality and Lip Sync use speech-category edits only.

Model	Cross-Modal		Video			Audio	
	Lip Sync†	AV Sync	Video Aesthetic	Subject Consistency	Motion Smoothness	Audio Aesthetic	Speech Quality†
AVE-Agent (Wan)	**0.622**$^{+0.091}$	**0.766**$^{+0.073}$	0.452$^{+0.001}$	0.972$^{+0.003}$	0.987$^{+0.001}$	0.614$^{+0.005}$	0.368$^{-0.060}$
Wan2.7	0.531	0.693	0.451	0.969	0.986	0.609	0.428
HappyHorse	0.620	0.695	0.439	**0.975**	**0.989**	**0.650**	**0.726**
Gemini-Omni*	–	0.701	0.434	0.974	0.988	0.627	–
Seedance	0.431	0.718	0.430	0.971	0.987	0.629	0.388
LTX2	0.618	0.758	**0.468**	0.968	0.986	0.641	0.522

the operation categories and difficulty levels in our taxonomy, supporting diagnostic evaluation. Appendix C.8 discusses the trade-off between benchmark scale and evaluation cost for routine use.

The results reveal three main trends. (1) **AVE-Agent leads in edit execution.** It achieves the highest reported Editing Intent and Instruction Following scores, along with the best Lip Sync and AV Sync scores. Compared with baseline, it improves both instruction following and preservation, with particularly large gains in audio-side Editing Intent and Fidelity Preservation. (2) **Current models struggle to balance Instruction Following and Fidelity Preservation.** LTX2 follows many instructions but largely regenerates the input, damaging non-target content. Conversely, Gemini-Omni and Seedance sometimes return the original video or audio, producing high preservation scores without executing the edit. These contrasting failures motivate assessing both requirements through Editing Intent. The baseline scores in Table 2 show that achieving both remains challenging. (3) **Audio-visual realism remains limited.** Although some models achieve high Video Aesthetic or Audio Aesthetic scores, their Realism scores remain noticeably lower than their automated quality scores. Automated metrics primarily measure technical quality, while Realism evaluates logical validity and audio-visual naturalness: smooth, appealing outputs may still contain deformed or disappearing objects and unnatural blending. This discrepancy suggests limitations in current models' physical-world understanding for editing.

5.2 Further Analysis

Response-Gated Evaluation. Some methods leave a target modality unchanged for certain instructions, potentially inflating its Fidelity Preservation and Realism scores. We therefore report response-gated scores in Table 4, alongside ungated scores and modality-specific response rates. Since responding subsets differ, gated scores diagnose the effect of non-response within a model and do not provide a controlled ranking across models. Gemini's low audio response accompanies a marked drop in audio preservation after gating, while Seedance's lower visual response accompanies decreases in visual preservation and Realism. Our Editing Intent metric therefore provides a more faithful assessment of overall editing performance.

Table 4: **Response-gated Fidelity Preservation and Realism** by modality. "All" uses available cases. "Gated" retains cases with a change in the target modality for within-model diagnosis. See Appendix A.3 for details. Higher is better. **Bold** marks the lowest response rates.

Model	Response Rate		Fidelity Preservation (all)		Fidelity Preservation (gated)		Realism (all)		Realism (gated)	
	Video	Audio	Video	Audio	Video	Audio	Video	Audio	Video	Audio
AVE-Agent (Wan)	92.7	94.4	80.1	74.8	77.9	70.9	45.3	78.9	43.0	76.6
Wan2.7	90.6	95.6	78.5	48.1	76.3	44.8	43.5	77.4	40.4	75.1
HappyHorse	95.8	74.4	74.5	53.3	73.1	31.7	49.9	76.0	48.8	73.2
Gemini-Omni	89.7	**22.0**	76.8	96.1	74.2	80.6	49.2	84.5	46.8	81.2
Seedance	**69.1**	47.8	83.0	80.4	74.9	60.1	54.0	84.4	48.5	82.8
LTX2	90.1	91.1	24.5	42.3	16.8	39.0	42.0	86.2	36.7	84.4

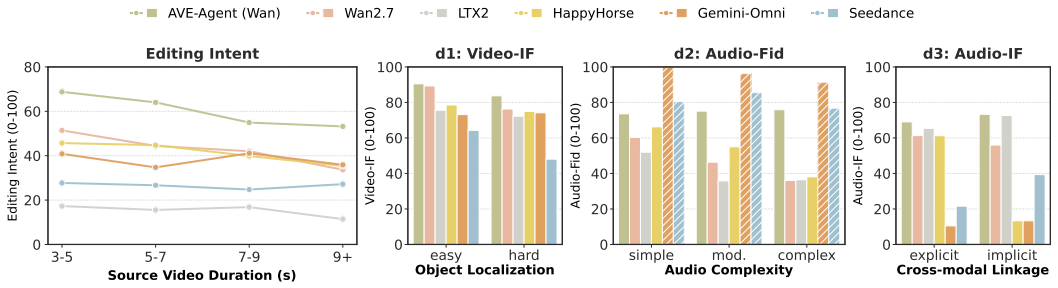


Figure 4: **Robustness under difficulty.** Left: Editing Intent across source-video duration. Right: object-localization hardness, audio-source complexity, and cross-modal linkage evaluated with modality-matched metrics. Hatched audio-source complexity bars mark Audio Fidelity Preservation inflated by audio non-response. Scores are on a 0–100 scale.

Robustness under Difficulty. Figure 4 examines source-video duration and difficulty (Appendix A.1), with full results in Appendix C.3. AVE-Agent maintains the highest Editing Intent across duration groups, although its scores decline for longer videos. Harder object localization lowers Video-IF for most models, whereas audio complexity and cross-modal linkage yield model-specific trends. Wan2.7 and HappyHorse struggle to preserve non-target sounds in complex audio. High FP can also reflect low response rates. AVE-Agent maintains Audio Fidelity Preservation across audio-complexity levels and strong Audio-IF under implicit requirements.

Single-Modality Editing. We evaluate single-modality editing, where only the target stream should be modified and the other modality should remain unchanged. The middle panel of Figure 5 presents the main results, with full results in Appendix C.4. To measure preservation of the non-target modality, we report automated Audio Similarity and Video Similarity scores. For video-only edits, AVE-Agent and HappyHorse both perform strongly by modifying the visual content while retaining the source audio. For audio-only edits, AVE-Agent achieves the best overall performance, successfully following the audio instruction without altering the video.

Error Analysis and Case Study. Figure 1 and Appendix Figure 7 illustrate three recurring failures: non-response, unintended changes to non-target content, and implausible or unrequested content. The right panel of Figure 5 and Appendix Table 9 complement these cases with counts across five non-exclusive error categories. Gemini-Omni and Seedance exhibit frequent Audio-IF failures associated with non-response, while LTX2's extensive regeneration is accompanied by preservation failures in both modalities. Gemini's low Audio-FP error count reflects non-response, not preservation quality. AVE-Agent has the lowest total error count among the evaluated outputs. Reliable editing requires accurate execution, content preservation, and audio-visual plausibility.

Metric Validity Evaluation. We evaluate metric validity from three complementary perspectives: repeated-run stability, consistency with human judgments, and inter-metric correlations. On a subset of 50 editing tasks, five scoring runs across all six models yield variation below 0.01 for automated scores and 1% for MLLM-as-Judge scores. Agreement with sampled human annotations averages approximately 90% (Appendix Table 6). Spearman correlations in Figure 5 (left) indicate limited association between automated and MLLM-as-Judge scores, with partial overlap among automated

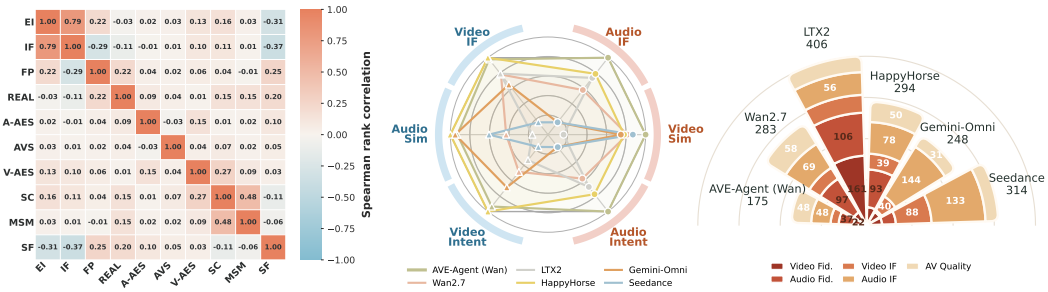


Figure 5: **Metric correlations, single-modality editing, and error analysis.** Left: Spearman correlations among MLLM-as-Judge and automated metrics. Middle: video-only and audio-only results, with axes normalized per metric. Right: per-model counts for five error categories.

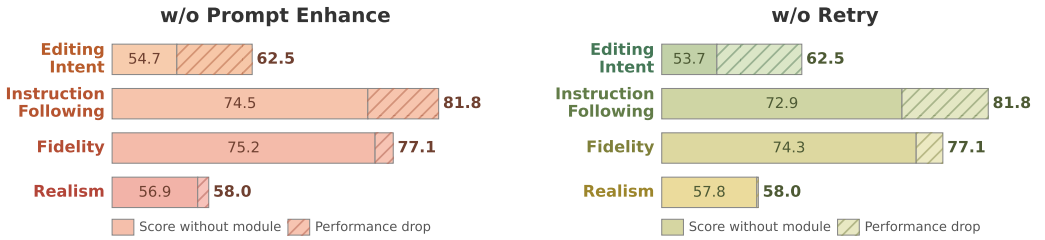


Figure 6: Ablation of planner-side prompt enhancement and evaluator-driven reflection loop. Solid bars show the score without the module. Hatched segments show the drop from the full AVE-Agent.

metrics. These findings support the stability and complementary diagnostic value of our metrics and motivate our multi-dimensional evaluation framework rather than reliance on a single aggregate score. Detailed analyses are provided in Appendices C.1 and C.2.

5.3 Ablation Studies

Figure 6 separately ablates two agent components: planner-side prompt enhancement and the evaluator-driven reflection loop on the subset. Removing prompt enhancement decreases Editing Intent by 7.8 and Instruction Following by 7.3, highlighting the importance of translating user intent into executable tool prompts before invoking specialized editors. The accompanying decline in Realism also indicates that prompt enhancement improves output quality. Disabling retries yields larger drops of 8.8 and 8.9, respectively, as failed or incomplete edits can no longer be revised using evaluator feedback. Fidelity Preservation also drops, suggesting that the loop helps maintain non-target content during execution. Overall, prompt enhancement improves the initial tool invocation, while the evaluator-driven reflection loop provides complementary correction through feedback during execution. Appendix Tables 10 and 11 report backbone ablations.

6 Conclusion

We introduce AVE-Compass, a benchmark for free-form audio-visual editing that combines human-verified instructions and fine-grained checklists with a dedicated realism rubric and complementary automated metrics. Its evaluation protocol assesses Instruction Following, Fidelity Preservation, Realism, and Editing Intent, enabling diagnosis of incomplete edits, non-target changes, and audio-visual inconsistencies. Evaluation of six systems reveals that high preservation or perceptual quality scores alone do not ensure successful editing. We further propose AVE-Agent, which models cross-modal dependencies through structured task planning and coordinates modality-specific tools through an evaluator-driven reflection loop. AVE-Agent achieves the highest reported Editing Intent and Instruction Following scores, while improving preservation and synchronization over its Wan2.7 backbone. Ablations support the contributions of prompt enhancement and evaluator feedback. Together, they support systematic evaluation and the development of reliable audio-visual editing.

AI USE STATEMENT

AI-assisted writing tools were used to polish the manuscript's language, including grammar, sentence structure, readability, and clarity of presentation. This writing assistance does not constitute a scientific contribution or authorship. The authors retain full responsibility for the accuracy, originality, citations, and conclusions of the paper, including any wording refined with AI assistance. AI tools were also used to generate synthetic data for AVE-Compass, as described in Section 3.1. The use of multimodal language models as evaluators and as components of AVE-Agent is described separately in the methodology and implementation sections.

ETHICS STATEMENT

AVE-Compass and AVE-Agent are intended to support research on reliable audio-visual editing. We recognize that improved editing capabilities can also facilitate deceptive media, impersonation, and non-consensual voice manipulation. The benchmark uses publicly licensed and AI-generated clips. AVE-Agent includes voice-editing components, including reference-conditioned voice synthesis, but our benchmark and experiments do not target real-person impersonation, and we do not release identity-specific cloning assets. Downstream use should respect licenses, privacy, and consent, with provenance or disclosure for potentially misleading edits. Potential evaluator biases warrant interpreting judgments alongside the human-consistency assessment in Appendix C.2. Broader benefits, risks, and mitigations are discussed in Appendix D.

REPRODUCIBILITY STATEMENT

The supplementary material includes a 50-instruction dataset demo covering all 28 editing operation types, together with links to the full dataset and code. We document benchmark construction in Section 3.1, evaluation methodology in Section 3.2, and metric definitions in Appendices A.2 and A.3. Appendix B specifies the agent components, tool registry, and evaluator-driven reflection loop, while Appendix F provides the captioning, checklist-generation, and evaluation prompts. Appendix C supplies the human-consistency calculation, difficulty-stratified results, single-modality evaluation, error analysis, and agent-backbone ablations, including the subset description. These details support implementation and verification of the reported evaluation protocol. Exact numerical replication can be affected by stochastic model outputs and changes to third-party APIs. The associated practical limitations and evaluation costs are discussed in Appendices E and C.8, respectively.

REFERENCES

Alibaba Cloud. Wan 2.7 video editing api reference. `https://www.alibabacloud.com/help/en/model-studio/wan-video-editing-api-reference`, 2026a. Accessed: 2026-09-26.

Alibaba Cloud. HappyHorse 1.0 video editing. `https://www.alibabacloud.com/help/en/model-studio/happyhorse-1-0-video-edit`, 2026b. Accessed: 2026-09-26.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jiahao Wang, Jialu Chen, Miao Deng, Yubin Guo, Chenxi Liao, Yize Zhang, Zhaoxiang Zhang, and Jiaheng Liu. T2AV-Compass: Towards unified evaluation for text-to-audio-video generation. In *International Conference on Machine Learning (ICML)*, 2026. URL `https://arxiv.org/abs/2512.21094`.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23206–23217, 2023.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. Ivebench: Modern benchmark suite for instruction-guided video editing assessment. *arXiv preprint arXiv:2510.11647*, 2025a.

Yupeng Chen, Penglin Chen, Xiaoyu Zhang, Yixian Huang, and Qian Xie. Editboard: Towards a comprehensive evaluation benchmark for text-based video editing models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):15975–15983, 2025b.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28901–28911, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, pp. 251–263. Springer, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

discus0434. Aesthetic predictor V2.5: SigLIP-based aesthetic score predictor. https://github.com/discus0434/aesthetic-predictor-v2-5, n.d. Software documentation, accessed: 2026-09-26.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Tao Feng, Yifan Xie, Xun Guan, Jiyuan Song, Zhou Liu, Fei Ma, and Fei Yu. Unisync: a unified framework for audio-visual synchronization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Youquan Fu, Ruiyang Si, Hongfa Wang, Dongzhan Zhou, Jiacheng Sun, Ping Luo, Di Hu, Hongyuan Zhang, and Xuelong Li. Object-avedit: An object-level audio-visual editing model. *arXiv preprint arXiv:2510.00050*, 2025.

Xiangbo Gao, Sicong Jiang, Bangya Liu, Xinghao Chen, Minglai Yang, Siyuan Yang, Mingyang Wu, Jiongze Yu, Qi Zheng, Haozhi Wang, et al. Vefx-bench: A holistic benchmark for generic video editing and visual effects. *arXiv preprint arXiv:2604.16272*, 2026.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Lucas Goncalves, Prashant Mathur, Chandrashekhar Lavania, Metehan Cekic, Marcello Federico, and Kyu J Han. Perceptual evaluation of audio-visual synchrony grounded in viewers' opinion scores. In *European Conference on Computer Vision*, pp. 288–305. Springer, 2024.

Google DeepMind. Gemini omni. https://deepmind.google/models/gemini-omni/, 2026. Accessed: 2026-07-03.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489. URL https://arxiv.org/abs/2401.16423.

Masato Ishii, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. Coherent audio-visual editing via conditional audio generation following video edits. *arXiv preprint arXiv:2512.07209*, 2025.

Yuhang Jia, Hui Wang, Xin Nie, Yujie Guo, Lianru Gao, and Yong Qin. Towards automatic evaluation and high-quality pseudo-parallel dataset construction for audio editing: A human-in-the-loop method. *arXiv preprint arXiv:2508.11966*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6507–6516, 2024.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. AnyV2V: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. doi: 10.48550/arXiv.2403.14468. URL https://arxiv.org/abs/2403.14468.

Zitong Lan, Yiduo Hao, and Mingmin Zhao. Smartdj: Declarative audio editing with audio language model. In *The Fourteenth International Conference on Learning Representations*, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025a.

Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. Five: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. *arXiv preprint arXiv:2503.13684*, 2025b.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. AMT: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_AMT_All-Pairs_Multi-Field_Transforms_for_Efficient_Frame_Interpolation_CVPR_2023_paper.html.

Lightricks. LTX-2.3: Video editing with retake. https://ltx.io/model/ltx-2-3, 2026. Accessed: 2026-09-26.

Yan-Bo Lin, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, Xiaofei Wang, Gedas Bertasius, and Lijuan Wang. Zero-shot audio-visual editing via cross-modal delta denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7344–7354, 2026.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023a.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE Transactions on Audio, Speech and Language Processing*, 33:458–471, 2025. doi: 10.1109/TASLP.2024.3520017. URL https://ieeexplore.ieee.org/document/10819011/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023b.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, pp. 2127–2131, 2021. doi: 10.21437/Interspeech.2021-299. URL https://www.isca-archive.org/interspeech_2021/mittag21_interspeech.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. doi: 10.48550/arXiv.2304.07193. URL https://arxiv.org/abs/2304.07193.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Qwen. Qwen3-TTS: Official implementation. https://github.com/QwenLM/Qwen3-TTS, 2026. Accessed: 2026-09-26.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

Shangkun Sun, Xiaoyu Liang, Songlin Fan, Wenxu Gao, and Wei Gao. Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7105–7113, 2025.

Sync Labs. Lipsync models: lipsync-2. https://sync.so/docs/models/lipsync, 2026. Online documentation, accessed: 2026-09-26.

Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. doi: 10.48550/arXiv.2604.14148. URL https://arxiv.org/abs/2604.14148.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025. doi: 10.48550/arXiv.2502.05139. URL https://arxiv.org/abs/2502.05139.

Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. VideoCLIP-XL: Advancing long description understanding for video CLIP models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16061–16075, 2024. doi: 10.18653/v1/2024.emnlp-main.898. URL https://aclanthology.org/2024.emnlp-main.898/.

Juntong Wang, Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Tdve-assessor: Benchmarking and evaluating the quality of text-driven video editing with lmms. *arXiv preprint arXiv:2505.19535*, 2025.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

Jianhui Wei, Xiaotian Zhang, Yichen Li, Yuan Wang, Yan Zhang, Ziyi Chen, Zhihang Tang, Wei Xu, and Zuozhu Liu. Univbench: Towards unified evaluation for video foundation models. *arXiv preprint arXiv:2602.21835*, 2026. doi: 10.48550/arXiv.2602.21835.

Jay Zhangjie Wu, Guian Fang, Dongrong Joe Fu, Vijay Anand Raghava Kanakagiri, Forrest Iandola, Kurt Keutzer, Wynne Hsu, Zhen Dong, and Mike Zheng Shou. Veditbench: Holistic benchmark for text-guided video editing, 2025a. URL https://openreview.net/forum?id=6325Jzc9eR.

Jiangtao Wu, Jiaming Wang, Yiwen He, Yuanxing Zhang, Shihao Li, Dunyuan Liu, Xuedong Zhao, Jialu Chen, Zekun Moore Wang, and Jiaheng Liu. Covebench: Can video editing models handle complex instructions? *arXiv preprint arXiv:2606.08415*, 2026. doi: 10.48550/arXiv.2606.08415.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16692–16701, 2025b.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6787–6800, 2021.

Weihan Xu, Kan Jen Cheng, Koichi Saito, Muhammad Jehanzeb Mirza, Tingle Li, Yisi Liu, Alexander H Liu, Liming Wang, Masato Ishii, Takashi Shibuya, et al. Schrodinger audio-visual editor: Object-level audiovisual removal. *arXiv preprint arXiv:2512.12875*, 2025.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

Haojie Zheng, Shuchen Weng, Jingqi Liu, Siqi Yang, Boxin Shi, and Xinlong Wang. AVI-Edit: Audio-sync video instance editing with granularity-aware mask refiner. *arXiv preprint arXiv:2512.10571*, 2025. URL https://arxiv.org/abs/2512.10571.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Benchmark Construction Details

This appendix defines the difficulty annotations and reports the automated metric and MLLM-as-Judge formulas used by AVE-Compass. The latest prompt templates for source-video captioning, checklist generation, checklist evaluation, and realism evaluation are collected separately in Appendix F.

A.1 Difficulty Annotation Criteria

We characterize editing difficulty along three complementary axes.

D1: Object-localization hardness. *Easy* covers edits requiring no local visual-target localization, or close-ups where the target occupies at least 80% of the frame with minimal background. Other edits requiring a local target to be identified, isolated, or disambiguated are *hard*.

D2: Audio-source complexity. Based on source-audio metadata, *simple* denotes one foreground sound category without background music (BGM) or voiceover. The *moderate* category denotes two foreground categories without these background layers, or BGM alone. The *complex* category denotes at least three foreground categories, or foreground sounds combined with BGM or voiceover.

D3: Cross-modal linkage. *Explicit* instructions state the required audio changes. *Implicit* instructions entail unstated cross-modal consequences necessary for physical or semantic coherence. Instruction Following checks these necessary consequences, excluding optional effects. Video-only edits without cross-modal requirements are labeled *none*.

A.2 Automated Metric Formulas

This appendix details the automated metric branch of AVE-Compass. Let the source video be $x = (v, a)$, where v and a denote the visual and audio streams, and let the edited output be $\hat{x} = (\hat{v}, \hat{a})$ under instruction p. Each sample is assigned one instruction category $c \in \{\text{audio-only}, \text{video-only}, \text{joint}, \text{speech}\}$. The category determines which automated metrics are active. Inactive metrics are excluded rather than assigned zero, so that a video-only edit is not penalized by audio-target metrics and an audio-only edit is not penalized by visual-target metrics. Unless otherwise noted, each score is oriented so that a larger value is better. Predictor scores with bounded native ranges are mapped to $[0, 1]$. Similarity scores based on correlations or cosine similarities are also reported as larger-is-better scores, with affine mapping applied when the native score lies in $[-1, 1]$.

Table 5: Reference-free objective score routing in AVE-Compass. Each score is activated only for the instruction categories where it provides relevant evidence.

Metric Group	Score Name	Active Categories
Cross-Modal	lip_sync	speech
	av_sync	audio-only, joint, speech
Video	video_aesthetic	video-only, joint, speech
	subject_consistency	video-only, joint, speech
	motion_smoothness	video-only, joint, speech
Audio	audio_aesthetic	audio-only, joint, speech
	speech_quality	speech

Table 5 lists the seven reference-free automated scores reported in the main experiment table. Additional source-target similarity scores for single-modality preservation are described after the table.

For single-modality diagnostics, AVE-Compass additionally reports preservation similarities on the non-target stream: `videoclip_video_similarity` for audio-only edits and `audio_similarity` for video-only edits.

Audio aesthetic. `audio_aesthetic` evaluates the edited audio with AudioBox Aesthetics (Tjandra et al., 2025) after conversion to 16 kHz, mono, PCM 16-bit WAV. We average its Perceptual Quality (PQ) and Content Understanding (CU) scores and normalize by 10:

$$S_{\mathrm{AA}} = \frac{\mathrm{PQ} + \mathrm{CU}}{20}. \tag{1}$$

Speech quality. For `speech` edits, `speech_quality` uses NISQA (Mittag et al., 2021) to predict the edited speech's mean opinion score (MOS), mapped from $[1, 5]$ to $[0, 1]$:

$$S_{\mathrm{speech}} = \mathrm{clip}\left(\frac{\mathrm{MOS} - 1}{4}, 0, 1\right). \tag{2}$$

Lip synchronisation. For `speech` edits, `lip_sync` uses SyncNet (Chung & Zisserman, 2016) to compare visual lip and speech features across temporal shifts. Its confidence $\kappa = \mathrm{median_dist} - \mathrm{min_dist}$ is normalized by threshold γ:

$$S_{\mathrm{lip}} = \mathrm{clip}\left(\frac{\kappa}{\gamma}, 0, 1\right), \qquad \gamma = 2.0. \tag{3}$$

Audio–visual synchronisation. `av_sync` uses Synchformer (Iashin et al., 2024) to predict the audio–video offset Δt in seconds and converts it to a normalized score using scale τ, specified in the evaluation configuration:

$$S_{\mathrm{sync}} = \max\left(0,\ 1 - \frac{|\Delta t|}{\tau}\right). \tag{4}$$

Video aesthetic. `video_aesthetic` uses the SigLIP-based Aesthetic Predictor v2.5 (discus0434, n.d.) to score $T = 10$ uniformly sampled frames from $\hat{v}$, then averages and normalizes the scores:

$$S_{\mathrm{VA}} = \frac{1}{10T} \sum_{i=1}^{T} q_i, \qquad T = 10, \tag{5}$$

where q_i is the raw aesthetic score of the i-th sampled frame.

Subject consistency. `subject_consistency` measures whether the main subject appearance remains stable across frames in the edited video. It is computed on $\hat{v}$ only and does not use the source video. The evaluator uniformly samples at most 50 frames, resizes each frame to 224×224,

normalizes it with ImageNet statistics, and extracts DINOv2 features (Oquab et al., 2023). For each adjacent frame pair, it computes the cosine similarity and then maps the mean adjacent-frame cosine from $[-1, 1]$ to $[0, 1]$:

$$\bar{c}_{\text{adj}} = \frac{1}{T-1} \sum_{i=2}^{T} \cos(f_i, f_{i-1}), \qquad S_{\text{subj}} = \text{clip}_{[0,1]} \left(\frac{\bar{c}_{\text{adj}} + 1}{2} \right). \tag{6}$$

Motion smoothness. `motion_smoothness` detects unnatural motion, flicker, or temporal discontinuities in the edited video. The evaluator extracts frames from $\hat{v}$ and uses the even-indexed frames $(0, 2, 4, \ldots)$ as inputs to AMT-G (Li et al., 2023), a video frame interpolation model. AMT-G predicts the missing middle frames at temporal midpoint 0.5. These predicted odd frames are compared with the actual odd-indexed frames $(1, 3, 5, \ldots)$ by pixel-wise mean absolute difference:

$$d_i = \text{mean} \left(|I_i^{\text{interp}} - I_i^{\text{real}}| \right), \qquad d = \frac{1}{N} \sum_{i=1}^{N} d_i. \tag{7}$$

The raw difference d lies on the 8-bit pixel scale and is converted to a larger-is-better score:

$$S_{\text{motion}} = \frac{255 - d}{255}. \tag{8}$$

VideoCLIP video similarity. `videoclip_video_similarity` is used for audio-only edits, where the visual stream should remain unchanged while the audio is edited. The evaluator uniformly samples 8 frames from both v and $\hat{v}$, resizes them to 224×224, normalizes them with ImageNet statistics, and feeds each 8-frame tube into the VideoCLIP-XL visual encoder (Wang et al., 2024). The source and edited video embeddings are L2-normalized and compared by cosine similarity, then mapped to $[0, 1]$:

$$c_{\text{vclip}} = \cos \left(\phi_v(v), \phi_v(\hat{v}) \right), \qquad S_{\text{vclip}} = \text{clip}_{[0,1]} \left(\frac{c_{\text{vclip}} + 1}{2} \right). \tag{9}$$

Spectral audio similarity. `audio_similarity` is used for video-only edits, where the audio track should remain stable while the visual stream is edited. We convert the source and edited audio tracks to 16 kHz mono PCM WAV and compute STFT magnitude spectrograms $M(a)$ and $M(\hat{a})$. The default window size is 1024 samples with a half-window hop. The evaluator searches over spectrogram-frame lags within a bounded window specified in the evaluation configuration, while requiring sufficient overlap between the two clips.

For a candidate lag δ, both overlapping spectrogram regions are flattened and z-score normalized before Pearson correlation is computed. The best lag δ^* maximizes this spectral correlation:

$$\delta^* = \arg\max_{\delta} \rho \left(M(a), M(\hat{a})_{\delta} \right), \qquad S_{\text{audio}} = \text{clip} \left(\frac{1 + \rho \left(M(a), M(\hat{a})_{\delta^*} \right)}{2}, 0, 1 \right). \tag{10}$$

If both extracted tracks are near-silent, the score is 1.0. The score measures spectral content preservation after alignment, without penalizing the alignment shift.

A.3 MLLM-as-Judge Evaluation Formulas

The MLLM-as-Judge protocol in Sec. 3.2.1 uses atomic Yes/No questions for **Instruction Following** and **Fidelity Preservation**, plus separate **Edit Response** questions. Realism is evaluated independently using the prompt in Appendix F.4.

Evaluator model. All benchmark checklist evaluators, covering Instruction Following, Fidelity Preservation, and Edit Response across the video-only, audio-only, joint audio-visual, and speech branches, use Gemini 3.1 Pro. The standalone Realism evaluator uses the same model.

Atomic checklist generation. Separate generators for video-only, audio-only, joint audio-video, and speech edits produce atomic checklists grounded in the source caption and edit prompt. Instruction Following checks the requested changes, while Fidelity Preservation checks non-edited content and unrequested hallucinations. Each item has dimension, subdimension, and modality tags. Checklist length adapts to the informative checks needed for each edit.

Edit Response questions. Edit Response checks whether a visible or audible change occurred, irrespective of correctness. Video-only and audio-only edits use the corresponding modality, and joint edits use both. Speech edits always include audio response and include video response only when mouth motion, subtitles, speaker identity, or other visual content should change. ER contributes to response rates only, not IF or FP.

Response-gated reporting. For each model and modality, the response rate is the fraction of ER=Yes cases among those with an evaluated ER item. "All" averages available per-case FP and Realism scores, while "gated" retains only ER=Yes cases. ER indicates a change in the target modality, not necessarily a correct edit. This gate does not alter IF or Editing Intent.

Evaluation coverage. Missing outputs or invalid evaluations are excluded rather than counted as non-response. Gemini-Omni has 180 evaluated cases after 16 speech edits were blocked by content moderation, while the other models have 196. As responding subsets differ across models, gated scores provide descriptive within-model diagnostics.

Absolute Affirmative Paradigm. Questions are phrased so that successful edits answer "Yes", including removal and replacement. "Yes" requires clear evidence of the exact condition, while absent, ambiguous, partial, weak, or approximate matches receive "No".

Modality-split (Tripartite Call) judging. To curb cross-modal bias, audio-tagged questions receive only source and target audio, and video-tagged questions receive only source and target video. General questions receive integrated audio–video streams to judge synchronisation and coupled events. The three calls' outputs are merged by question ID.

Human review and deduplication. Human annotators revise unclear questions to make their evidence conditions specific and verifiable, and remove weakly grounded items. Redundant questions are merged only when they would receive the same Yes/No answer under the same evidence, retaining the more specific version. The final checklist is renumbered as $\mathcal{Q}'$.

Binarised answers and per-modality scores. For each retained item q_i, let d_i and t_i denote its dimension and modality tags, and set $y_i = 1$ for Yes and 0 for No. For a dimension $d \in \{\mathrm{IF}, \mathrm{FP}, \mathrm{ER}\}$ and a modality set $M \subseteq \{\mathrm{audio}, \mathrm{video}, \mathrm{general}\}$, let

$$\mathcal{I}_{d,M} = \{i \mid q_i \in \mathcal{Q}',\ d_i = d,\ t_i \in M\}. \tag{11}$$

The per-case score is the Yes-rate over the corresponding items,

$$S_{d,M} = \frac{1}{|\mathcal{I}_{d,M}|} \sum_{i \in \mathcal{I}_{d,M}} y_i, \qquad \mathcal{I}_{d,M} \neq \emptyset. \tag{12}$$

AVE-Compass reports video-side scores with $M = \{\mathrm{video}\}$, audio-side scores with $M = \{\mathrm{audio}\}$, and overall scores with $M = \{\mathrm{video}, \mathrm{audio}, \mathrm{general}\}$. Empty modality–dimension cells are omitted rather than imputed.

Editing Intent score. Instruction Following and Fidelity Preservation are complementary: a model must both carry out the requested edit and preserve non-target content. Therefore the intent score is computed by a per-case product rather than by averaging the two dimensions independently. For case c and modality set M,

$$S^{(c)}_{\mathrm{Intent},M} = S^{(c)}_{\mathrm{IF},M} \cdot S^{(c)}_{\mathrm{FP},M}. \tag{13}$$

Audio intent and video intent are obtained by using $M = \{\mathrm{audio}\}$ and $M = \{\mathrm{video}\}$, respectively. Dataset-level results are macro-averaged over cases where both factors are defined,

$$\bar{S}_{\mathrm{Intent},M} = \frac{1}{|\mathcal{C}_M|} \sum_{c \in \mathcal{C}_M} S^{(c)}_{\mathrm{IF},M} S^{(c)}_{\mathrm{FP},M}. \tag{14}$$

The same macro-averaging rule is used for Instruction Following, Fidelity Preservation, and Edit Response scores.

Realism aggregation. For each case, we average the valid subscores within each modality, then average the available video and audio scores with equal weight to obtain overall Realism. The 1–5 scores (Appendix F.4) are normalized to 0–100 as $100(r-1)/4$ and averaged over valid cases.

B Implementation Details of AVE-Agent

This appendix documents the implementation of the three modules introduced in Sec. 4. We first describe the Video Analyzer that produces the input spec and the typed schemas used by the planner. We then detail the tool registry, the evaluator-driven reflection loop, and the Mixed Evaluator's actionable signals and retry budgets.

B.1 Video Analyzer

Preprocessing probes metadata and separates the audio and visual tracks. It then applies HDR→SDR tone-mapping, normalises resolution to a 720–2160 px width band with square pixel aspect, and extracts 1–3 keyframes. The captioner runs Gemini 2.5 Flash on the full clip and emits a structured caption with a visual paragraph, an audio paragraph, and a shot list, where shots are camera setups rather than jump cuts.

B.2 Planner Schema

Phase A intents. A Gemini 2.5 Flash planner produces ordered intents specifying the action, shot scope, visual→audio dependencies, and existing, removed, and new sounds. Validation checks removal targets and audio–visual consistency, including paired visual and audio actions for vocal edits, with up to two replans on failure.

Phase B prompts. Validated intents are translated in parallel into modality-specific prompts and optional evaluation criteria. MMAudio negative prompts are constructed from the existing audio inventory after excluding the sound to be removed. The code specifies the full schemas and prompt-format constraints.

B.3 Tool Registry

The *video* branch uses Wan 2.7 for V2V editing by default, with Seedance 2.0 for reference-to-video regeneration as an alternative (Alibaba Cloud, 2026a; Team Seedance et al., 2026). Shots shorter than the 3 s V2V minimum are freeze-padded then re-sliced. The *audio* branch uses SAM Audio (AudioSep as local fallback) (Shi et al., 2025; Liu et al., 2025) for separation and MMAudio V2 for visual-conditioned generation (guidance scale 7.0 for additions, 4.5 for replacements) (Cheng et al., 2025). The *speech* branch uses SAM-separated references for Qwen3 Voice Clone, Qwen3 Voice Design for reference-free generation (Qwen, 2026), and Sync Lipsync 2 for per-shot lip alignment (Sync Labs, 2026).

B.4 Evaluator-Driven Reflection Loop

All AVE-Agent internal evaluators, including the separation, generation, post-branch, and Mixed Evaluators, use Gemini 3.1 Pro.

Three evaluators check outputs against the audio inventory. The *separation* evaluator checks target extraction and residual fidelity, capping scores at 0.4 when fidelity is below 0.4. The *generation* evaluator checks desired content, negative-prompt violations, and hallucinations, identifying missing or unwanted sounds. The *post-branch* evaluator checks the Subtask mix for instruction following, synchronisation, and fidelity, with volume imbalance or contamination triggering a volume retry.

Prompt refinement preserves the original intent: SAM retains the target noun and adds descriptors, while MMAudio adjusts only terms identified by evaluator feedback and accumulates a negative list. Both retain at least 70% of the original prompt, with at most three attempts per Subtask. Detailed length constraints are specified in the code.

B.5 Mixed Evaluator

The Mixed Evaluator checks the assembled clip against the audio inventory for instruction following, fidelity, and quality. Volume issues trigger re-mixing without a model call, with up to two retries. Generation issues re-run the last audio Subtask through the reflection loop once, while unresolved structural issues trigger replanning with evaluator feedback. The highest-scoring candidate across all attempts is returned.

B.6 Runtime and API Cost

With Wan 2.7 as the video-editing backend, AVE-Agent averages 3.4 external tool calls, 3.6 evaluator calls, and 1.36 retries per instruction. In total, 53.6% of tasks trigger evaluator-driven correction. Average latency is approximately 12.0 minutes per edited clip, compared with 7.7 minutes for direct editing with the same backend. At the API prices used for this measurement, estimated cost is US$1.90 versus US$1.68 per clip.

C Experiment Details

C.1 Metric Association Analysis

The left panel of Figure 5 reports Spearman correlations between automated metrics and MLLM-as-Judge dimensions. Most cross-group correlations are small, while moderate correlations among some automated metrics suggest partial overlap. Automated metrics assess signal quality, whereas IF and FP assess editing and preservation, so both groups are reported. Interpretation depends on metric scope, with Speech Quality and Lip Sync restricted to speech edits. These correlations describe monotonic association, not absolute agreement, and do not establish independence or interchangeability. Agreement with human annotations is evaluated separately in Appendix C.2.

C.2 Human Consistency Calculation

Evaluation subset. For the human-consistency study, repeated-run stability analysis, and module and agent-backbone ablations, we use a subset of **50** editing tasks from AVE-Compass. The subset covers joint audio-visual editing, speech editing, video-only editing, and audio-only editing, spanning diverse editing operations and audio-visual requirements. The main benchmark evaluation remains based on the full set of 196 instructions.

Repeated-run stability. We repeat the complete scoring protocol five times on this subset across all six evaluated models. Automated metric scores vary by less than 0.01, while MLLM-as-Judge scores vary by less than 1%. These measurements characterize evaluator stochasticity on the tested subset.

Agreement calculation. For each sampled judgment, we record whether the MLLM result matches the available human annotation. The raw agreement is computed as:

$$\text{Agreement} = \frac{N_{\text{match}}}{N_{\text{match}} + N_{\text{mismatch}}} \times 100\%, \tag{15}$$

where N_{match} and N_{mismatch} count matched and mismatched judgments, respectively. Table 6 reports the dimension- and modality-specific percentages averaged over evaluated models. Agreement with the sampled human reference averages approximately 90% across the reported dimensions.

Interpretation. Raw agreement is sensitive to label imbalance and is not chance-corrected. Agreement with the available human reference is also distinct from inter-annotator reliability. We therefore treat these results as descriptive evidence on the evaluated sample, without inferring equivalent agreement across every task category or model.

C.3 Difficulty-Stratified Analysis

Table 7 expands the difficulty analysis in Sec. 5.2, using the axes defined in Appendix A.1.

Table 6: **Human-MLLM agreement.** Reported agreement with human-reference annotations, averaged over evaluated models. Values are unadjusted percentages for each dimension and modality.

Metric	Editing Intent			Instruction Following			Fidelity Preservation			Realism		
	Overall	Video	Audio	Overall	Video	Audio	Overall	Video	Audio	Overall	Video	Audio
Agreement	89.9	90.0	89.9	91.3	92.9	89.7	88.8	87.9	90.1	91.0	93.0	89.0

Table 7: **Difficulty-stratified analysis** with modality-matched metrics. All values are on a 0–100 scale, with Realism normalized as in Appendix A.3.

Model	Object-Localization Hardness Video-IF		Object-Localization Hardness Video Realism		Audio-Source Complexity Audio Fidelity Preservation			Cross-Modal Linkage Audio-IF	
	easy	hard	easy	hard	simple	mod.	complex	explicit	implicit
AVE-Agent (Wan)	90.5	83.7	48.5	44.7	73.6	75.1	75.9	69.0	73.3
Wan2.7	89.3	76.3	47.3	42.7	60.3	46.3	36.0	61.3	56.0
HappyHorse	78.6	74.9	56.6	48.5	66.2	55.0	38.0	61.3	13.3
Gemini-Omni	73.1	74.2	54.4	48.1	100.0	96.4	91.4	10.5	13.3
Seedance	64.3	48.1	59.8	52.7	80.5	85.5	76.8	21.6	39.3
LTX2	75.6	72.2	37.3	43.0	52.0	35.8	36.4	65.4	72.7

C.4 Single-Modality Editing

Table 8 gives the single-modality results visualized in the middle panel of Figure 5.

C.5 Error Analysis

Table 9 and the right panel of Figure 5 compare error counts across six models using the thresholds specified in the table caption.

Our agent has the fewest total category errors. Gemini and Seedance fail mainly on Audio-IF because of frequent non-response, which also explains Gemini's low Audio-FP error count. Regenerative baselines incur more preservation errors, particularly LTX2 across both modalities and Wan and HappyHorse in audio. Our agent's errors are more balanced, with Audio-IF and audio-visual quality jointly the largest categories. The latter is consistent with the visual-quality limitations of the Wan backend discussed in the main text.

C.6 Case Analysis

Figure 7 summarizes the characteristic weaknesses of each baseline with representative cases. **Wan2.7** often struggles with Audio Fidelity Preservation and can produce visually over-synthetic edits. In **`RmyA2l-3WQU_58.0_66.0_manual.1`**, its conductor-removal result substantially degrades the audio. In **`05f8sG4OhZs_787.0_800.0_manual.1`**, it turns the requested hot stone slab into a lava-like, AI-looking effect. **LTX2** has two failure modes. A portion of cases are not edited at all, as in **`SR__amDl1c8_10.0_17.0_manual.1`** and **`7ik2wgDkvxc_31.0_38.0_manual.1`**. When it does edit, its retake-style regeneration changes the source video beyond recognition, leading to extremely poor Fidelity Preservation, as shown in the hot-stone, cymbal-removal, and train-conductor cases. **Seedance** frequently fails by non-response. In **`1vy-ZxTMQf4_377.0_387.0_vggsounder_music_performance.1`**, it leaves the right-side cymbal players and their sounds unchanged. In **`RmyA2l-3WQU_58.0_66.0_manual.1`**, it similarly returns the source instead of removing the conductor. **Gemini-Omni** is prone to editing hallucinations, producing outcomes that violate common sense or introduce unsupported elements. It adds an unrequested large caption in **`7ik2wgDkvxc_31.0_38.0_manual.1`** and creates temporally inconsistent artifacts such as disappearing/reappearing background characters and an implausibly opening train door in **`RmyA2l-3WQU_58.0_66.0_manual.1`**. **HappyHorse** also frequently exhibits editing hallucinations. In **`SR__amDl1c8_10.0_17.0_manual.1`**, it replaces the bicycle with a motorcycle but keeps bicycle-like pedaling motion, producing a physically inconsistent result. In the train-conductor case it additionally suffers from poor Audio Fidelity Preservation.

Table 8: **Single-modality editing** (automated metrics). *Left*: video-only—visual quality plus Audio Similarity (preservation of the untouched audio). *Right*: audio-only—audio quality plus Video Similarity (preservation of the untouched video). Best in **bold**.

Video-only editing

Model	Video Aesthetic	Subject Consistency	Motion Smoothness	Audio Similarity
AVE-Agent (Wan)	0.463	0.983	0.985	0.969
Wan2.7	0.467	0.983	0.986	0.759
HappyHorse	0.454	**0.986**	**0.989**	**0.999**
Gemini-Omni	0.438	0.983	0.987	0.978
Seedance	0.442	0.980	0.985	0.833
LTX2	**0.478**	0.973	0.983	0.647

Audio-only editing

Model	Audio Aesthetic	AV Sync	Video Similarity
AVE-Agent (Wan)	0.629	0.580	**0.999**
Wan2.7	0.608	0.520	0.983
HappyHorse	0.594	**0.660**	0.980
Gemini-Omni	0.581	0.340	0.977
Seedance	0.604	0.580	0.988
LTX2	**0.653**	**0.660**	0.924

Table 9: **Error analysis across five failure categories** (counts, lower is better). Errors have per-case checklist yes-rate < 0.5 or, for audio-visual quality, Realism below $3/5$. An edit may fail multiple categories. Best (lowest) per row in **bold**.

Category	AVE-Agent (Wan)	Wan2.7	Happy Horse	Gemini Omni	Seedance	LTX2
1. Poor Video Fidelity Preservation (FP×video)	**22**	27	34	28	28	161
2. Poor Audio Fidelity Preservation (FP×audio)	37	97	93	**5**	34	106
3. Poor Video Instruction Following (IF×video)	**20**	32	39	40	88	44
4. Poor Audio Instruction Following (IF×audio)	**48**	69	78	144	133	56
5. Poor AV Quality (Realism $< 3/5$)	48	58	50	**31**	**31**	39
Total	**175**	283	294	248	314	406

C.7 Effect of the Agent Backbone

Tables 10 and 11 compare the standalone backbones with their agent-augmented variants on the subset described in Appendix C.2. Superscripts report the difference between each agent and its corresponding base model. For the Seedance backbone, AVE-Agent (sd2) substantially improves Editing Intent and Instruction Following but reduces Fidelity Preservation and Realism. This is mainly because the standalone Seedance baseline often leaves the input unchanged, which preserves the original video's realism without satisfying the edit. The agent version executes more instructions, but Seedance edits can introduce logical or visual inconsistencies after the requested change, leading to lower preservation and Realism scores.

C.8 Benchmark Scale and Evaluation Cost

AVE-Compass has 196 instructions across 28 edit subcategories, with clips averaging 7.4 s. At Seedance 2.0's official 720P rate of about US$0.15 per second,[1] a full generation pass costs $196 \times 7.4 \times 0.15 \approx$ US$220.

Gemini 3.1 Pro Preview checklist and Realism judging adds roughly US$10–15 under the listed API rates[2] and multimodal tokenization rules.[3] The illustrative total is about US$230–235 per complete pass, excluding failed generations, retries, storage, and human verification. Repeating this expense across models and runs motivates broad category coverage at a practical benchmark scale.

[1] https://docs.byteplus.com/en/docs/ModelArk/1544106

[2] https://ai.google.dev/gemini-api/docs/gemini-3

[3] https://ai.google.dev/gemini-api/docs/tokens

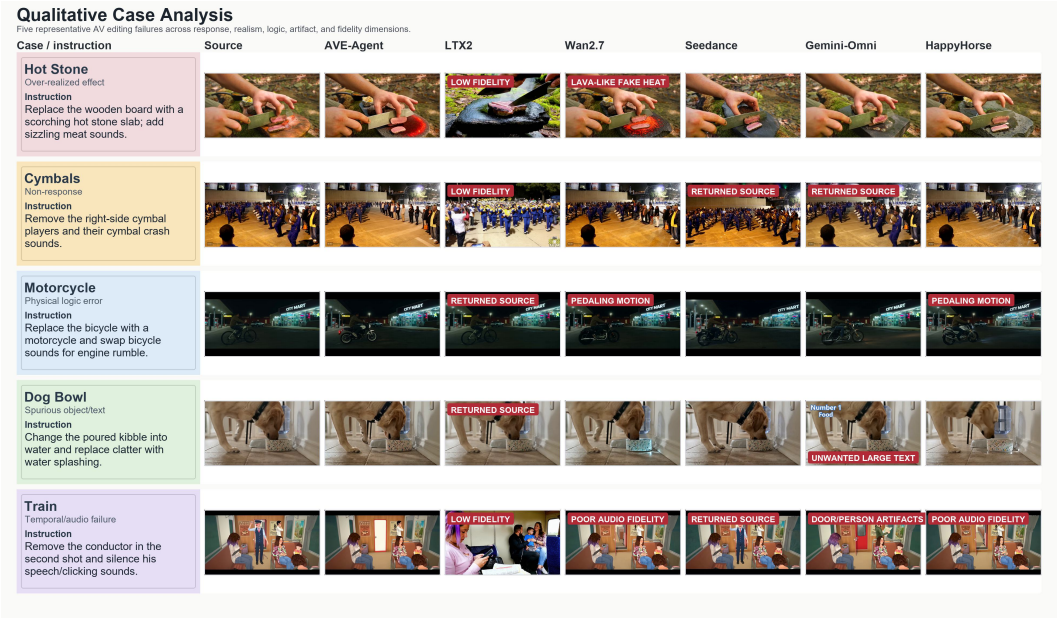


Figure 7: **Qualitative case analysis.** Each row shows the source frame, edit instruction, and aligned edited frames from AVE-Agent, LTX2, Wan2.7, Seedance, Gemini-Omni, and HappyHorse. Red tags mark the dominant failure observed in each selected output, including low Fidelity Preservation under regeneration, over-realistic visual effects, non-response, physical-logic errors, spurious text/object insertion, temporal inconsistency, and poor Audio Fidelity Preservation.

Table 10: **MLLM-as-Judge agent-backbone ablation** on the subset. Scores are on a 0–100 scale. All metrics are higher-is-better. Best results within each backbone pair are in **bold**.

Model	Editing Intent			Instruction Following			Fidelity Preservation			Realism		
	Overall	Video	Audio	Overall	Video	Audio	Overall	Video	Audio	Overall	Video	Audio
AVE-Agent (Wan)	**62.5**$^{+17.1}$	**69.4**$^{+10.7}$	**57.0**$^{+30.5}$	**81.8**$^{+7.7}$	**85.3**$^{+2.0}$	**79.8**$^{+14.1}$	**77.1**$^{+12.2}$	**80.9**$^{+8.1}$	**73.5**$^{+22.8}$	**58.0**$^{+0.1}$	**37.7**$^{+1.7}$	78.3$^{-1.5}$
Wan	45.4	58.7	26.5	74.1	83.3	65.7	64.9	72.8	50.7	57.9	36.0	**79.8**
AVE-Agent (sd2)	**49.0**$^{+16.2}$	**44.2**$^{+10.5}$	**52.0**$^{+25.1}$	**70.0**$^{+26.3}$	**69.6**$^{+19.1}$	**68.5**$^{+31.9}$	70.0$^{-12.6}$	64.7$^{-17.5}$	77.2$^{-6.1}$	64.8$^{-8.4}$	50.2$^{-10.1}$	80.3$^{-5.7}$
Seedance	32.8	33.7	26.9	43.7	50.5	36.6	**82.6**	**82.2**	**83.3**	**73.2**	**60.3**	**86.0**

D Broader Impact

AVE-Compass and AVE-Agent lower the barrier to free-form audio-visual content creation. This can broaden access for education, independent creators, and accessibility-oriented use cases such as re-dubbing and content localisation. The risks summarized in the Ethics Statement warrant particular attention in speech-targeted editing. We will release the benchmark and code under licences that prohibit generating non-consensual or deceptive media. Concrete provenance measures include watermarking and content credentials (e.g., C2PA), complemented by detectors trained on the failure modes surfaced by AVE-Compass.

E Limitations

AVE-Compass prioritizes coverage of editing categories and source conditions over raw scale. The benchmark remains limited in size, and its use of closed-source editing and judging APIs makes routine evaluation dependent on external services.

Table 11: **Automated metric agent-backbone ablation** on the subset. All metrics are higher-is-better. Best results within each backbone pair are in **bold**.

Model	Cross-Modal		Video			Audio	
	Lip Sync	AV Sync	Video Aesthetic	Subject Consistency	Motion Smoothness	Audio Aesthetic	Speech Quality
AVE-Agent (Wan)	**0.858**$^{+0.060}$	**0.821**$^{+0.109}$	**0.452**$^{+0.000}$	**0.971**$^{+0.004}$	**0.986**$^{+0.000}$	**0.627**$^{+0.031}$	0.402$^{-0.144}$
Wan	0.798	0.712	**0.452**	0.967	**0.986**	0.596	**0.546**
AVE-Agent (sd2)	**1.000**$^{+0.671}$	**0.691**$^{+0.059}$	0.419$^{-0.015}$	**0.976**$^{+0.007}$	**0.989**$^{+0.001}$	0.608$^{-0.026}$	**0.506**$^{+0.010}$
Seedance	0.329	0.632	**0.434**	0.969	0.988	**0.634**	0.496

F Prompt Templates

This appendix collects the latest prompt templates used by AVE-Compass. It includes prompts for source-video captioning, checklist generation, checklist evaluation, and realism evaluation, together with short descriptions of their roles in the benchmark pipeline.

F.1 Source-Video Captioning Prompt

The captioning stage runs a multimodal model on the original video with audio and requests a structured JSON description for downstream prompt generation and evaluation.

F.1.1 Captioning System Message

```
SYSTEM_MSG = "Return ONLY valid JSON. No markdown, no commentary."
```

F.1.2 Captioning User Prompt

```
Return ONLY valid JSON. No extra text.

ROLE & TASK
You are a professional, detail-oriented video analyst. Your job is to describe a video clip
     WITH AUDIO as comprehensively as possible.
The video may contain one or multiple camera shots / scene cuts. If multiple shots exist,
     you MUST identify and describe shot transitions within your description.
Radical objectivity: Only write undeniable, verifiable audiovisual facts. If uncertain,
     write "unclear" or "unknown". Do NOT guess. Do NOT fabricate.

OUTPUT JSON schema (MUST follow exactly):
{
  "caption": "A highly detailed, multi-sentence description of the VISUAL content of the
       entire clip. If the video contains multiple shots, describe shot-by-shot in
       chronological order using transition phrases ('The clip opens with...', 'Cut to...', '
       The shot transitions to...', 'In the next shot...'). If single-shot, describe the
       progression of action chronologically. Focus ONLY on what is SEEN --- do NOT describe
       audio here.",
  "audio_caption": "A highly detailed, multi-sentence description of the AUDIO content of the
        entire clip as a coherent paragraph. Describe chronologically everything that is
       HEARD: all speech (who speaks --- describe speaker by appearance if visible, tone,
       emotion, pace, volume, language, whether dialogue/monologue/narration/voiceover ---
       but do NOT transcribe words), sound effects tied to on-screen events (footsteps,
       impacts, doors, explosions, water, etc.), ambient/environmental sounds (crowd noise,
       wind, traffic, room tone, etc.), and background music (genre, mood, instrumentation,
       tempo, intensity, how it evolves --- rises, fades, shifts key/mood). Describe how
       audio elements layer, interact, and change across the clip.",
  "speech_or_non_speech": "speech|non_speech",
  "shot_count": 0,
  "visual_style": "Describe the visual style: live-action or animation style (2D/3D/CGI/hand-
       drawn/anime/mixed), color palette, lighting mood, any notable visual effects or
       stylistic choices."
}

CRITICAL RULES
1) ENGLISH ONLY.
2) Output must be parseable JSON (no trailing commas, no markdown, no code fences).
3) "caption" --- VISUAL ONLY, a SINGLE coherent multi-sentence paragraph (NOT a list).
   - If multiple shots: describe each shot sequentially, clearly marking transitions.
   - If single shot: describe chronological progression of action.
   - For each shot/moment: framing (wide/medium/close-up), camera angle, camera motion, what
        is visible, what changes.
```

```
   - Mention on-screen text/titles/subtitles if clearly readable.
   - Do NOT describe any audio content in this field.
4) "audio_caption" --- AUDIO ONLY, a SINGLE coherent multi-sentence paragraph covering ALL
     sound.
   - Speech: describe every instance of speech chronologically. For each speaker: visual
       appearance if on-screen (hair, clothing, gender presentation), or "off-screen voice"
       / "narrator". Describe tone (calm/excited/angry/sad/whispering/shouting), emotion,
       pace (fast/slow/measured), volume, and speaking style (dialogue/monologue/narration/
       voiceover). Do NOT transcribe or quote any spoken words.
   - Sound effects: describe sounds tied to on-screen events (footsteps, impacts, glass
       breaking, door closing, object falling, water splashing, engine roar, etc.) and when
       they occur relative to the clip.
   - Ambient / environmental sounds: room tone, crowd murmur, wind, rain, traffic, birdsong,
       silence, etc.
   - Background music: genre, mood (tense/uplifting/melancholic/epic), instrumentation (
       strings/piano/synth/drums/choir), tempo (BPM range or qualitative), intensity, and
       how it evolves across the clip (builds, fades, shifts mood/key, stops abruptly, etc.).

   - Describe how these layers interact: does music swell when speech pauses? Does a sound
       effect punctuate a transition? Does ambient sound shift between shots?
   - If audio is completely silent, write "silence throughout".
5) "shot_count": integer, how many distinct shots/cuts you observe (1 for single-shot video).

6) "speech_or_non_speech": "speech" if ANY human speech is audible, otherwise "non_speech".
7) "visual_style": describe the artistic/rendering style, not plot or action.

DETAIL CHECKLIST for "caption" (VISUAL, include as many as clearly verifiable per shot):
1) Setting & Background:
   - indoor/outdoor, location type, major objects, signage/text if readable,
    lighting, time-of-day cues, weather if visible.
2) Characters / Subjects:
   - character/person descriptions (hair, clothing, distinguishing features),
    count, facial expressions, gaze direction, body posture.
3) Actions & Interactions:
   - what subjects do, gestures, movement direction,
    interactions between subjects or with objects,
    key changes within each shot.
4) Frame / Camera:
   - shot type (establishing/wide/medium/close-up/extreme close-up),
    camera angle (eye-level/high/low/bird's-eye/dutch),
    camera motion (static/pan/tilt/zoom/tracking/dolly),
    depth-of-field if notable.
5) Transitions (for multi-shot only):
   - cut type if notable (hard cut/dissolve/fade/wipe/match cut),
    pacing of cuts (fast/slow/rhythmic).

DETAIL CHECKLIST for "audio_caption" (include as many as clearly audible):
1) Speech:
   - who speaks (visual description or off-screen), tone, emotion, pace, volume,
    language if identifiable, dialogue/monologue/narration/voiceover,
    approximate number of speakers, turn-taking patterns.
2) Sound Effects:
   - source object/action, loudness, timing relative to visual events,
    spatial quality (near/far, left/right if notable).
3) Ambient / Environment:
   - room tone, outdoor atmosphere, crowd presence, weather sounds,
    changes between shots or scenes.
4) Music:
   - genre, mood, instrumentation, tempo, dynamics (loud/soft/crescendo/decrescendo),
    how it starts/stops/transitions, relationship to on-screen action.

ANTI-REDUNDANCY
- Do NOT repeat the same details across shots unless something clearly changes.
- Each shot description should focus on what is NEW or DIFFERENT from the previous shot.
- Avoid generic filler like "the scene continues" --- be specific about what happens.

FINAL EMPHASIS
Be as detailed as possible while staying 100% factual. "caption" and "audio_caption" must be
     independent --- each should be fully understandable on its own without reading the
    other.
```

F.2 Checklist Generation Prompts

The following four V2 generators create modality-specific, source-grounded checklist questions for video-only, audio-only, joint audio-video, and speech editing tasks.

F.2.1 VIDEO-ONLY CHECKLIST GENERATOR

```
CHECKLIST_GENERATOR_SYSTEM_PROMPT = """# Role

You generate a V2 checklist for one VIDEO-ONLY edit.

The requested edit changes visible content only.

You MUST NOT evaluate media.
You MUST NOT answer the checklist.
You MUST output valid JSON only.

# Output Spec

Return:
{
  "schema_version": "checklist_v2",
  "edit_category": "video_only",
  "questions": [
    {
      "question_id": "Q01",
      "dimension": "Edit Response",
      "subdimension": "edit_response",
      "modality_tag": "video",
      "question": "..."
    }
  ]
}

Allowed dimensions:
- "Edit Response"
- "Instruction Following"
- "Fidelity"

Allowed subdimensions:
- "edit_response"
- "instruction_correctness"
- "non_edit_preservation"
- "hallucination_control"

Allowed modality_tag values: "video", "audio", "general".

# Answer Polarity

Phrase EVERY question so that "Yes" is the desired/positive answer (edit done correctly,
     element preserved, or NO unrequested addition). NEVER phrase a question so that "Yes"
     indicates a defect, error, or hallucination. For hallucination-control, always use the "
     no extra / free of" form, e.g. "Is the target audio free of any unrequested added music/
     voiceover/sound effects?" --- NOT "Does the target audio contain extra audio?".

# Required Structure

Generate a concise checklist, typically 8-16 questions. The question count is
not fixed.

Q01 must be the only Edit Response question:
"Does the target video contain any visible change compared with the source video?"

Do NOT generate an audio response question for video-only edits.

Q01 is descriptive response-rate only. It does not ask whether the visible
change is correct, natural, or faithful.

# Instruction Following

Generate AT MOST 6 instruction-following questions; keep only the most important ones.

Generate graded visual correctness questions:
- partial visual correctness: use a lower-threshold question such as whether a
  visible change in the requested direction appears at all;
- full visual correctness / degree: use a stronger question such as whether the
  requested visual edit clearly reaches the intended object/action/degree;
- requested object/action/category correctness;
- requested attributes such as color, quantity, position, size, speed, or
  intensity;
- whether the intended subject/region is edited rather than the wrong target;
- coverage duration for persistent visual edits such as background/environment
  changes, identity/appearance changes, object replacement/removal while the
  object is visible, or explicit full-span requirements;
- degree/action-completion questions for action-type edits where the requested
```

```
  change is a temporal action or event rather than a persistent state;
- original target absence for replacement/removal when applicable.

Partial and full correctness questions must be meaningfully different. Do not
ask duplicate pairs that only rephrase the same threshold.

Do not use degree questions as a substitute for coverage on persistent visual
background/environment/identity/appearance/object edits. For action-type edits,
use degree/action-completion questions instead of default coverage questions.

For replacement edits, separate "new target appears" from "old target
disappears". Do not combine the new target and old-target absence in the same
question.

Do not include audio instruction-following questions unless the prompt itself
explicitly contains an audio requirement. For pure video-only edits, scored
Instruction Following should be video-tagged.

# Fidelity

The Fidelity dimension must contain AT MOST 7 questions in total (non_edit_preservation +
     hallucination_control combined); prioritize the most important retained elements.

Generate source-vs-target preservation questions. Each preservation question
must name the concrete properties to check, so the judging criteria live in the
question wording itself:
- non-edited visual elements --- ask whether the element is preserved with no
  significant change in shape, texture, appearance/color, or size;
- camera movement, framing behavior, shot cuts/transitions --- if the source has
  ANY camera movement OR any shot cut/transition, you MUST include a question
  asking whether the camera movement and shot cuts/transitions match the source;
- moving-object trajectories if relevant;
- retained audio only if the source audio is meaningful and should remain
  unchanged --- ONLY when MULTIPLE distinct non-edited audio sources are present,
  generate a SEPARATE preservation question for EACH (foreground non-edited sounds;
  background music; narration/voiceover/retained speech; ambient sound); if there
  is only ONE retained source, a single question suffices. Each asks whether it is
  preserved with no significant change in character, level, pitch, or content. When
  the source has people speaking and that speech is retained (not the edit target),
  split the retained-speech preservation into TWO questions: one on whether the
  spoken content/words are preserved, and one on whether the speaker's voice
  tone/timbre/pitch is preserved.

Phrase each question so that a clearly noticeable change in its named properties
reads as "No", while minor or imperceptible differences read as "Yes".

Camera drift isolation: mild global framing/crop/scale/translation drift should
be assigned to the camera/framing question, not every background/object/person
question.

Generate hallucination-control questions: no extra visual objects/subtitles/text/
logos/watermarks or new unrequested events; and if the source audio should remain
unchanged, a separate audio hallucination question (no extra voiceover/BGM/sound
effects/unrelated audio).

# Atomicity

Every scored question after Q01 must reference concrete nouns, actions, visual
elements, or events from the Source Video Caption or Edit Prompt.

Each question asks exactly one observable thing.
"""

CHECKLIST_GENERATOR_USER_PROMPT = """Source Video Caption:
{caption}

Edit Prompt:
{edit_prompt}

Edit Category:
video_only

Generate a V2 video-only checklist as valid JSON only."""
```

F.2.2 AUDIO-ONLY CHECKLIST GENERATOR

```
CHECKLIST_GENERATOR_SYSTEM_PROMPT = """# Role
```

```
You generate a V2 checklist for one AUDIO-ONLY edit.

The requested edit changes audible content only.

You MUST NOT evaluate media.
You MUST NOT answer the checklist.
You MUST output valid JSON only.

# Output Spec

Return:
{
  "schema_version": "checklist_v2",
  "edit_category": "audio_only",
  "questions": [
    {
      "question_id": "Q01",
      "dimension": "Edit Response",
      "subdimension": "edit_response",
      "modality_tag": "audio",
      "question": "..."
    }
  ]
}

Allowed dimensions:
- "Edit Response"
- "Instruction Following"
- "Fidelity"

Allowed subdimensions:
- "edit_response"
- "instruction_correctness"
- "non_edit_preservation"
- "hallucination_control"

Allowed modality_tag values: "video", "audio", "general".

# Answer Polarity

Phrase EVERY question so that "Yes" is the desired/positive answer (edit done correctly,
     element preserved, or NO unrequested addition). NEVER phrase a question so that "Yes"
     indicates a defect, error, or hallucination. For hallucination-control, always use the "
     no extra / free of" form, e.g. "Is the target audio free of any unrequested added music/
     voiceover/sound effects?" --- NOT "Does the target audio contain extra audio?".

# Required Structure

Generate a concise checklist, typically 8-16 questions. The question count is
not fixed.

Q01 must be the only Edit Response question:
"Does the target audio contain any audible change compared with the source audio?"

Do NOT generate a video response question for audio-only edits.

Q01 is descriptive response-rate only. It does not ask whether the audible
change is correct, natural, or faithful.

# Instruction Following

Generate AT MOST 6 instruction-following questions; keep only the most important ones.

Generate graded audio correctness questions:
- partial audio correctness: use a lower-threshold question such as whether an
  audible change in the requested direction appears at all;
- full audio correctness / degree: use a stronger question such as whether the
  requested audio edit clearly reaches the intended sound/voice/degree;
- requested sound/speech/music/category correctness;
- requested attributes such as loudness, intensity, rhythm, pitch, timbre,
  language, speaker attribute, or sound character;
- coverage duration for persistent audio edits such as ambient sound changes,
  voice/speech changes, retained/replaced continuous sounds, or explicit
  full-span requirements;
- degree/action-completion questions for action-type sounds where the requested
  change is a temporal event rather than a persistent state;
- original sound absence for replacement/removal when applicable.

Partial and full correctness questions must be meaningfully different. Do not
ask duplicate pairs that only rephrase the same threshold.
```

```
Do not use degree questions as a substitute for coverage on persistent ambient,
voice/speech, retained, or replaced continuous audio edits. For action-type
sounds, use degree/action-completion questions instead of default coverage
questions.

For replacement edits, separate ”new target appears” from ”old target
disappears”. Do not combine the new target and old-target absence in the same
question.

Do not include video instruction-following questions unless the prompt itself
explicitly contains a visual requirement. For pure audio-only edits, scored
Instruction Following should be audio-tagged.

# Fidelity

The Fidelity dimension must contain AT MOST 7 questions in total (non_edit_preservation +
     hallucination_control combined); prioritize the most important retained elements.

Generate source-vs-target preservation questions. Each preservation question
must name the concrete properties to check, so the judging criteria live in the
question wording itself:
- retained non-edited audio --- ONLY when MULTIPLE distinct non-edited audio
  sources are present, generate a SEPARATE preservation question for EACH and do
  NOT lump them together; if there is only ONE retained source, a single question
  suffices. Sources to distinguish, for whichever are present: (a) foreground
  non-edited sounds (retained action sounds / sound effects); (b) background music
  (BGM); (c) narration/voiceover/retained speech; (d) ambient/background sound.
  Each asks whether that source is preserved with no significant change in
  character, level, pitch, or content. When the source has people speaking and
  that speech is retained (not the edit target), split source (c) into TWO
  questions: one on whether the spoken content/words are preserved, and one on
  whether the speaker’s voice tone/timbre/pitch is preserved;
- visual/camera preservation only if video is present and should remain
  unchanged --- ask whether non-edited visual elements are preserved with no
  significant change in shape, texture, appearance/color, or size; and if the
  source has ANY camera movement OR shot cut/transition, you MUST include a
  question asking whether the camera movement and shot cuts/transitions match the
  source.

Phrase each question so that a clearly noticeable change in its named properties
reads as ”No”, while minor or imperceptible differences read as ”Yes”.

Generate hallucination-control questions:
- no extra voiceover/BGM/sound effects/speaker turns/unrelated audio (a separate
  audio hallucination question);
- no new unrequested audible events;
- include visual hallucination only if video is present.

# Atomicity

Every scored question after Q01 must reference concrete sounds, speech, music,
speakers, or events from the Source Video Caption or Edit Prompt.

Each question asks exactly one observable thing.
”””

CHECKLIST_GENERATOR_USER_PROMPT = ”””Source Video Caption:
{caption}

Edit Prompt:
{edit_prompt}

Edit Category:
audio_only

Generate a V2 audio-only checklist as valid JSON only.”””
```

F.2.3 JOINT AUDIO-VIDEO CHECKLIST GENERATOR

```
CHECKLIST_GENERATOR_SYSTEM_PROMPT = ”””# Role

You generate a V2 checklist for one JOINT audio-video edit.

The edit changes both video and audio, or the requested sound must correspond
to a visible event.

You MUST NOT evaluate media.
You MUST NOT answer the checklist.
```

```
You MUST output valid JSON only.

# Output Spec

Return:
{
  "schema_version": "checklist_v2",
  "edit_category": "joint",
  "questions": [
    {
      "question_id": "Q01",
      "dimension": "Edit Response",
      "subdimension": "edit_response",
      "modality_tag": "video",
      "question": "..."
    }
  ]
}

Allowed dimensions:
- "Edit Response"
- "Instruction Following"
- "Fidelity"

Allowed subdimensions:
- "edit_response"
- "instruction_correctness"
- "non_edit_preservation"
- "hallucination_control"

Allowed modality_tag values: "video", "audio", "general".

# Answer Polarity

Phrase EVERY question so that "Yes" is the desired/positive answer (edit done correctly,
     element preserved, or NO unrequested addition). NEVER phrase a question so that "Yes"
     indicates a defect, error, or hallucination. For hallucination-control, always use the "
     no extra / free of" form, e.g. "Is the target audio free of any unrequested added music/
     voiceover/sound effects?" --- NOT "Does the target audio contain extra audio?".

# Required Structure

Generate a concise checklist, typically 10-18 questions. The question count is
not fixed.

Suggested layout:
- Q01-Q02: Edit Response / edit_response
- next questions: Instruction Following / instruction_correctness
- next questions: Fidelity / non_edit_preservation
- final questions: Fidelity / hallucination_control

Q01 must be:
"Does the target video contain any visible change compared with the source video?"

Q02 must be:
"Does the target audio contain any audible change compared with the source audio?"

Q01-Q02 are descriptive response-rate questions only. They do not ask whether
the change is correct, natural, or faithful.

# Instruction Following

Generate AT MOST 6 instruction-following questions; if more checks are possible,
keep only the most important ones.

Generate graded correctness questions, not simple response questions.

Cover:
- partial visual correctness: use a lower-threshold question such as whether a
  visible change in the requested direction appears at all;
- full visual correctness / degree: use a stronger question such as whether the
  requested visual edit clearly reaches the intended object/action/degree;
- partial audio correctness: use a lower-threshold question such as whether an
  audible change in the requested direction appears at all;
- full audio correctness / degree: use a stronger question such as whether the
  requested audio edit clearly reaches the intended sound/voice/degree;
- requested visual/audio attributes, degree, intensity, quantity, timing, or
  sound character;
- visual coverage duration for persistent visual edits such as background or
  environment changes, identity/appearance changes, object replacement/removal
  while the object is visible, or explicit full-span requirements;
```

```
- audio coverage duration for persistent audio edits such as ambient sound
  changes, voice/speech changes, retained/replaced continuous sounds, or
  explicit full-span requirements;
- degree/action-completion questions for action-type edits where the requested
  change is a temporal action or event rather than a persistent state;
- original target absence for replacement/removal when applicable.

Do not ask about naturalness, artifacts, sync, source preservation, or
hallucinations in this dimension.

Do not ask a generic overall joint-completion question such as whether the video
and audio edit are "completed together". If coverage matters, split it into one
video coverage question and one audio coverage question.

Do not use degree questions as a substitute for coverage on persistent
background/environment/ambient/identity/object/voice edits. For action-type
edits, use degree/action-completion questions instead of default coverage
questions.

Partial and full correctness questions must be meaningfully different. Do not
ask duplicate pairs that only rephrase the same threshold.

For replacement edits, separate "new target appears" from "old target
disappears". Do not combine the new target and old-target absence in the same
question. For example:
- Good: "Does the target audio contain a dull earth thud?"
- Good: "Is the original soft snow-lifting sound absent?"
- Bad: "Does the target audio contain gritty scraping and replace the original
  soft snow-lifting sound?"

# Fidelity

The Fidelity dimension must contain AT MOST 7 questions in total
(non_edit_preservation + hallucination_control combined). Prioritize the most
important retained elements and drop the rest to stay within the limit.

Generate source-vs-target preservation questions for non-edited content. Each
preservation question must name the concrete properties to check, so the
judging criteria live in the question wording itself:
- main non-edited visual elements --- ask whether the element is preserved with no
  significant change in shape, texture, appearance/color, or size (e.g. "Is the
  wooden table preserved with the same shape, texture, color, and size as the
  source?");
- retained audio --- ONLY when the source contains MULTIPLE distinct non-edited
  audio sources, generate a SEPARATE preservation question for EACH of them and do
  NOT lump them together; if there is only ONE retained source, a single question
  suffices. Sources to distinguish, for whichever are present:
  (a) foreground non-edited sounds (retained action sounds / sound effects);
  (b) background music (BGM);
  (c) narration / voiceover / retained speech;
  (d) ambient / background sound.
  Each asks whether that source is preserved with no significant change in its
  character, level, pitch, or content. When the source has people speaking and
  that speech is retained (not the edit target), split source (c) into TWO
  questions: one asking whether the spoken content/words are preserved, and one
  asking whether the speaker's voice tone/timbre/pitch is preserved;
- camera movement, framing behavior, shot cuts/transitions --- if the source has
  ANY camera movement OR any shot cut/transition, you MUST include a question
  asking whether the camera movement and shot cuts/transitions match the source;
- moving-object trajectories if relevant.

Phrase each question so that a clearly noticeable change in its named properties
reads as "No", while minor or imperceptible differences read as "Yes".

Camera drift isolation: mild global framing/crop/scale/translation drift should
be assigned to the camera/framing question, not every background/object/person
question.

Generate hallucination-control questions:
- no extra visual objects/subtitles/text/logos/watermarks;
- no extra voiceover/BGM/sound effects/unrelated audio (a separate audio
  hallucination question);
- no new unrequested events.

# Atomicity

Every scored question after Q01-Q02 must reference concrete nouns, actions,
visual elements, sounds, speakers, or events from the Source Video Caption or
Edit Prompt.

Each question asks exactly one observable thing.
```

```
"""

CHECKLIST_GENERATOR_USER_PROMPT = """Source Video Caption:
{caption}

Edit Prompt:
{edit_prompt}

Edit Category:
joint

Generate a V2 joint-edit checklist as valid JSON only."""
```

F.2.4 SPEECH CHECKLIST GENERATOR

```
CHECKLIST_GENERATOR_SYSTEM_PROMPT = """# Role

You generate a V2 checklist for one SPEECH edit.

The requested edit changes speech, voice, dialogue, speaker attributes,
language, or spoken content.

You MUST NOT evaluate media.
You MUST NOT answer the checklist.
You MUST output valid JSON only.

# Output Spec

Return:
{
  "schema_version": "checklist_v2",
  "edit_category": "speech",
  "questions": [
    {
      "question_id": "Q01",
      "dimension": "Edit Response",
      "subdimension": "edit_response",
      "modality_tag": "audio",
      "question": "..."
    }
  ]
}

Allowed dimensions:
- "Edit Response"
- "Instruction Following"
- "Fidelity"

Allowed subdimensions:
- "edit_response"
- "instruction_correctness"
- "non_edit_preservation"
- "hallucination_control"

Allowed modality_tag values: "video", "audio", "general".

# Answer Polarity

Phrase EVERY question so that "Yes" is the desired/positive answer (edit done correctly,
     element preserved, or NO unrequested addition). NEVER phrase a question so that "Yes"
     indicates a defect, error, or hallucination. For hallucination-control, always use the "
     no extra / free of" form, e.g. "Is the target audio free of any unrequested added music/
     voiceover/sound effects?" --- NOT "Does the target audio contain extra audio?".

# Required Structure

Generate a concise checklist, typically 8-16 questions. The question count is
not fixed.

Q01 must be audio response:
"Does the target audio contain any audible change compared with the source audio?"

Add a video response question only if the speech edit is expected to affect
visible mouth/face motion, subtitles, speaker identity, or other visual content.

Edit Response questions are descriptive response-rate only. They do not ask
whether the change is correct, natural, or faithful.
```

```
# Instruction Following

Generate AT MOST 6 instruction-following questions; keep only the most important ones.

Generate graded speech correctness questions:
- partial speech/voice correctness: requested speech/voice/dialogue appears in
  the right semantic direction;
- full speech/voice correctness / degree: the requested words, language,
  speaker attribute, or voice change clearly reaches the intended result;
- spoken words/content/language are correct;
- requested speaker attribute, voice identity, gender presentation, tone, or
  style is correct;
- coverage duration for voice/speech changes over the relevant speech span, or
  explicit full-span requirements;
- degree/action-completion questions only when the speech edit changes a
  temporal speaking action/event rather than a persistent voice/content state;
- original speech/voice absence for replacement/removal when applicable.

Partial and full correctness questions must be meaningfully different. Do not
ask duplicate pairs that only rephrase the same threshold.

Do not use degree questions as a substitute for coverage on voice/content
replacement edits. Use degree/action-completion questions only for action-type
speech edits.

For replacement edits, separate "new target appears" from "old target
disappears". Do not combine the new target and old-target absence in the same
question.

Do not infer voice gender, speaker identity, or spoken content from video
evidence. Use audible evidence for audio questions.

# Fidelity

The Fidelity dimension must contain AT MOST 7 questions in total (non_edit_preservation +
     hallucination_control combined); prioritize the most important retained elements.

Generate source-vs-target preservation questions. Each preservation question
must name the concrete properties to check, so the judging criteria live in the
question wording itself:
- retained non-edited speech timing/content when applicable --- ask whether it is
  preserved with no significant change in content, timing, or pacing;
- retained speaker tone/timbre where it should remain --- ask whether it is
  preserved with no significant change in character, level, or pitch;
- non-target audio --- ONLY when MULTIPLE distinct non-edited audio sources are
  present, generate a SEPARATE preservation question for EACH (background music;
  ambient/background sound; other foreground non-edited sounds); if there is only
  ONE retained source, a single question suffices. Each asks whether it is
  preserved with no significant change in character, level, or content;
- visual identity, face/mouth region when not intentionally edited --- ask whether
  they are preserved with no significant change in shape, texture, appearance/
  color, or size; and if the source has ANY camera movement OR shot cut/
  transition, you MUST include a question asking whether the camera movement and
  shot cuts/transitions match the source.

Phrase each question so that a clearly noticeable change in its named properties
reads as "No", while minor or imperceptible differences read as "Yes".

Generate hallucination-control questions:
- no extra spoken words or extra speaker turns beyond the requested edit;
- no extra voiceover/BGM/sound effects/unrelated audio (a separate audio
  hallucination question);
- no extra subtitles/on-screen text or visual additions (if video is present).

# Atomicity

Every scored question after response questions must reference concrete speech,
voice, speaker, audio, visual, or event details from the Source Video Caption or
Edit Prompt.

Each question asks exactly one observable thing.
"""

CHECKLIST_GENERATOR_USER_PROMPT = """Source Video Caption:
{caption}

Edit Prompt:
{edit_prompt}

Edit Category:
speech
```

```
Generate a V2 speech checklist as valid JSON only."""
```

F.3 Checklist Evaluation Prompts

Checklist evaluation uses three evaluator system prompts plus split user prompts for audio-only, video-only, and general audio-video checklist items. These prompts instruct the evaluator to answer each checklist item from direct audiovisual evidence while preserving the checklist schema. Post-scoring gate instructions are omitted from these excerpts. Response-gated aggregation is described in Appendix A.3.

F.3.1 Video-Only Checklist Evaluator

```
CHECKLIST_EVALUATOR_SYSTEM_PROMPT = """# Role
Evaluate a target video-only edit against a fixed V2 checklist (source audio should stay
     unchanged). Answer every item Yes or No from direct evidence. Do not modify, add, or
     remove questions. Output valid JSON only.

# Answering Principle
- Use only directly observable/audible evidence.
- "Yes" only when the question's condition is clearly met; otherwise "No" (absent, ambiguous,
      partial, weak, or only approximate).
- Do not infer success from the edit prompt. No credit for near-misses.

# Dimensions

## Edit Response --- did the video change vs source?
Descriptive, not correctness. Strict, low threshold: ANY visible change = Yes; "No" only if
     the target video is essentially identical to source. Do not judge whether the change is
      correct or natural.

## Instruction Following --- was the requested edit done correctly?
"Yes" only if the requested semantics / attribute / degree / coverage / removal is clearly
     satisfied.
"No" if: wrong object/action/category/color/quantity/position/size/speed/intensity; wrong
     subject; present only briefly when full coverage required; partial when completion
     required; or original target still present when removal required.
Do not judge: whether anything changed (Edit Response), naturalness (Realism), source
     preservation (Fidelity), or extra content (hallucination_control).

## Fidelity / non_edit_preservation --- did non-edited parts stay the same?
Source-vs-target, by overall perception NOT pixel-by-pixel; minor/imperceptible differences
     = Yes.
Each question names the element and the properties to check --- judge exactly that: a
     clearly noticeable change in the named property = No; trivial difference = Yes. Do not
     check the edited target itself here.
Edit-consequence (applies to BOTH video and audio): ignore changes to a non-edited element
     that are a necessary/direct consequence of the requested edit --- answer Yes. For video,
      e.g. shadows, reflections, lighting, occlusion caused by the edited target; for audio,
      e.g. masking, mixing, or level changes of retained sounds caused by the edited sound.
     Only count changes unrelated to the edit.
Camera drift: a single uniform global framing/scale/position offset -> mark only the camera/
     framing question No, not every element; it does not excuse a real local change.
Aspect-ratio / frame completion: ignore any change caused by a different aspect ratio or by
     outpainting/padding that fills newly exposed areas. Do not answer "No" because of the
     aspect-ratio change itself or anything in the filled-in regions; judge only the
     originally-visible content.
If a retained-audio question is present, "No" when the audio element is significantly
     altered in character/level/pitch/content.

## Fidelity / hallucination_control --- any unrequested additions?
"No" for: extra visual objects/subtitles/text/logos/watermarks; new unrequested visual
     events; and --- for audio (if a retained-audio hallucination question is present) ---
     penalize ONLY clearly NEW added audio: added voiceover/narration, added background
     music, or a significant new independent event sound. Do NOT penalize leftover ORIGINAL
     audio that was not fully removed (e.g., partially retained voice or residual background
      noise from the source), nor minor/faint noise or low-level artifacts. Content that
     appears only because of an aspect-ratio change/frame completion, or any sound that is a
      direct consequence of the requested edit, is NOT an unrequested addition --- ignore it.


# Output
JSON only; keep each item's question_id / dimension / subdimension / modality_tag / question
      unchanged:
{
```

```
  "schema_version": "checklist_v2",
  "edit_category": "video_only",
  "visual_discrepancy_analysis": "short comparison summary",
  "evaluations": [
    {"question_id": "Q01", "dimension": "Edit Response", "subdimension": "edit_response", "
        modality_tag": "video", "question": "...", "observation": "short evidence", "answer":
         "Yes", "justification": "short reason"}
  ],
  "summary": {"total_questions": 0, "yes_count": 0, "no_count": 0, "score": 0.0}
}
"""
```

F.3.2 AUDIO-ONLY CHECKLIST EVALUATOR

```
CHECKLIST_EVALUATOR_SYSTEM_PROMPT = """# Role
Evaluate a target audio-only edit against a fixed V2 checklist (source video should stay
     unchanged). Answer every item Yes or No from direct evidence. Do not modify, add, or
     remove questions. Output valid JSON only.

# Answering Principle
- Use only directly observable/audible evidence.
- "Yes" only when the question's condition is clearly met; otherwise "No" (absent, ambiguous,
      partial, weak, or only approximate).
- Do not infer success from the edit prompt. No credit for near-misses.

# Dimensions

## Edit Response --- did the audio change vs source?
Descriptive, not correctness. Strict, low threshold: ANY audible change = Yes; "No" only if
     the target audio is essentially identical to source. Do not judge whether the change is
      correct or natural.

## Instruction Following --- was the requested edit done correctly?
"Yes" only if the requested semantics / attribute / degree / coverage / removal is clearly
     satisfied.
"No" if: wrong sound/voice/language/category; wrong loudness/intensity/rhythm/pitch/timbre/
     speaker/content; wrong target; present only briefly when full coverage required;
     partial when completion required; or original sound still present when removal required.

Do not judge: whether anything changed (Edit Response), naturalness/sync (Realism), source
     preservation (Fidelity), or extra content (hallucination_control).

## Fidelity / non_edit_preservation --- did non-edited parts stay the same?
Source-vs-target, by overall perception NOT sample-by-sample; minor/imperceptible
     differences = Yes.
Each question names the element and the properties to check --- judge exactly that: a
     clearly noticeable change in the named property = No; trivial difference = Yes. Do not
     check the edited target itself here.
Edit-consequence (applies to BOTH audio and video): ignore changes to a non-edited element
     that are a necessary/direct consequence of the requested edit --- answer Yes. For audio,
      e.g. masking, mixing, or level changes of retained sounds caused by the edited sound;
     for video, e.g. shadows/lighting/occlusion caused by the edited target. Only count
     changes unrelated to the edit.
For retained audio (timbre/tone/ambience/music/SFX/retained speech), "No" when significantly
      altered in character/level/pitch/content.
If a visual-preservation question is present, "No" only when a non-edited visual element
     changes significantly in shape, texture, appearance/color, or size. Ignore any change
     caused by a different aspect ratio or by frame completion/outpainting --- judge only
     the originally-visible content.

## Fidelity / hallucination_control --- any unrequested additions?
Penalize ONLY clearly NEW added audio: "No" for added voiceover/narration, added background
     music, a significant new independent event sound, or extra speaker turns; and extra
     visual objects/subtitles/text (if a visual hallucination question is present). Do NOT
     penalize leftover ORIGINAL audio that was not fully removed (e.g., partially retained
     voice or residual background noise from the source), nor minor/faint noise or low-level
      artifacts. Any sound that is a direct consequence of the requested edit is NOT an
     unrequested addition --- ignore it.

# Output
JSON only; keep each item's question_id / dimension / subdimension / modality_tag / question
      unchanged:
{
  "schema_version": "checklist_v2",
  "edit_category": "audio_only",
  "visual_discrepancy_analysis": "short comparison summary",
  "evaluations": [
```

```
   {"question_id": "Q01", "dimension": "Edit Response", "subdimension": "edit_response", "
       modality_tag": "audio", "question": "...", "observation": "short evidence", "answer":
        "Yes", "justification": "short reason"}
  ],
  "summary": {"total_questions": 0, "yes_count": 0, "no_count": 0, "score": 0.0}
}
"""
```

F.3.3 RESTRUCTURED JOINT/SPEECH CHECKLIST EVALUATOR

```
CHECKLIST_EVALUATOR_SYSTEM_PROMPT = """# Role
Evaluate a target output (joint or speech edit) against a fixed V2 checklist. Answer every
     item Yes or No from direct evidence. Do not modify, add, or remove questions. Output
     valid JSON only.

# Answering Principle
- Use only directly observable/audible evidence.
- "Yes" only when the question's condition is clearly met; otherwise "No" (absent, ambiguous,
      partial, weak, or only approximate).
- Do not infer success from the edit prompt. No credit for near-misses.

# Dimensions

## Edit Response --- did this modality change vs source?
Descriptive, not correctness. Strict, low threshold: ANY detectable change = Yes; "No" only
     if essentially identical to source. Do not judge whether the change is correct or
     natural.
- Video response: compare source vs target video.
- Audio response: compare source vs target audio.
- Speech: audio response expected; include video response only if the edit affects visible
     mouth/face/subtitles.

## Instruction Following --- was the requested edit done correctly?
"Yes" only if the requested semantics / attribute / degree / coverage / removal is clearly
     satisfied.
"No" if: wrong object/action/sound/voice/language/color/quantity/speed/intensity/content;
     wrong subject; present only briefly when full coverage required; partial when
     completion required; or original target still present when removal required.
Do not judge: whether anything changed (Edit Response), naturalness/sync (Realism), source
     preservation (Fidelity), or extra content (hallucination_control).

## Fidelity / non_edit_preservation --- did non-edited parts stay the same?
Source-vs-target, by overall perception NOT pixel-by-pixel; minor/imperceptible differences
     = Yes.
Each question names the element and the properties to check --- judge exactly that: a
     clearly noticeable change in the named property = No; trivial difference = Yes. Do not
     check the edited target itself here.
Edit-consequence (applies to BOTH video and audio): ignore changes to a non-edited element
     that are a necessary/direct consequence of the requested edit --- answer Yes. For video,
      e.g. shadows, reflections, lighting, occlusion caused by the edited target; for audio,
      e.g. masking, mixing, or level changes of retained sounds caused by the edited sound.
     Only count changes unrelated to the edit.
Camera drift: a single uniform global framing/scale/position offset -> mark only the camera/
     framing question No, not every element; it does not excuse a real local change.
Aspect-ratio / frame completion: ignore any change caused by a different aspect ratio or by
     outpainting/padding that fills newly exposed areas. Do not answer "No" because of the
     aspect-ratio change itself or anything in the filled-in regions; judge only the
     originally-visible content.

## Fidelity / hallucination_control --- any unrequested additions?
"No" for: extra visual objects/subtitles/text/logos/watermarks; extra spoken words or
     speaker turns; new unrequested events; and --- for audio --- penalize ONLY clearly NEW
     added audio: added voiceover/narration, added background music, or a significant new
     independent event sound. Do NOT penalize leftover ORIGINAL audio that was not fully
     removed (e.g., partially retained voice or residual background noise from the source),
     nor minor/faint noise or low-level artifacts. Content that appears only because of an
     aspect-ratio change/frame completion, or any sound that is a direct consequence of the
     requested edit, is NOT an unrequested addition --- ignore it.

# Output
JSON only; keep each item's question_id / dimension / subdimension / modality_tag / question
      unchanged:
{
  "schema_version": "checklist_v2",
  "edit_category": "joint",
  "visual_discrepancy_analysis": "short comparison summary",
  "evaluations": [
```

```
   {"question_id": "Q01", "dimension": "Edit Response", "subdimension": "edit_response", "
       modality_tag": "video", "question": "...", "observation": "short evidence", "answer":
        "Yes", "justification": "short reason"}
  ],
  "summary": {"total_questions": 0, "yes_count": 0, "no_count": 0, "score": 0.0}
}
"""
```

F.3.4 SPLIT EVALUATION USER PROMPTS

```
# Split Evaluation User Prompts

These templates are used as user prompts during checklist evaluation. The evaluator splits
     checklist items by `modality_tag` and sends separate calls for audio-only, video-only,
     and general A/V questions.

## Audio-Only

```text
# Evaluation Input (Audio-Only)

You are provided with SOURCE audio and TARGET audio only. No video frames are provided.

Source Video Caption:
{source_caption}

Edit Prompt:
{edit_prompt}

Edit Type:
{edit_type}

Checklist subset (audio questions only):
{checklist_text}

Instructions:
1. Answer every checklist item above.
2. Keep question_id, question, dimension, subdimension, and modality_tag unchanged.
3. Base judgments only on audible evidence from source/target audio.
4. Judge each dimension by the rules in the system prompt; for Fidelity, always compare
     source and target audio.

Respond in JSON:
{
  "visual_discrepancy_analysis": "short audio-focused discrepancy analysis",
  "evaluations": [
    {
      "question_id": "Q01",
      "dimension": "Instruction Following",
      "subdimension": "instruction_correctness",
      "modality_tag": "audio",
      "question": "...",
      "observation": "...",
      "answer": "Yes/No",
      "justification": "..."
    }
  ],
  "summary": {"total_questions": {len(checklist_items)}, "yes_count": 0, "no_count": 0, "
       score": 0.0}
}
```

Video-Only

```text
# Evaluation Input (Video-Only)

You are provided with SOURCE video and TARGET video only. No standalone audio files are
     provided.

Source Video Caption:
{source_caption}

Edit Prompt:
{edit_prompt}

Edit Type:
{edit_type}
```
```

````
Checklist subset (video questions only):
{checklist_text}

Instructions:
1. Answer every checklist item above.
2. Keep question_id, question, dimension, subdimension, and modality_tag unchanged.
3. Base judgments only on visual evidence from source/target video.
4. For Instruction Following state questions, target-only evaluation is allowed.
5. Judge each dimension by the rules in the system prompt; for Fidelity, always compare
     source and target video.

Respond in JSON:
{
  "visual_discrepancy_analysis": "short visual-focused discrepancy analysis",
  "evaluations": [
    {
      "question_id": "Q01",
      "dimension": "Instruction Following",
      "subdimension": "instruction_correctness",
      "modality_tag": "video",
      "question": "...",
      "observation": "...",
      "answer": "Yes/No",
      "justification": "..."
    }
  ],
  "summary": {"total_questions": {len(checklist_items)}, "yes_count": 0, "no_count": 0, "
       score": 0.0}
}
```

## General A/V

```text
# Evaluation Input (General A/V)

You are provided with SOURCE and TARGET multimodal inputs (video and audio).

Source Video Caption:
{source_caption}

Edit Prompt:
{edit_prompt}

Edit Type:
{edit_type}

Checklist subset (general questions only):
{checklist_text}

Instructions:
1. Answer every checklist item above.
2. Keep question_id, question, dimension, subdimension, and modality_tag unchanged.
3. Use both visual and audio evidence.
4. For Instruction Following state questions, target-only evaluation is allowed.
5. Judge each dimension by the rules in the system prompt; for Fidelity, always compare
     source and target.

Respond in JSON:
{
  "visual_discrepancy_analysis": "short multimodal discrepancy analysis",
  "evaluations": [
    {
      "question_id": "Q01",
      "dimension": "Instruction Following",
      "subdimension": "instruction_correctness",
      "modality_tag": "general",
      "question": "...",
      "observation": "...",
      "answer": "Yes/No",
      "justification": "..."
    }
  ],
  "summary": {"total_questions": {len(checklist_items)}, "yes_count": 0, "no_count": 0, "
       score": 0.0}
}
```
````
````

F.4 Realism Evaluation Prompt

The realism evaluator judges whether the edited target video remains physically and temporally plausible after the requested modification. Its audio fields are the Audio Artifact Score (`AAS`), which rates signal purity, and Material–Timbre Consistency (`MTC`), which rates the agreement of sound timbre with visible materials and room acoustics.

F.4.1 Realism Evaluator

```
# Realism Evaluation Prompt

```text
# Task
Rate the Realism of the TARGET edited audio-video on 5 sub-dimensions (each 1-5).
Realism = whether the target is natural, well-formed and coherent --- NOT whether it follows
      the instruction.

# Rules
- Judge the TARGET only. Edit instruction / caption are context, not requirements.
- Do NOT penalize an edit that was not performed; if an attempted change is simply absent,
     ignore it.
- Audio dims: output "NA" if there is no audio / no clear sound source. Do not default to 5
     or 1.
- Motion blur is scene-adaptive: blur from fast motion/camera is fine; blur on static
     content is a defect.

# Shared 1-5 scale
5 = fully natural & coherent, no issues.
4 = minor: slight warp/distortion/blur, recognition & perception unaffected.
3 = noticeable: looks AI-generated / illogical, or a physics violation, but no glaring error.

2 = clear error: deformation, abnormal/extra object, sudden appearance/disappearance, third
     hand,
   floating object, clipping-through, or a visibly mismatched/unblended edit region.
1 = severe / broken.

# Sub-dimensions
VISUAL
- object_integrity: each subject's own form & existence. Penalize structural errors (reverse
      joints,
  rubber limbs, extra/fused fingers, melting faces, rigid bodies bending like jelly, texture
      morphing)
  and existence MUTATIONS (objects popping in/out without occlusion, flicker, category/
      identity change).
  Reasonable occlusion/exit is fine.
- interaction_physics: physical & commonsense logic of interactions. Penalize violations of
     gravity,
  contact/collision, liquids, slicing/breaking/pouring, support (feet on ground), clipping-
      through,
  and action-object mismatch (e.g. a person on a motorcycle making a pedaling motion).
- naturalness: does it look like a real, coherently-styled video rather than an AI edit.
     Penalize an
  edited region that does not match the background (lighting, shadow, color, seam),
      inconsistent art
  style across the frame, an obvious AI-generated look, and warping/flicker/jitter artifacts.
AUDIO
- AAS: signal purity --- clipping, static, dropout, robotic/distorted sound. NA if no audio.
- MTC: does each sound's timbre match the visible material and the room acoustics. NA if no
     sound source.

# Context (scene understanding only)
Edit instruction: {edit_prompt}
Source caption: {caption}

# Output - return ONLY this JSON:
{{"object_integrity": <1-5>, "interaction_physics": <1-5>, "naturalness": <1-5>,
  "AAS": <1-5 or "NA">, "MTC": <1-5 or "NA">,
  "reason": "<one concise sentence naming the dominant issue, or 'none'>"}}
```
```