AVENUE: Audio-Video EditiNg Understanding and Evaluation

Hayeon Kim [*1] **Yoojin Jang** [*1] **Jaejun Yoo** [1]

Abstract

Audio-video (AV) editing requires models to infer a *modality-selective edit scope* from the prompt alone—determining not only what should change, but also which modality should be preserved. However, existing benchmarks provide limited edit-type and modality coverage, while current evaluation systems are modality-blind and sample-agnostic. We introduce AVENUE (**A**udio-**V**ideo **E**diti**N**g **U**nderstanding and **E**valuation), comprising: (1) a benchmark of 1,291 source clips and 7,957 editing instructions across audio-targeted, video-targeted, and AV-coupled edit types, curated and human-verified from VGGSound; and (2) a sample-specific, modality-aware evaluation framework. Our findings reveal that when editing one modality, existing models frequently induce unintended changes in the other, regardless of paradigm. The benchmark and evaluation framework are publicly available at https://huggingface.co/datasets/AVENUE-dataset/AVENUE.

1. Introduction

Recent advances in multimodal generative models have extended their applications beyond content creation to a broad range of content editing tasks. Among them, audio-video (AV) editing aims to modify both audio and video content according to a given target prompt. While recent work has proposed unified AV editing frameworks with diverse architectures, the task poses a challenge that single-modality editing does not: a model must infer **modality-selective edit-scope** from the prompt alone—determining not only what should change, but also how strongly each modality should be edited or preserved. This asymmetry is pervasive in practice: adding thunder sound to a scene requires a substantial acoustic change while the video content remains entirely intact; converting a scene to cartoon style demands a complete visual transformation with no change to the audio; and transporting a scene into a snowy setting requires both the video appearance and ambient soundscape to change in tandem. A capable AV editing model must therefore adaptively decide when to preserve or substantially modify each modality in accordance with user intent.

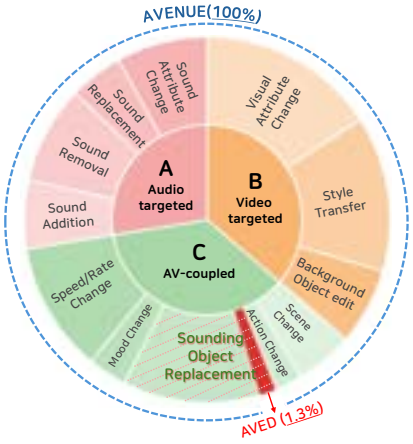

Figure 1. **AVENUE taxonomy.** AVENUE defines a three-part edit taxonomy: Ⓐ audio-targeted, Ⓑ video-targeted, and Ⓒ av-coupled edits. The outer sectors categorize the 12 fine-grained editing types across the three groups.

These challenges also demand a benchmark that explicitly evaluates the core capabilities of AV editing models. Yet existing benchmarks remain limited in both edit diversity and modality coverage. In terms of edit diversity, prior work (Lin et al., 2026) is largely confined to object-level manipulation—covering only 1.3% of the taxonomy we define (110 samples; see the red region in Table 1)—leaving attributes such as speed, action, and scene change largely unaddressed. In terms of modality coverage, existing benchmarks address only AV-coupled edits (see Table 1), overlooking audio-targeted and video-targeted edits entirely. This narrow scope makes it impossible to evaluate whether a model can selectively edit one modality while preserving the other.

Beyond data limitations, existing evaluation frameworks are fundamentally inadequate for AV editing. Current au-

[*]Equal contribution [1]Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology (UNIST). Correspondence to: Jaejun Yoo <jaejun.yoo@unist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

tomated evaluation approaches—whether based on Vision-Language Models (VLM) judges, Large Multimodal Models (LMM) scoring rubrics, or hybrid metric pipelines (Li et al., 2025; Chen et al., 2025; Zhao et al., 2025)—apply a fixed, sample-agnostic prompt to all test cases. This design is *modality-blind* (it does not explicitly measure whether the unintended modality was faithfully preserved) and *sample-agnostic* (without grounded per-sample criteria specifying what must remain intact).

To address these limitations, we introduce **AVENUE**: **A**udio-**V**ideo **E**diti**N**g **U**nderstanding and **E**valuation—a large-scale benchmark with comprehensive edit-type coverage and a modality-aware and sample-specific evaluation framework. AVENUE is built on a curated subset of VGGSound (Chen et al., 2020), yielding **7,957 editing instructions** across 4 audio-targeted, 3 video-targeted, and 5 AV-coupled edit types. Each sample is equipped with human-verified annotations specifying what changes and what must remain intact, grounding our proposed **Selective Controllability** metric alongside Edit Accuracy, Modality Selectivity, Perceptual Quality, and AV Consistency.

Our main contributions are:

1. We introduce **AVENUE**, with 7,957 editing instructions spanning 12 edit types across audio-targeted, video-targeted, and AV-coupled settings.
2. We provide per-sample human-verified annotations that specify the intended edit and elements that must remain intact, functioning as pseudo-GT to ground automated LMM judges.
3. We introduce Selective Controllability as a dedicated metric and define a modality-type-specific evaluation framework.
4. We conduct a comprehensive study of existing AV editing models across three paradigms: joint, sequential and separate.

2. AVENUE-Bench

A core observation motivating our design is that real-world editing tasks are inherently *selective*: not all edits require modifying both modalities; unnecessary changes to an irrelevant modality indicate a lack of modality-aware understanding. We organize our benchmark along two axes: *modality scope* (Audio-targeted, Video-targeted, or AV-Coupled) and *semantic category* (e.g., Instruments, Animals, Human).

2.1. Dataset Curation

We construct AVENUE through a rigorous multi-stage curation pipeline built on VGGSound (Chen et al., 2020). We source all clips exclusively from the **VGGSound test set** to avoid data leakage.

Table 1. **Comparison with existing audio-video editing benchmarks.** A, V, and AV denote audio-targeted, video-targeted, and audio-video coupled editing, respectively.

Benchmark	# Samples	Edit Types	Edit Modality	Data Access
OAVE (Liang et al., 2024)	44	add, adjust	AV	✓
Object-AVEdit (Fu et al., 2025)	-	add, remove, replace	AV	✗
VGG-Edit (Guo et al., 2026)	150	add, remove, replace	AV	✗
AvED (Lin et al., 2026)	110	replace	AV	✓
AVENUE	**7,957**	12 edit-types (add, remove, replace, ...)	Audio Video, AV	✓

Category Filtering. We group VGGSound labels into eight semantic categories: *Instruments*, *Animals*, *Human*, *Vehicles & Engines*, *Domestic & Tools*, *Sports & Leisure*, *Nature & Environment*, and *Weapons & Explosions*. We exclude *Speech* and *Others*.

Audio-Video Quality Filtering. We apply two-stage quality filtering: (1) Audio-Video quality filtering using ImageBind and CLAP scores, and (2) Editing suitable filtering comprising semantic consistency, person density, ROI consistency, and static video checks. Full details are provided in the Appendix B.1.

2.2. Editing Taxonomy

Our taxonomy is organized into three modality-level groups:

A. Audio-targeted Edits. Audio is targeted for editing; the video content must remain unchanged. Edit types include sound addition, removal, replacement, and intensity or pitch adjustment.

B. Video-targeted Edits. Video is targeted for editing; the audio must remain unchanged. Edit types include attribute change, style transfer, and background object manipulation.

C. AV-coupled Edits. Both modalities must change together in a semantically coherent manner. Edit types include sounding object replacement, scene/weather/mood changes, speed changes, and action changes.

A central design decision is that edit templates vary by semantic category. We define a *category-aware edit template* for each semantic category (see Appendix E.2), specifying which edit types are valid.

Editing Instruction Generation and Filtering. For each retained clip, we generate editing instructions using **gemini-3-flash-preview** (Team et al., 2023) with a *category-specific edit template*. Generated instructions are then filtered to retain only those that align with a valid taxonomy entry, discarding vague or malformed outputs.

2.3. Selective Controllability Evaluation

We evaluate audio-video editing quality across four complementary dimensions, grounded in human-verified per-sample annotations: `what_changed` and `what_preserved`. We employ Qwen3-Omni-30B (Xu et al., 2025) as an evaluator.

- **Edit Accuracy (EA)** measures whether the target edit instruction has been faithfully applied. Scored on a 1–5 Likert scale.
- **Perceptual Quality (PQ)** assesses the intrinsic quality and naturalness of the edited output. Scored on a 1–5 Likert scale.
- **Modality Selectivity (MS)** evaluates whether the unintended modality remains intact via five binary questions plus one sample-specific question grounded in `what_preserved`.
- **AV Consistency (AV-C)** measures the synchrony and semantic coherence between the audio and visual streams. Scored on a 1–5 Likert scale.

For all metrics, the input modality supplied to the evaluator is aligned with the edit category. Detailed evaluation prompts are provided in the Appendix E.3.

3. Experiments

3.1. Preliminaries

Audio-Video Editing Paradigms. We analyze AV editing models under three paradigms. Let v, a, and p denote the source video, source audio, and editing prompt, and let f_V, f_A, and f_{AV} denote the video, audio, and joint editing functions.

Separate paradigm independently applies single-modality models: $\hat{v} = f_V(v, p)$, $\hat{a} = f_A(a, p)$. We include RAVE+ZETA and RAVE+SDEdit.

Sequential paradigm adopts a video-to-audio pipeline: $\hat{v} = f_V(v, p)$, $\hat{a} = f_A(a, \hat{v}, p)$. We evaluate CAVE (Ishii et al., 2025) and MMAudio (Cheng et al., 2025).

Joint paradigm processes both streams within a single framework: $(\hat{v}, \hat{a}) = f_{AV}(v, a, p)$. We include AvED (Lin et al., 2026).

All models are evaluated across all three edit taxonomies. The sequential and separate paradigms share the same RAVE (Kara et al., 2024)-edited video as input.

Self-prompt Preservation. We construct a *null-edit* setting where the source and target prompts are identical (src = trg), under which a well-behaved model should produce output indistinguishable from the input.

Table 2. **Self-prompt preservation evaluation.** Higher similarity and lower LPIPS & LPAPS indicate better performance. **Bold**: best, underline: second best. Dash (–): source video used directly without video editing.

Type	Model	Video Preserv.				Audio Preserv.		
		V-CL↑	IB↑	DINO↑	LP↓	CL↑	IB↑	LPA↓
Sep.	RAVE+SDEdit	.956	.911	.939	.204	.721	.534	4.11
	RAVE+ZETA					.859	.757	**2.77**
Seq.	RAVE→CAVE					.736	.589	4.32
	Src v→CAVE	-	-	-	-	.731	.580	4.32
Joint	AvED	**.995**	**.939**	**.967**	**.110**	**.886**	**.846**	3.13

Table 2 reports preservation scores under the null-edit condition. **Separated** models enforce strict modality independence by design. **Sequential** models exhibit similar audio instability (CLAP ≈ 0.73). **Joint** achieves the best preservation scores across both modalities (V-CLIP: 0.995, CLAP: 0.886), but this raises a critical question: *does this reflect genuine editing quality, or merely a tendency to preserve the original input unchanged?*

3.2. Comprehensive Analysis

3.2.1. AV-COUPLED EDIT RESULTS

As shown in Figure 3(a), the Edit Accuracy of each paradigm varies considerably depending on the edit subtype. Sequential and Separate methods outperform Joint by a large margin on C1 and C4, yet Joint reverses this gap on C3. AV Consistency scores are nearly identical across all paradigms (∼78.5). The key differentiators are Edit Accuracy, where Joint trails by 7–8 points, and Perceptual Quality, where the gap widens to ∼15 points.

Focusing on C1, Joint achieves the lowest scores across all three metrics (EA: 48.9, PQ: 45.9, AV-C: 69.9). In contrast, C3 represents the most challenging sub-type where Joint ranks first in both EA (53.5) and AV-C (70.5), suggesting that *holistic cross-modal understanding* becomes critical for sounding object replacement.

3.2.2. RESULTS ON AUDIO-TARGETED AND VIDEO-TARGETED EDITING

Table 3 summarizes the audio modality evaluation. AVED achieves the highest modality selectivity (97.5), yet records the lowest edit accuracy (64.1) among all models. This confirms that AVED's high self-prompt preservation scores reflect a tendency to preserve rather than an ability to edit.

Separated and sequential paradigms achieve substantially higher edit accuracy (76–80), though their modality selectivity varies dramatically depending on the audio model. Under the same RAVE video backbone, ZETA attains 93.0 in audio preservation while SDEdit scores only 51.4—a gap driven entirely by the choice of audio model. The source v ablation

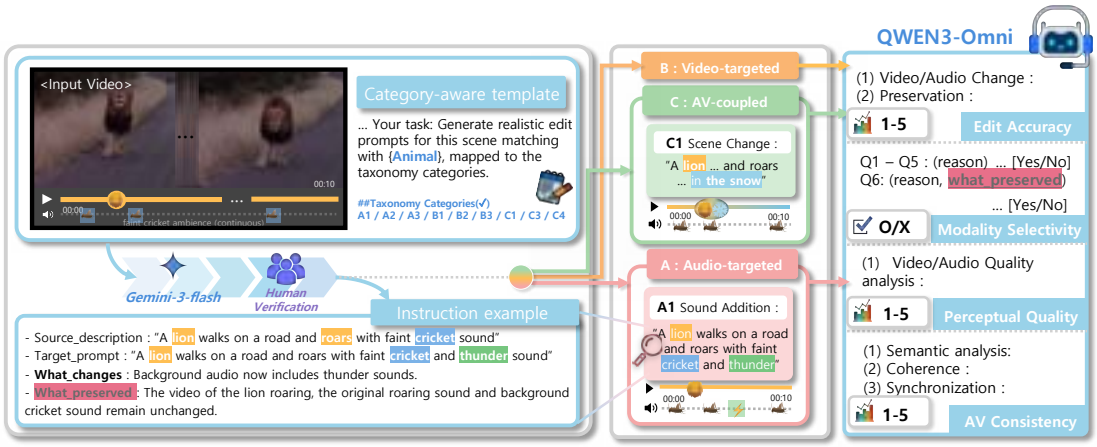

Figure 2. **AVENUE Bench.** The pipeline consists of three components: (a) category-aware instruction generation with human verification; (b) instruction examples with explicit target prompt; and (c) LMM-as-a-Judge evaluation via QWEN3-Omni.

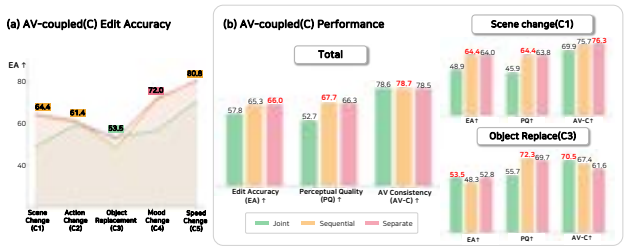

Figure 3. **Evaluation of AV-coupled editing under Cat.C.** All scores out of 100. (a) Edit Accuracy across five sub-categories. (b)–(d) Comparison across EA, PQ, and AV-C for all C, C1, and C3.

yields nearly identical scores, indicating that upstream video editing introduces minimal accumulated error.

ZETA, an inversion-based model, achieves strong performance in both edit accuracy (79.7) and modality selectivity (93.0) by preserving the source audio structure. SDEdit perturbs the input with Gaussian noise, achieving comparable edit accuracy (78.7) but substantially lower preservation (51.4). MMAudio, as a generation model, shows high perceptual quality (89.5) but the lowest modality selectivity (39.7) as it has no mechanism to retain the source audio.

These results highlight that the audio model's underlying mechanism—whether inversion-based, noise-based, or generative—is the key factor determining both editing and preservation performance. Among the evaluated models, the inversion-based approach (ZETA) is the only one that achieves competitive performance across both axes.

3.2.3. Consistency with Embedding-based Evaluation

To validate our LMM-based evaluation, we compute the Spearman rank correlation (ρ) between LMM scores and embedding-based metrics (see Appendix A5). All 12 correlations are positive and statistically significant ($p < .05$), confirming that our LMM-based scores consistently reflect the quality signal captured by embedding-based metrics.

Table 3. **Audio modality evaluation overview.** All scores are scaled to 100. EA and PQ on Cat.A; MS on Cat.B. Higher is better.

Type	Model		EA	PQ	MS
	Video	Audio	(A)↑	(A)↑	(B)↑
Sep.	RAVE +	ZETA	79.7	88.2	93.0
	RAVE +	SDEdit	78.7	82.4	51.4
Seq.	RAVE →	CAVE	79.6	89.3	58.0
	Src v →	CAVE	**79.8**	**89.7**	–
	RAVE →	MMAudio	75.8	89.5	39.7
	Src v →	MMAudio	77.6	88.8	–
Joint	AvED		64.1	87.7	**97.5**

4. Conclusion

We presented **AVENUE**, a benchmark and evaluation framework for audio-video editing, enabling the first systematic assessment of modality-selective editing across diverse edit types and model paradigms. Our evaluation reveals that no existing paradigm fully resolves this challenge: each exhibits a distinct trade-off between edit fidelity and modality preservation. Among the evaluated models, inversion-based mechanisms show the most promise. These findings indicate that modality-selective controllability remains a fundamental open challenge, and that future AV editing models should be designed with explicit awareness of modality scope—knowing not only what to change, but what to leave intact.

References

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. 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.

Chen, Y., Zhang, J., Hu, T., Zeng, Y., Xue, Z., He, Q., Wang, C., Liu, Y., Hu, X., and Yan, S. Ivebench: Modern benchmark suite for instruction-guided video editing assessment. *arXiv preprint arXiv:2510.11647*, 2025.

Cheng, H. K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., and Mitsufuji, Y. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28901–28911, 2025.

Fu, Y., Si, R., Wang, H., Zhou, D., Sun, J., Luo, P., Hu, D., Zhang, H., and Li, X. Object-avedit: An object-level audio-visual editing model. *arXiv preprint arXiv:2510.00050*, 2025.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Guo, X., Yang, X., Zhang, L., Yang, J., Wang, Z., and Luan, J. Av-edit: Multimodal generative sound effect editing via audio-visual semantic joint control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21504–21512, 2026.

Ishii, M., Hayakawa, A., Shibuya, T., and Mitsufuji, Y. Coherent audio-visual editing via conditional audio generation following video edits. *arXiv preprint arXiv:2512.07209*, 2025.

Jocher, G., Chaurasia, A., and Qiu, J. Ultralytics yolov8, 2023. URL https://github.com/ultralytics/ultralytics.

Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J. M., and Yanardag, P. 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.

Li, M., Xie, C., Wu, Y., Zhang, L., and Wang, M. Five: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. *arXiv preprint arXiv:2503.13684*, 2025.

Liang, S., Huang, C., Tian, Y., Kumar, A., and Xu, C. Language-guided joint audio-visual editing via one-shot adaptation. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1011–1027, 2024.

Lin, Y.-B., Lin, K., Yang, Z., Li, L., Wang, J., Lin, C.-C., Wang, X., Bertasius, G., and Wang, L. 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.

Manor, H. and Michaeli, T. Zero-shot unsupervised and text-based audio editing using ddpm inversion. *arXiv preprint arXiv:2402.10009*, 2024.

Mao, Y., Shen, X., Zhang, J., Qin, Z., Zhou, J., Xiang, M., Zhong, Y., and Dai, Y. Tavgbench: Benchmarking text to audible-video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6607–6616, 2024.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., and Lin, J. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Zhao, X., Zhang, P., Tang, K., Zhu, X., Li, H., Chai, W., Zhang, Z., Xia, R., Zhai, G., Yan, J., et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

A. AVENUE: Audio-Video EditiNg Understanding and Evaluation Appendix

A. AVENUE Dataset Details

A.1. Dataset Access

We provide AVENUE dataset annotation file. These are available for download from the following links: dataset annotation file (LINK) Detailed dataset statistics are discussed in Section A.2.

A.2. Dataset Statistics for AVENUE

Statistics for Semantic Category. We define eight semantic categories to enable category-aware instruction generation, as described in main Section 2.1. Figure **A**3 shows the distribution of source videos across these categories. Instruments dominates with 481 clips, followed by Animals (265) and Domestic (181), while Nature (14) and Explosions (18) are relatively underrepresented. This imbalance naturally reflects the underlying distribution of VGGSound (Chen et al., 2020), from which all clips are sourced.

Word Distribution for AVENUE. In this section, we verify the diversity of edit instructions in our dataset by visualizing the semantic distribution of editing vocabulary through word clouds constructed separately for each edit category. As shown in Figure **A**1, the edit-specific terms are distinctly distributed across audio-targeted, video-targeted, and AV-targeted categories, confirming that our dataset covers a broad and varied range of editing semantics rather than converging on a narrow set of expressions.

(a) Audio-targeted edits (A1–A4) *(b)* Visual-targeted edits (B1–B3) *(c)* AV targeted edits (C1–C5)

Figure A1. Word clouds of content words added in editing prompts, grouped by editing modality: audio-targeted (A), visual-targeted (B), and AV-coupled (C). Words are extracted by aligning source descriptions and target prompts via sequence matching, then filtered to retain nouns and adjectives targeted, excluding domain-generic terms.

B. Dataset Curation

B.1. Dataset Filtering

In this section, we describe the rigorous multi-stage filtering pipeline employed to construct a high-quality and diverse audio-video editing dataset, as mentioned in the main Section 2.1.

B.1.1. Audio-Video Quality Filtering.

To ensure that each clip contains well-aligned audio-video content, we apply two complementary audio-video correspondence metrics. We compute audio-video similarity using **ImageBind** (Girdhar et al., 2023) and audio-language alignment using **CLAP** (Wu et al., 2023), retaining only clips for which both scores exceed a threshold of 0.3. This step removes clips where the audio and video streams are poorly correlated, which would undermine the validity of AV-Coupled edit evaluation.

B.1.2. Editing Suitable Filtering.

We apply four filters to ensure that retained clips are suitable for video editing tasks.

Semantic Consistency Filter. We compute pairwise CLIP (Radford et al., 2021) similarity across 10 uniformly sampled frames (one per second) and discard clips whose mean inter-frame similarity falls below 0.75, excluding clips composed of multiple spliced scenes.

Person Density Filter. To avoid complex occlusion and identity ambiguity, we apply YOLOv8n (Jocher et al., 2023) on the 10 sampled frames and reject a clip if $\geq$3 persons are detected in $\geq$5 out of 10 frames.

ROI Consistency Filter. To ensure a coherent primary subject throughout the clip, we require that the most frequently detected object class appears in at least 9 out of 10 sampled frames. Clips failing this criterion are discarded.

Static Video Filter. To remove near-static clips that lack meaningful temporal dynamics, we compute the mean SSIM between consecutive sampled frame pairs and discard clips whose average SSIM exceeds 0.85. This threshold empirically captures clips where frames are near-identical throughout, which are unsuitable for video editing tasks. Additionally, for the high-quality test set used in model evaluation, all remaining clips were manually verified to ensure the complete absence of static content.

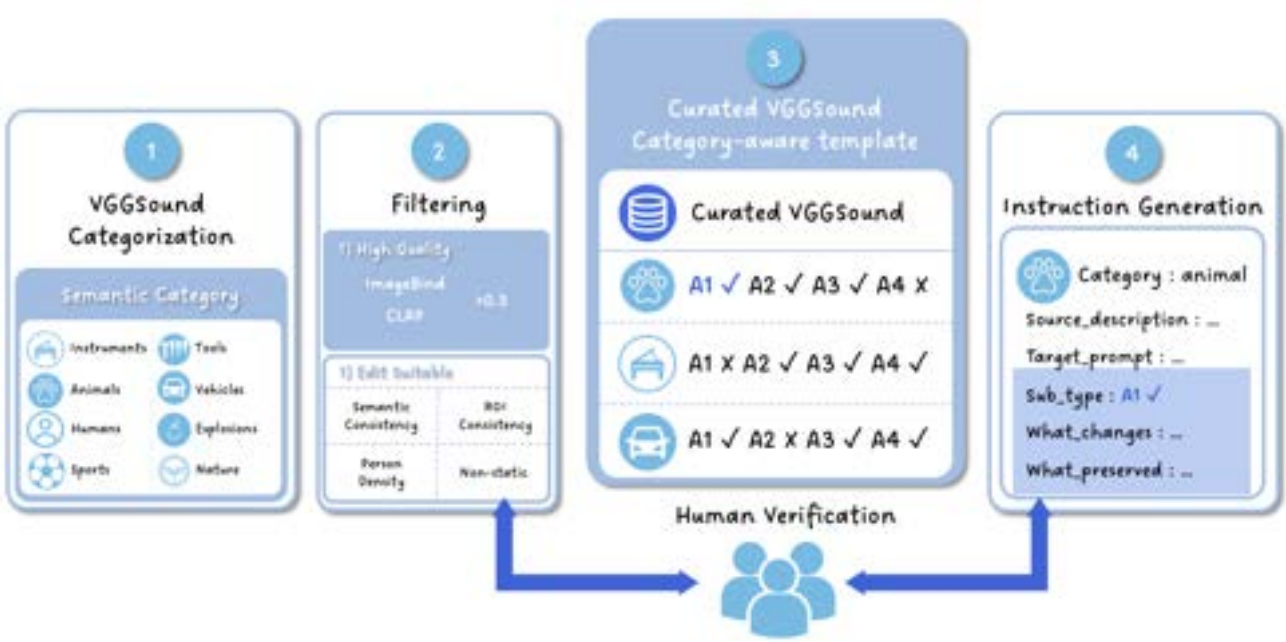

Figure A2. **Overview of AVENUE Data Curation** (1) VGGSound clips are categorized into 8 semantic categories. (2) A two-stage filtering pipeline retains high-quality, edit-suitable clips based on audio-visual alignment and edit compatibility criteria (semantic consistency, ROI consistency, person density, non-static). (3) Human annotators verify which edit types are applicable per clip, producing a curated set with category-aware edit type assignments. (4) Valid clips are paired with structured instruction templates to generate per-sample annotations including target prompt, edit sub-type, what changed and what preserved.

B.2. Human Verification

This section presents an example of the user interface employed during the human verification stage of data curation. Two expert annotators participated in verifying the edit instructions generated by the MLLM. Each annotator was presented with the MLLM-generated source prompt and target prompt corresponding to the source video, and flagged an edit instruction for the filtering pool if it satisfied any of the following criteria:

- The target prompt generated by the MLLM already exists in the source video or audio.
- The target prompt is designated as an audio-targeted edit, yet its modification would affect other modalities.
- The target prompt is semantically ambiguous.

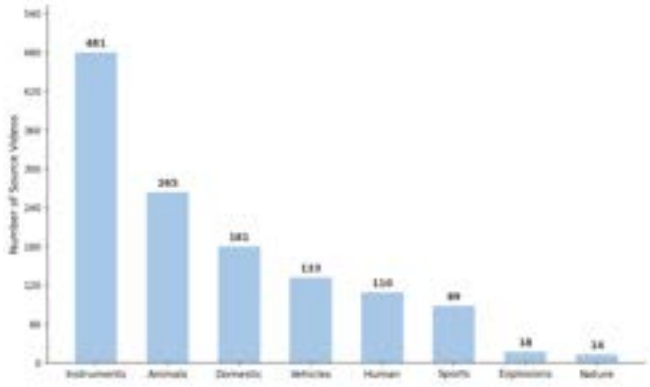

Figure A3. Statistics for semantic categories

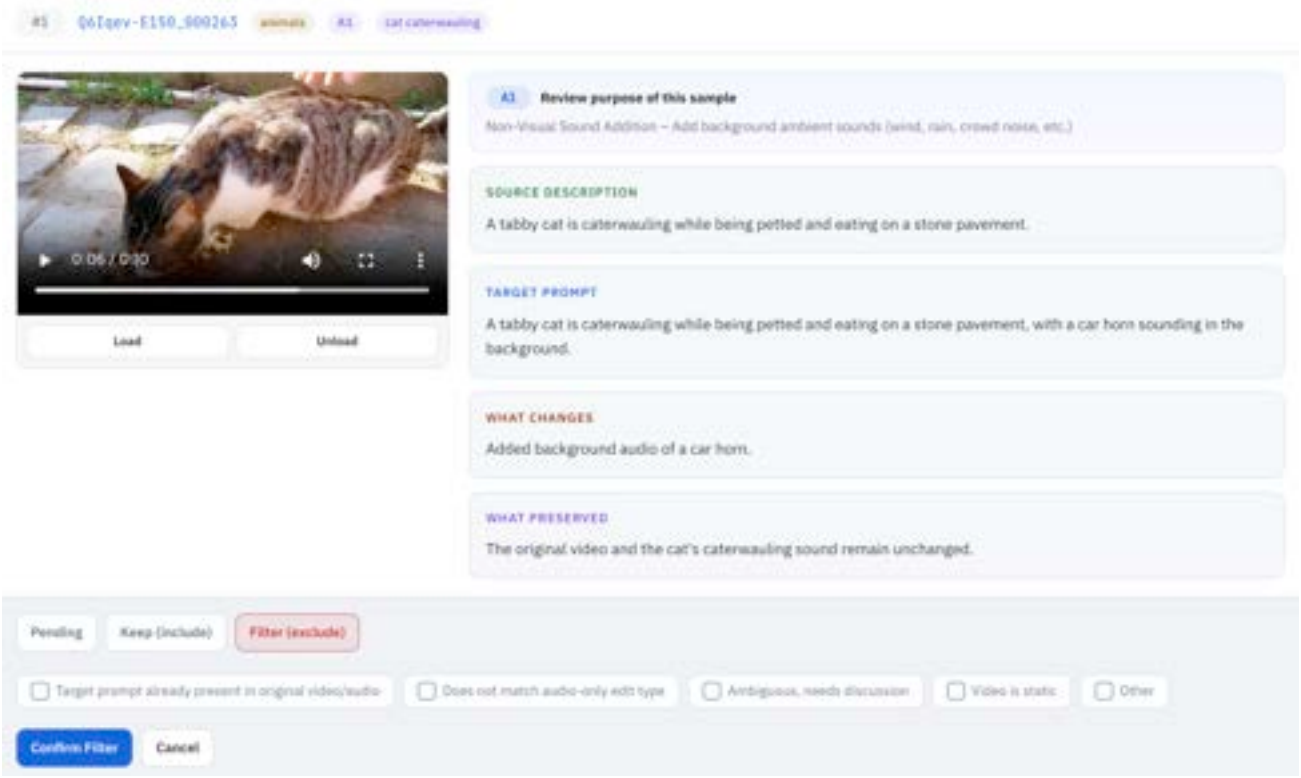

Figure A4. **Illustration of the user interface used for human verification.**

- The source video is a static (i.e., no-motion) video.
- Other reasons.

Among the flagged samples, annotators manually regenerated appropriate edit instructions for all entries except those involving static videos, which were discarded entirely. Figure A4 illustrates an example of the human verification user interface.

C. Additional Analysis

This section presents additional model analysis on our dataset, along with visual illustrations of the analyses discussed in the main. Figure A5 visualizes the mean Spearman rank correlation between LLM-based scores and embedding-based metrics,

Table A1. **Category-aware editing taxonomy.** ✓ indicates that the edit type is applicable to the given category. Edit types are grouped by modality scope: A (Audio-targeted), B (Video-targeted), and C (AV-coupled).

Group	Edit Type	Instruments	Animals	Human	Vehicles	Domestic	Sports	Nature	Explosions
A. Audio-targeted	A1: Non-visual Sound Addition	–	✓	–	–	✓	✓	✓	–
	A2: Non-visual Sound Removal	✓	✓	–	✓	✓	✓	✓	–
	A3: Non-visual Sound Replacement	✓	✓	–	–	–	–	✓	–
	A4: Sound Intensity/Pitch Change	✓	–	–	✓	–	–	✓	✓
B. Video-targeted	B1: Attribute/Color Change	✓	✓	✓	✓	✓	✓	–	–
	B2: Style Transfer	✓	✓	✓	✓	–	✓	✓	–
	B3: Background Object Edit	–	✓	✓	–	✓	–	–	–
C. AV-coupled	C1: Scene/Weather/Time Change	–	✓	–	✓	–	–	✓	–
	C2: Action Change	–	–	✓	–	–	✓	–	–
	C3: Sounding Object Swap	✓	✓	–	✓	✓	–	–	✓
	C4: Mood Change	–	✓	–	–	–	–	✓	–
	C5: Speed/Rate Change	✓	–	✓	✓	✓	–	–	✓

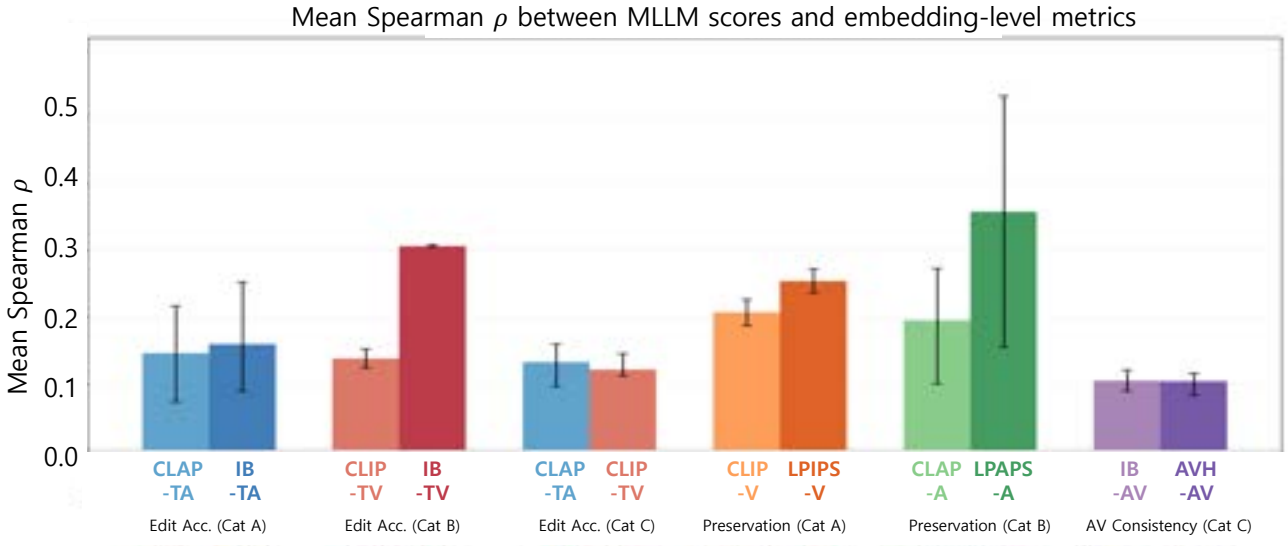

Figure A5. **Mean Spearman ρ between our LLM-based evaluation scores and embedding based metrics across all evaluation dimensions and edit categories.** Error bars indicate the min–max range across generation models. All reported correlations are statistically significant ($p < .05$). Higher ρ indicates stronger agreement between LLM-based scores and embedding-based metrics. The suffixes `ta`, `tv`, and `av` denote target text–target audio, target text–target video, and target audio–target video alignment, respectively. The single-modality suffixes `a` and `v` instead denote preservation metrics, measuring source audio–target audio and source video–target video alignment, respectively.

averaged across AV editing models, for audio-targeted, video-targeted, and AV-targeted edit categories. In Figure A5, AVHScore (Mao et al., 2024) quantifies audio-visual semantic consistency by averaging the cosine similarity between each video-frame embedding and the audio embedding in a shared embedding space. Higher values indicate that the generated audio and video are more semantically aligned. Suffixes denote the modalities used in each metric. Figure A6 visualizes the overall scores of AV editing models for each sub-category of Cat. C.

D. Implementation Details

Baselines We evaluate the models listed in Table A2, running each under its proposed default configuration without modification. For **AvED**, since its temporal shuffling operates on a 2×2 image grid, the frame rate is fixed at 4 fps. For **RAVE**, the fps is adjusted to match the original clip length, ensuring the output is generated at the source fps. For **MMAudio (CAVE)**, since MMAudio is trained on 8-second clips, inputs are automatically truncated to 8 seconds upon entry.

All models are constrained to produce outputs at the original source fps. For fair evaluation, when comparing results across models on the same sample, all outputs are trimmed to the shortest duration among all models for that sample.

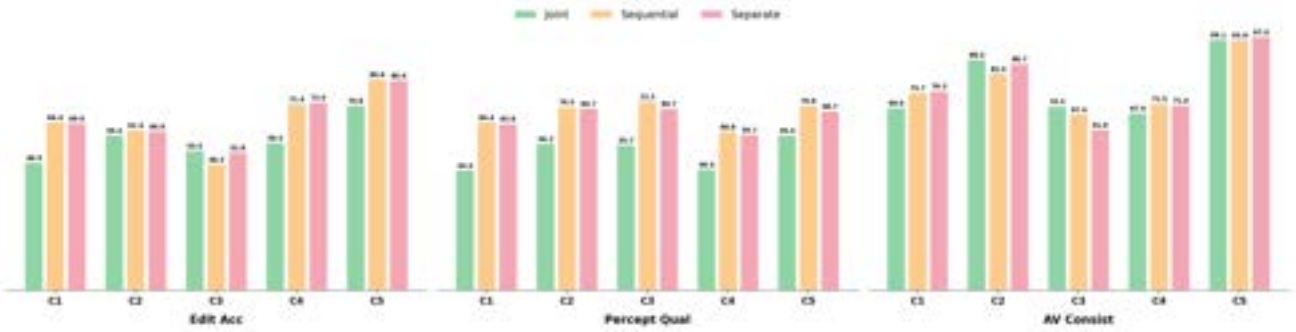

Figure A6. **Category C overall performance**

Table A2. **Existing AV editing models evaluated across our proposed edit taxonomies.** *Video editing via shared RAVE (Kara et al., 2024) backbone.

Model	Paradigm	A(Audio)	B(Video)	C(AV)
AvED (Lin et al., 2026)	Joint	✓	✓	✓
CAVE (Ishii et al., 2025)	Sequential	✓	✓*	✓
MMAudio (Cheng et al., 2025)	Sequential	✓	✓*	✓
RAVE + SDEdit (Kara et al., 2024; Meng et al., 2021)	Separate	✓	✓*	✓
RAVE + ZETA (Kara et al., 2024; Manor & Michaeli, 2024)	Separate	✓	✓*	✓

Regarding audio sampling rates, **ZETA** and **SDEdit** are based on AudioLDM2 and thus operate at 16,000 Hz. **MMAudio** uses the `small_16k` variant to match this rate. **CoherentAVEdit** provides a fine-tuned model built on top of MMAudio, which also operates at 16,000 Hz.

Computational Cost All models are evaluated on a single NVIDIA A6000 (48GB) GPU with a batch size of 1. Audio-based models — **CoherentAVEdit**, **MMAudio**, **ZETA**, and **SDEdit** — complete inference within 1 minute per sample. **AvED** requires approximately 20 minutes per sample due to its 2×2 grid-based temporal shuffling mechanism, while **RAVE** takes approximately 35 minutes per sample.

Experimental Setup To comprehensively assess how each editing paradigm handles the *audio modality*, we evaluate models using samples from two benchmark categories. First, using Cat.A (audio-targeted editing) samples, we measure edit accuracy and perceptual quality to quantify how faithfully and naturally each model performs the intended audio edit. Second, using Cat.B (video editing) samples, where only the video is edited, we measure modality selectivity to assess whether the audio is preserved intact—a critical requirement for practical editing. For sequential paradigms, where the audio model takes the video model's output as input, we additionally include an source v condition that bypasses the video editing stage. This ablation isolates the audio model's inherent capability from any artifacts propagated through the sequential pipeline, revealing whether performance differences stem from the audio model itself or from upstream video editing.

E. Prompt Detail

E.1. Instruction Generation

Prompt Instruction Generation(Gemini-3-Flash-preview)

You are an expert in audio-visual editing evaluation. You will be given a video clip and its audio label describing the sound in the scene.
Your task: Generate realistic edit prompts for this scene matching with category, mapped to the taxonomy categories below.
Taxonomy Categories : {category_taxonomies}

Instructions
1. Analyze the video and audio label to understand the scene.
2. For EACH sub-type above, determine whether a plausible edit exists for this specific scene.
3. If a plausible edit exists, generate it. If not (e.g., no background object to remove), skip that sub-type.
4. Use the exact format below for each edit.

Output Format
Return a JSON list. Each entry must follow this structure:
”category”: ”A—B—C”,
”sub_type”: ”A1—A2—A3—A4—B1—B2—B3—C1—C2—C3—C4—C5”,
”source_description”: ”describe the original scene in one sentence”,
”target_prompt”: ”the editing instruction or target description”,
”what_changes”: ”briefly describe what should change¿”,
”what_preserved”: ”briefly describe what must stay the same”

Example
...(ommision)
Important Rules
- Generate ONLY edits that are semantically plausible for the given scene. Do not force unnatural edits.
- The source_description should accurately reflect the video and audio label.
- The target_prompt must be a complete scene description (not just the delta), written naturally.
- For Audio-Only edits: the target_prompt should describe the full audio scene.
- For Video-Only edits: the target_prompt should describe the full visual scene.
- For AV-Coupled edits: the target_prompt should describe both modalities.

Now, here is your input:
Audio Label: {audio_label}
Video: {attached video}
Generate all plausible edit prompts mapped to the taxonomy.

E.2. Category-aware editing taxonomy(category_taxonomies.txt)

Prompt animal.txt

Category A: Audio-Only Edits (video must stay unchanged) A1. Ambient Sound Addition – Add background environmental sounds (e.g., wind, rain, rustling leaves) A2. Background Sound Removal – Remove background sounds while keeping the animal's sound intact A3. Background Sound Replacement – Replace background ambient sounds with different environmental sounds
Category B: Video-Only Edits (audio must stay unchanged) B1. Animal Attribute Change – Modify the animal's appearance (e.g., fur color, size, pattern) B2. Style Transfer – Apply a visual style to the scene (anime, watercolor, sketch, etc.) B3. Background Vegetation Edit – Add or remove background elements like trees, grass, or bushes
Category C: AV-Coupled Edits (both audio and video must change together) C1. Scene Change – Change the environment or setting (e.g., mountain → beach, forest → desert) C3. Animal Swap/Duplicate – Replace the animal with a different species, or duplicate the same animal in the scene C4. Mood Change – Modify the overall atmosphere or emotional tone of the scene

Prompt domestic_and_tools.txt

Category A: Audio-Only Edits (video must stay unchanged) A1. Domestic Ambient Addition – Add everyday background sounds (e.g., TV noise, clock ticking, kitchen hum) A2. Noise Removal – Remove extraneous noise or unwanted background sounds
Category B: Video-Only Edits (audio must stay unchanged) B1. Tool Color Change – Modify the color or appearance of the tool/appliance B3. Background Object Addition – Add a new object to the background of the scene
Category C: AV-Coupled Edits (both audio and video must change together) C3. Tool Swap/Add/Remove – Add, remove, or replace the tool or appliance with a different one C6. Machine Speed Change – Modify the operating speed of the machine or tool

Prompt human_sounds.txt

Category B: Video-Only Edits (audio must stay unchanged) B1. Person Attribute Change – Modify clothing color, hairstyle, or other human appearance attributes B2. Style Transfer – Apply a visual style to the scene (anime, watercolor, sketch, etc.) B3. Indoor Object Swap – Add, remove, or replace indoor objects/furniture in the background
Category C: AV-Coupled Edits (both audio and video must change together) C2. Action Change – Replace the person's action with a different one (e.g., clapping → snapping fingers) C6. Motion Speed Change – Modify the speed of the person's movement or action (e.g., slow clap → fast clap)

Prompt instruments.txt

Category A: Audio-Only Edits (video must stay unchanged) A2. Background Instrument Removal – Remove other instruments playing together in the background, isolating the main instrument A3. Background Instrument Replacement – Replace background ensemble instrument sounds with different instruments A4. Pitch/Loudness Shift – Change the pitch or loudness of the performance
Category B: Video-Only Edits (audio must stay unchanged) B1. Instrument Color Change – Modify the color or finish of the instrument (e.g., black piano → white piano) B2. Style Transfer – Apply a visual style to the scene (anime, watercolor, sketch, etc.)
Category C: AV-Coupled Edits (both audio and video must change together) C3. Instrument Swap/Add – Remove or replace the instrument with a different one; if the scene is filmed from a distance, add another instrument to the ensemble C5. Instrument Variant Change – Change the variant or type of the same instrument family (e.g., acoustic guitar → electric guitar) C6. Performance Speed Change – Modify the tempo/speed of the performance (e.g., slow ballad → fast virtuoso)

Prompt nature_and_environment.txt

Category A: Audio-Only Edits (video must stay unchanged) A1. Sound Addition – Add a new natural sound to the scene (e.g., bird calls, insect chirping) A2. Sound Removal – Remove a specific sound layer from the scene A3. Sound Replacement – Replace a natural sound with a different one (e.g., seagull calls → crow cawing) A4. Sound Intensity Change – Adjust the intensity or loudness of a natural sound (e.g., gentle waves → crashing waves)
Category B: Video-Only Edits (audio must stay unchanged) B2. Style Transfer – Apply a visual style to the scene (anime, watercolor, sketch, etc.)
Category C: AV-Coupled Edits (both audio and video must change together) C1. Weather Change – Change the weather or atmospheric conditions (e.g., sunny → rainy, calm → stormy) C4. Mood Change – Modify the atmosphere to evoke a different emotion (e.g., peaceful forest → eerie forest)

Prompt sports_and_leisure.txt

Category A: Audio-Only Edits (video must stay unchanged) A1. Crowd Sound Removal – Remove crowd cheering or audience noise from the scene A2. Crowd Sound Addition – Add crowd cheering, applause, or audience reactions
Category B: Video-Only Edits (audio must stay unchanged) B1. Attribute Change – Modify appearance attributes (e.g., jersey color, equipment color) B2. Style Transfer – Apply a visual style to the scene (anime, watercolor, sketch, etc.)
Category C: AV-Coupled Edits (both audio and video must change together) C2. Intensity Change – Reduce or increase the intensity of the athletic action (e.g., weaker hit, slower run, softer kick)

Prompt vehicles_and_engines.txt

Category A: Audio-Only Edits (video must stay unchanged) A2. Engine Sound Removal – Remove or silence the engine/motor sound A4. Engine Loudness Change – Adjust the loudness of the engine sound (e.g., louder revving, quieter idle)
Category B: Video-Only Edits (audio must stay unchanged) B1. Vehicle Color Change – Modify the color or paint of the vehicle B2. Style Transfer – Apply a visual style to the scene (anime, watercolor, sketch, etc.)
Category C: AV-Coupled Edits (both audio and video must change together) C1. Time-of-Day Change – Change the lighting/time of the scene (e.g., daytime → nighttime) C3. Vehicle Type Swap – Replace the vehicle with a different type (e.g., car → police car, sedan → ambulance) C6. Vehicle Speed Change – Modify the speed of the vehicle (e.g., cruising → racing, slow → fast)

Prompt weapons_and_explosions.txt

Category A: Audio-Only Edits (video must stay unchanged) A4. Explosion/Gunshot Intensity Change – Adjust the loudness or pitch of explosion or gunfire sounds
Category C: AV-Coupled Edits (both audio and video must change together) C3. Weapon Swap – Replace the weapon with a different type (e.g., machine gun → laser gun, pistol → shotgun) C6. Firing Rate Change – Modify the speed of firing or detonation (e.g., single shot → rapid fire)

E.3. LMM-as-a-Judge

Prompt Edit Accuracy(EA) - Audio targeted edit

You are an expert evaluator for audio-video editing quality. Your task is to assess whether an AUDIO edit was performed correctly.

[Source audio]: (attached)
[Edited audio]: (attached)

[Source description]: "source_description"
[Target description]: "target_prompt"

=== EVALUATION CRITERIA (human-verified) ===
Changes intended:
"what_changed"

Elements that must remain intact:
"what_preserved"
==========

Evaluate the edited audio based on the criteria above.

Scoring rubric (1–5):
5: All intended changes are fully and precisely realized, and all preserved elements remain completely intact.
4: Intended changes are mostly realized with minor inaccuracies (e.g., correct sound type but slightly off in quality or timing), preserved elements are intact.
3: Intended changes are partially realized (e.g., some changes applied correctly, others missing or weak), OR preserved elements show minor unintended damage.
2: Intended changes are attempted but mostly incorrect or incomplete, OR preserved elements are significantly damaged.
1: No meaningful change was made, OR changes go in the wrong direction, OR preserved elements are destroyed.

Provide your response in this format:
(1) Change assessment: For each intended change, describe whether and how well it was achieved.
(2) Preservation assessment: For each preserved element, describe whether it remains intact.
(3) Score: [1–5]

Prompt Perceptual Quality (PQ) - Audio targeted edit

You are an expert evaluator for audio quality assessment. Your task is to assess the perceptual quality of an edited audio track.
Do NOT evaluate whether the edit content is correct — focus ONLY on technical and perceptual quality.
[Edited audio]: (attached)

Assess the following aspects:
- Clarity and crispness of the audio
- Smoothness of transitions (no abrupt cuts or jumps)
- Natural timbre and realistic sound texture
- Temporal continuity (no sudden gaps, loops, or repetitions)
- Absence of audio artifacts (clipping, noise, distortion, echo, robotic sound, unnatural reverb)

Scoring rubric (1-5):
5: Fully natural and artifact-free. Sounds like a professional real-world recording.
4: Mostly natural with minor artifacts (e.g., slight background hiss, barely noticeable clipping).
3: Noticeable artifacts but content is clearly recognizable (e.g., intermittent noise, partial distortion, slightly unnatural transitions).
2: Severe artifacts that significantly degrade listening experience (e.g., persistent noise, heavy distortion, robotic timbre, repeated audio glitches).
1: Nearly unrecognizable or completely broken (e.g., extreme static, fully corrupted audio, incomprehensible sound).
Provide your response in this format:
(1) Quality analysis: Describe the audio's quality, noting specific strengths and artifacts found.
(2) Score: [1-5]

Prompt Modality Selectivity (MS) - Audio targeted edit

You are an expert evaluator for modality preservation in audio-video editing.
This is an AUDIO-ONLY edit. The video must remain UNCHANGED. Your task is to assess how well the video was preserved.
[Source video]: (attached) [Edited video]: (attached)
[Source description]: ”source_description” [Target description]: ”target_prompt”
=== PART A: General Preservation Check ===
Answer each question with Yes (1) or No (0).

Q1: Does the same primary subject from the source video still exist in the edited video?
Q2: Is the primary subject’s action/motion preserved identically to the source?
Q3: Is the background (location, structure, layout) identical to the source?
Q4: Is the edited video free from any new blur, artifacts, or distortions not present in the source?
Q5: Are the overall color tone, lighting, and visual style preserved identically to the source?
=== PART B: Sample-Specific Preservation Check ===
The following elements were verified by human annotators as elements that MUST be preserved:
”what_preserved”

Q6: Are ALL of the specified preserved elements fully maintained in the edited video without any alteration? (Yes=1 / No=0) If No, explain which specific elements were altered and how.

==========

Provide your response in this format:
Q1: (brief evidence) - [Yes/No] Q2: (brief evidence) - [Yes/No] Q3: (brief evidence) - [Yes/No] Q4: (brief evidence) - [Yes/No] Q5: (brief evidence) - [Yes/No]
Q6: (brief evidence, reference specific elements from preservation criteria) - [Yes/No]

Total score: [sum] / 6

Prompt AV Consistency (AV-C)

You are an expert evaluator for audio-video coherence assessment. Your task is to assess whether the audio and video in an edited result are semantically coherent and well-synchronized.
[Edited video with audio]: (attached)
=== EDIT CONTEXT (human-verified) ===
The following changes were intended in this AV-coupled edit: "what_changed"
==========

Assess the following aspects:
- Semantic coherence: Do the audio and video content match each other meaningfully? (e.g., if a dog is shown, is a dog-related sound heard?)
- Temporal synchronization: Are audio events aligned with visual events in timing? (e.g., impact sound at the moment of visual contact)
- Contextual consistency: Given the intended edit, do both modalities reflect the changes in a unified, believable way?

Scoring rubric (1-5):
5: Perfect coherence. Audio and video correspond precisely — sounds match visual content, timing is synchronized, and the intended edit is reflected coherently in both modalities. 4: Mostly coherent with minor misalignment (e.g., slight timing offset, or one subtle element doesn't quite match between modalities).
3: Related but noticeable mismatch (e.g., generally correct category of sound for the scene, but specific details don't align, or timing is clearly off).
2: Weak relation between audio and video (e.g., both modalities seem edited independently without awareness of each other).
1: Completely unrelated or contradictory (e.g., visual shows one scene while audio represents an entirely different context).

Provide your response in this format:
(1) Semantic analysis: Describe what is seen in the video and what is heard in the audio.
(2) Coherence assessment: Analyze how well the audio and video correspond to each other, with specific examples.
(3) Synchronization assessment: Comment on temporal alignment between audio and visual events.
(4) Score:]1-5]

F. Limitations

While AVENUE enables systematic evaluation and analysis of existing AV editing models across audio-targeted and video-targeted modalities with diverse edit categories — aspects largely overlooked in prior benchmarks — our work has several limitations. First, the edit instructions in our benchmark are generated by an Omni LLM, which may introduce model-specific biases into the dataset. Second, although we apply rigorous multi-stage filtering, the inherent noisiness of the source dataset VGGSound may leave residual artifacts in audio or video quality. Third, our edit instructions are currently composed in declarative prompt form, and do not cover diverse instruction styles such as imperative commands. Fourth, our paradigm analysis is limited to sequential, joint, and separated architectures, whereas a broader range of AV editing paradigms may exist. We aim to address these limitations in future work.