AVFakeBench: A Comprehensive Audio-Video Forgery Detection Benchmark for AV-LMMs

Supplementary Material

A. More details of AVFakeBench

A.1. Statistics of each scenario

AVFakeBench includes 3,000 audio-video pairs across 11 real-world scenarios, with distribution statistics presented in Table 6.

Table 6. Statistics of scenarios and sample counts in AVFakeBench.

Scenarios	Subjects	Samples
Human Speech	Human Subject	1500
Natural Landscapes	General Subject	145
Animals	General Subject	232
Social Activities	General Subject	117
Music Performances	General Subject	230
Transportation	General Subject	186
Daily-life Scenes	General Subject	245
Sports	General Subject	144
Industrial Operations	General Subject	68
Alarm Signals	General Subject	68
Science	General Subject	65

A.2. Data Sources

The audio–video clips categorized under the Human Subject are primarily sourced from existing public datasets and reorganized according to our 7-category taxonomy. Below are the detailed description for these datasets.

DDL [23]. The DDL dataset originates from The Deepfake Detection and Localization Challenge, encompassing cross-modal forgery types and precise temporal localization of forgeries. From this dataset, we select 500 real samples, which are subsequently categorized under the "Human Subject-Real" class.

DigiFakeAV [36]. DigiFakeAV is a large-scale multimodal Deepfake benchmark dataset generated using diffusion models. The construction of this dataset begins with the selection of high-quality real videos from HDTF and CelebV-HQ. During the synthesis phase, five state-of-the-art video diffusion models along with the audio generation model are employed. By conditioning the audio (real or synthesized) and reference images as inputs, the synthesis pipeline generates high-fidelity video frames through diffusion sampling, followed by rigorous quality control to ensure audiovisual consistency. Since the dataset provides open access to real audio and real video, real audio and synthesized video, and single-modal synthesized video, we directly select 150 samples of real audio paired with synthesized video, categorizing them as the "Human Subject-Real Audio & Synthesized Video" class. For the single-modal synthesized videos, 200 samples are chosen and re-synthesized using LipVoicer, which are classified under the "Human Subject-Synthesized Audio & Synthesized Video" category.

AV-DeepFake1M [5]. AV-DeepFake1M is a large-scale multimodal benchmark specifically tailored for the task of temporal deepfake localization. The dataset construction employs a sophisticated three-stage pipeline: First, Large Language Models (ChatGPT) are leveraged to perform context-aware word-level manipulations—including replacement, deletion, and insertion—on original transcripts to generate semantically inverted text. Subsequently, high-fidelity Text-to-Speech models synthesize the corresponding audio, which is fused with background noise to enhance realism. Finally, the Audio-to-Video model generates visual frames with precise lip-synchronization to the synthetic audio, resulting in high-quality forgery data spanning three distinct modality configurations: "Edited Audio-Edited Video," "Edited Audio-Real Video," and "Real Audio-Edited Video". We select 150 real videos from AVDeepfake1M and generate synthetic audios using LipVoicer and TTS, followed by temporal alignment with the original video. These samples are reorganized into "Synthesized Audio & Real Video".

LAVDF [4]. Similar to AV-DeepFake1M, LAV-DF is a large-scale multimodal Deepfake benchmark dataset designed to address the task of content-driven temporal forgery localization and detection, containing over 136,000 video samples. From this dataset, we select 150 Edited Audio & Real Video, 150 Real Audio & Edited Video, and 200 Edited Audio & Edited Video clips. Among them, the distribution of insertion, deletion, and replacement manipulations is balanced across all forgery types.

Data Selection Criteria. To ensure the robustness and reliability of the constructed benchmark, we implement a rigorous manual screening process to filter high-quality samples from the aforementioned source datasets. The selection criteria for "high-quality" data are defined as follows:

- **Perceptual Fidelity:** The video frames must maintain high resolution without significant compression artifacts, blurring, or lighting inconsistencies. Similarly, audio

You are provided with a real video frame depicting a specific scene. Based on this frame, imagine a plausible continuation of the video, particularly focusing on the dynamic changes in action within the scene. Describe how the scene evolves naturally, including the following aspects:

1. What are the key actions taking place in the scene (e.g., people moving, objects interacting, environmental changes)?
2. How do these actions progress logically over time (e.g., how would a person move or how objects would behave)?
3. What new elements or events might logically appear based on the context of the current frame (e.g., additional characters, background changes, environmental interactions)?
4.How do the visual and physical actions transform or escalate as the scene progresses?

Please ensure that your description focuses on maintaining continuity in terms of motion, character behavior, and environmental interaction. The goal is to create a coherent and realistic extension of the given video frame.

Figure 5. The prompt template used for generating dynamic descriptions.

You are given a list of 10 real-world scenes. For each scene, imagine and describe the visual details in a way that allows an artist or a generative model to create a corresponding image. Your task is to provide a detailed scene description for each of the following categories:

1. Natural Landscapes: Describe a breathtaking landscape, including its natural elements like mountains, rivers, forests, and skies. Consider the lighting, weather conditions, and time of day.
2. Animals: Describe a scene featuring animals in their natural environment. Be specific about the animals' actions, interactions, and surroundings.
3. Social Activities: Imagine a social gathering, such as a family reunion, a festival, or a casual meeting in a park. Focus on the people's emotions, interactions, clothing, and the setting.
4. Music Performances: Visualize a live music performance. What type of music is being performed? Describe the stage, performers, audience, lighting, and mood.
5. Transportation: Imagine a busy transportation scene, such as a bustling train station, a traffic jam on a highway, or an airplane taking off at sunrise. Describe the vehicles, people, and any notable features like weather conditions, motion, or architecture.
6. Daily-Life Scenes: Depict a moment from everyday life. This could include scenes like a person walking their dog in the park, cooking in the kitchen, or working at a desk. Focus on the ordinary but specific details that make the scene feel real.
7. Sports: Picture a sporting event. Describe the players, the action of the game, the stadium or field, and the audience. Pay attention to the movement, the intensity, and the environment.
8. Industrial Operations: Visualize a scene from an industrial setting, like a factory floor, a construction site, or a warehouse. Describe the machines, workers, equipment, and any movement or process taking place. How does the environment look (e.g., machinery, raw materials, lighting)?
9. Alarm Signals: Picture a scene involving an alarm or emergency situation. This could be a fire drill, a car accident, or an emergency alert in a city. Describe the urgency, the individuals involved, the emergency response, and the surrounding environment.
10. Science: Imagine a scientific experiment or laboratory setting. Describe the equipment, scientists, and the process being conducted.

For each scene, ensure the description includes enough detail to generate an image that conveys the mood, atmosphere, and key elements of the environment.

Figure 6. The prompt template used for generating static scenes.

clips must be clear, with minimal background noise (unless intentionally added for realism) and devoid of robotic or metallic artifacts typical of low-quality synthesis.

- **Audiovisual Synchronization:** For samples involving speech, there must be precise alignment between lip movements and audio streams. Samples exhibiting noticeable desynchronization or unnatural lip-sync jitter are excluded to ensure the challenge arises from subtle forgery traces rather than obvious alignment errors.
- **Semantic Consistency:** For text-edited or context-driven forgeries (e.g., from AV-DeepFake1M), the manipulated content must maintain semantic fluidity. We discard samples where the manipulated speech resulted in grammatical incoherence or logical breaks that would make the forgery trivially detectable by humans.

Based on these criteria, we curate a balanced subset of samples that represent the challenging and realistic scenarios in current deepfake generation.

A.3. Forgery Framework

The construction of AVFakeBench leverages a proprietary model to assist in generating dynamic descriptions, generating static scenes and proposing a plausible manipulation. The specific prompt templates used for dataset construction are summarized below.

Prompt Templates to generate dynamic descriptions.

Fig. 5 illustrates the prompt template used to instruct LMMs to generate dynamic descriptions based on the first frame extracted from a video.

Prompt Templates to generate static scenes. Fig. 6 illustrates the prompt template used to instruct LMMs to generate static scenes based on the 10 scenarios belonging to General Subject.

Prompt Templates to propose a plausible manipulation. Fig. 7 illustrates the prompt template used to instruct LMMs to propose a plausible manipulation based on 8-frames extracted from a real video.

Table 7. Generation parameters for proprietary commercial models

Model Name	Model Version	Key Settings		
		Setting 1	Setting 2	Setting 3
Midjourney	Default Model (V6.1)	Raw Mode	Stylize Med	Subtle Variation
KLING	Video 2.1	Resolution: 720P	Length: 10s	Aspect Ratio: 16:9

Generation parameters for proprietary commercial models. As shown in the table 7, we have listed in detail a series of generation parameters for the closed-source business model in the data forgery framework to ensure reproducibility.

Human Supervision Details. To ensure the high perceptual quality, logical consistency, and precise localization of the General Subject component, we implement a robust human-in-the-loop supervision protocol spanning the entire generation pipeline. This protocol involves human supervision at three key phases:

- **Phase 1: Data Preparation and Filtering.** During the collection of real samples from VGGSound, we conduct a comprehensive screening process to ensure the data's relevance and quality. The screening involves assessing the *scenario alignment*, which verifies that the content precisely matches one of the 10 defined scenarios, and the *audio-visual quality*, which involves discarding clips that exhibit low resolution, excessive motion blur, or background noise that interfers with the primary audio event. These stringent criteria ensured the data's consistency and suitability for further processing.
- **Phase 2: Synthesis Consistency Check.** In the synthesis branch, supervision is directed at ensuring both physical plausibility and semantic accuracy. We rigorously review the static visual anchors generated by the T2I model in Stage 1, rejecting any samples that contain obvious artifacts, unrealistic textures, or discrepancies in visual fidelity. Additionally, the dynamic descriptions generated by the LMM are carefully examined to ensure they conform to physical laws, such as proper gravity and motion logic. For the final output in Stage 2, we meticulously verify the synchronization between the generated video dynamics and the Foley-synthesized audio, ensuring seamless cross-modal coherence and preventing any temporal mismatches that may disrupt the audiovisual experience.
- **Phase 3: Editing Precision and Constraints.** In the editing phase, human intervention plays a crucial role in ensuring spatial and temporal precision. Initially, we validate the *feasibility* of the LMM-proposed editing instructions to prevent any logical conflicts, such as attempting to remove an object that does not exist in the scene. During the execution stage (Stage 2), annotators manually refine the segmentation masks provided by SAM2 to correct any boundary errors. Specifically, we specify the exact coordinates of the editing bounding boxes, ensuring that the editing operations are confined to the correct regions and that no unintended elements are affected. This step is critical for maintaining the spatial accuracy of the edits and preventing "hallucinations"—unrealistic artifacts introduced into non-target regions of the background. Furthermore, the final output underwent a meticulous quality control process, where we check the transitions between the real and forged segments to reduce visual inconsistencies, such as jumps or distortions in the frame, or acoustic artifacts, such as mismatches between sound and visual cues, are present.

A.4. Dataset Annotation

Different types of input used for annotation: As shown in Fig. 8, we present examples of the four different types of input used for annotation: Video Frames, Motion Heatmaps, Log-Mel Spectrograms, and High-Frequency Zooms. Each input type contributes distinct evidence that the LMM uses for analysis.

Video Frames (Top Row): These frames provide spatial appearance and serve as the primary visual evidence for forgery detection. Each frame shows a different moment in time within the video clip, offering a detailed view of the manipulated content.

Motion Heatmaps (Second Row): These heatmaps highlight temporal motion patterns, showing areas of significant change in the video. They are particularly useful for detecting anomalies in object movement or forgeries that involve dynamic elements, such as unnatural shifts in the scene or motion artifacts.

Log-Mel Spectrogram (Third Row): The full Log-Mel Spectrogram reveals the frequency-domain characteristics of the audio, helping to identify anomalies in the audio track that correspond to visual forgeries. It offers insights into the temporal structure of the sound, making it valuable for detecting mismatches between the audio and visual content.

High-Frequency Zoom (Fourth Row): This zoomed-in view of the spectrogram focuses on the high-frequency range, capturing subtle audio artifacts such as blending traces, tampering noise, or unnatural transitions in the sound.

You are provided with 8 evenly sampled frames from a real video, each with a timestamp labeled at the top-left corner. Your task is to generate a reasonable edit operation based on these frames. The possible editing operations include insertion, deletion, and replacement of specific elements within the frames. For each edit, you must specify the following:

Edit Type: Choose one of the following edit types:
1. Insert: Insert a new element (e.g., an object, person, or background feature) into the scene.
2. Delete: Remove an existing element (e.g., an object, person, or part of the background).
3. Replace: Replace an existing element with a new one (e.g., replacing an object with a different version, or changing the background).

Time Period for Edit: Specify the exact time range (start time and end time) of the edit. The edit should be applied between two or more frames, so provide the timestamps that indicate where the change should begin and end.

Edited Region: Identify the specific region(s) in the frame that will be edited. This could refer to:

A specific object or person in the frame (e.g., replacing a car with a different model, deleting a person walking, etc.).
A particular part of the background (e.g., replacing a cloudy sky with a clear one, or removing an object from the background).

Description of Edit: Provide a detailed description of what is being modified. For example:

For insertion, describe what new element is being added to the frame (e.g., inserting a tree in the foreground or adding a new character walking through the scene).
For deletion, explain what element is being removed (e.g., deleting a person walking on the street).
For replacement, describe what element is being swapped (e.g., replacing a blue car with a red one).

Ensure the edit you propose is coherent with the overall scene and timeline of the video. The goal is to maintain the natural flow of the video while making the specified changes to the scene elements.

Figure 7. The prompt template used for proposing a plausible manipulation.

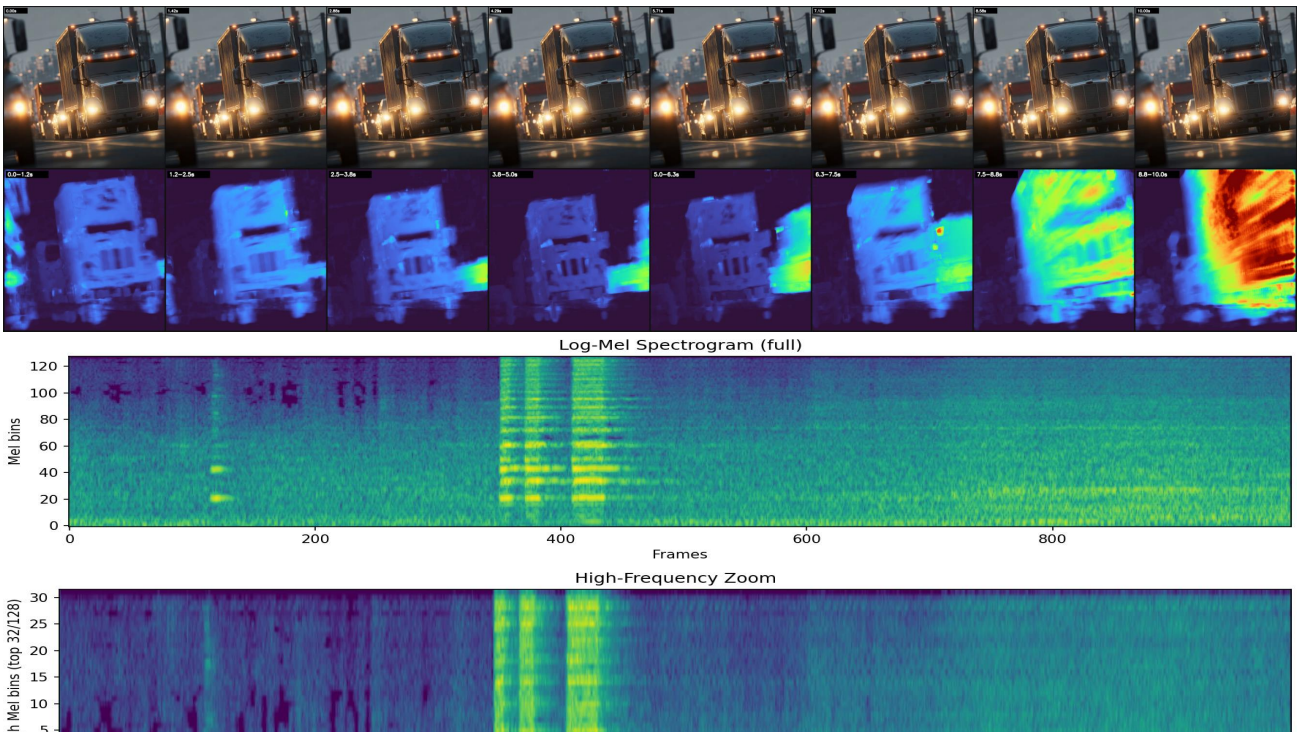


Figure 8. The prompt template used for proposing a plausible manipulation.

L4 Annotation Prompts for Human and General Subjects. We introduce the distinct prompts used to generate L4 (Explanatory Reasoning) annotations for Human Subject and General Subject. Given the inherent differences in

You are an audio-video authenticity analysis expert.
Your task is to analyze a {**forgery type**} audio-video clip and provide a clear, concise, and insightful explanation of its authenticity, focusing on the most critical and distinctive evidence rather than listing every minor detail.

You will be given 4 images, each representing a different modality of the clip:
Video frame collage (8 sampled frames in chronological order, with timestamps)
Motion heatmap collage (optical-flow-based motion intensity for the same 8 frames, showing facial/body dynamics)
Full audio spectrogram (Log-Mel Spectrogram across the entire duration)
High-frequency zoom (zoomed-in high-frequency region, highlighting sibilance/breath/noise details)

Please evaluate authenticity from two levels:
(1) Single-modality authenticity analysis
(2) Cross-modality consistency analysis

You may briefly reference key observations in each modality, but you do NOT need to explain every checklist item.
Focus on the most important features that strongly support or challenge authenticity.
If there are other key features not mentioned, feel free to include them.

I. Video Frame Authenticity (static visual)
(Briefly summarize realism: lighting, texture, expressions, absence of face warping, etc.)
Key angles to consider:
Lighting & color consistency; Natural textures and clean edges; Resolution and compression coherence; Realistic facial expressions and muscle movement;Stable perspective and natural depth-of-field

II. Motion Heatmap Authenticity (dynamic visual)
(Analyze whether motion patterns align with natural facial/body movement, speech rhythm, and background stability.)
Key angles:
Motion concentrated in mouth/jaw during speech; Blinking/head movement visible; Background static as expected; Motion intensity follows speech rhythm; No abrupt spikes or melting/block artifacts; Motion hotspots align with facial structure

III. Audio Spectrogram Authenticity (overall speech & environment)
(Highlight whether the spectral structure, environment sounds, and temporal flow look natural and human-like.)
Key angles:
Balanced low/mid/high frequencies; Energy rises and falls with speech rhythm; No repetitive/template-like spectral blocks; Background noise, breathing, room reverb present; Smooth temporal evolution, no abrupt cuts; Natural formant shifts, emotion, speech rate variation

IV. High-Frequency Zoom Authenticity (fine-grained audio detail)
(Comment on sibilance, breath noise, high-frequency variation, and absence of AI artifacts.)
Key angles:
Clear sibilance, breath, crackling; Energy decreases during pauses; No grid-like or repetitive high-frequency patterns; High-frequency details align with lower-frequency structure

V. Cross-Modality Consistency
(Determine if modalities reinforce each other. Point out strong coherence or mismatches.)
Key angles:
Lip movement timing matches audio speech rhythm; Speech pauses align with silent segments; Motion heatmap peaks match spectrogram energy peaks; Facial expressions align with vocal tone; Indoor scene matches indoor reverberation and background audio; Common AI inconsistencies to check: Video real but audio too clean/mechanical; Motion natural but face shows reconstruction artifacts; Audio real but lip/motion sync is off

Figure 9. The prompt template used for generating L4 annotations for Human Subject.

the types of forgeries that can occur in these two categories, the prompts are tailored to guide the model to focus on the most relevant features for each case.

As shown in Fig. 9, for Human Subject, the focus is on human actions, facial expressions, body language, and interactions with the environment. The prompts are structured to direct the model to detect and explain inconsistencies related to these human-centered aspects, such as unnatural gestures, mismatched lip-syncing, or unnatural facial expressions.

As shown in Fig. 10, for General Subject, the emphasis shifts to spatial and temporal changes in the scene. In these cases, the forgery may involve subtle changes to the environment, such as altered weather conditions, object placements, or unnatural motion. The prompts guide the model to focus on discrepancies in visual and motion patterns, as

You are an audio-video authenticity analysis expert.
Your task is to analyze a {**forgery type**} audio-video clip and provide a clear, concise, and insightful explanation of its authenticity, focusing on the most critical and distinctive evidence rather than listing every minor detail.

1 Global Directives (must follow)

G1 | Subject extraction

Begin by listing the top 1–3 subjects/events (e.g., wave & shoreline, two dogs running, road traffic, rotating beacon). Use these names consistently.

G2 | Evidence anchoring

Anchor every claim to timestamps (mm:ss) and visual regions (quadrant/area). For audio claims, provide the corresponding time. Use: Subject + time/region + phenomenon.

G3 | Partial anomalies can be decisive

Synthesis may surface in only a few key symptoms. Do not try to fill a checklist—focus on the strongest, most specific evidence.

G4 | Examples are patterns, not exhaustive

If new subjects appear (animals, factory siren, sports, weather), apply the same physical/semantic/temporal-consistency principles and name the subjects.

2 Three Analysis Pillars

I. Video: detect visual synthesis (within-frame + across-frames)

Select the most telling and locatable cues:
Texture & edges: over-smooth skin/fur/foam/foliage; tiling or repeated micro-patterns; halos, color bleed, doubled contours; “sticker-like” elements drifting relative to scene.
Lighting & shadows: light direction vs. shadow direction/length mismatch; contact shadows not hugging geometry; sky/ground brightness relation feels globally filtered.
Materials & reflections: water highlights not evolving with shape/view; metal/glass/wet surfaces reflecting the wrong content or staying “locked” as the view changes.
Geometry/perspective/depth: inconsistent convergence of building/road lines; far objects oddly sharp/saturated; depth cues fluctuate between frames without cause.
Temporal continuity: looped segments; pop-in/pop-out; melting/blocky patches through time; gait/vehicle path/crest advance that jumps phases.
Text/symbols (if present): inconsistent letter shapes across frames, broken strokes, implausible spelling.
Reflections/occlusion/interaction (if present): mirror/transparent surfaces with wrong parallax; contact/occlusion order that violates scene logic.

II. Audio: detect audio synthesis (listening; semantics & timing)

Focus on what you can hear, emphasizing semantics and temporal continuity:
Soundscape–visual mismatch: sounds with no visible cause (e.g., steady siren with no beacon, engine idle with no vehicle), or missing sounds for salient visible events.
Temporal discontinuities: audible cut points (sudden silence, hard jump), unnatural resets, repeating chunks, or template-like loops.
Event correspondence failures: breaking waves, impacts, landings, sharp turns without matching audible change—or audible bursts with no visual trigger.
Speech (if present): semantic or prosodic seams (spliced words/phrases), irregular pacing, emotional jumps, or misaligned mouth timing (heard vs. seen).

III. Cross-modal Consistency

Synthesis is revealed by aligning visible events with audible changes:
Event-sound synchronization: Visible events (waves breaking, animals taking off and landing, vehicles starting/turning, police lights flashing) should be accompanied by audible changes at the same timeframe; absence or misalignment is suspicious.
Motion-loudness relationship: Stronger visible motion is generally accompanied by more noticeable auditory changes; sequence and causality should be reasonable (e.g., post-action attenuation).
Environmental and material logic: The auditory perception of indoor/outdoor, open/narrow, hard/wet surfaces should align with the image space.
Camera movement and auditory perception: Zooming in/out or tracking/static shots correspond to the subjective auditory perception of approaching/removing; if they do not change with the image, it is suspicious.
Multi-subject layering: When waves, vehicles, and pedestrians coexist, the audio should exhibit multiple rhythms rather than a single template.

Figure 10. The prompt template used for generating L4 annotations for General Subject.

well as any audio-visual misalignments that could indicate a forgery. The goal is to encourage the model to explain the forgery in terms of spatial anomalies, such as object distortions, motion inconsistencies, or environmental artifacts.

Based on the explainability content, construct one MCQ that tests the model's ability to identify the most relevant forgery trace(s) for the declared subtype.

Question:
What looks inauthentic in this audio-video clip?

Options:
A. There is nothing inauthentic in this clip.
B. (Visual: subject/time/region + concrete symptom)
C. (Audio: subject/time + audible symptom)
D. (Cross-modal: event–sound alignment failure with time)
E. (Any axis, subject-anchored, time-anchored)

Rules:
Include exactly one correct option that captures the most salient anomaly in one modality (or a clear cross-modal mismatch pointing to it).

The other three must be plausible but wrong distractors.

After the options, explicitly state the correct answer(s), e.g.:
Correct Answer(s): C or Correct Answer(s): B and D.

Figure 11. The prompt template used for generating L3 annotations from L4 annotations.

Extracting L3 Annotations from L4. As shown in Fig. 11, we proceed to extract L3 (Forgery Detail Selection) annotations from L4. The goal of the L3 annotation is to distill the detailed explanation provided in the L4 annotation into a concise multiple-choice question that captures the most salient piece of evidence for the forgery.

Human Verification and Revision. After the LMM generates L3 and L4 annotations, these annotations undergo a thorough review by human annotators. The verification and revision process is designed to assess the plausibility, correctness, and clarity of the generated annotations. Specifically, annotators check for:

- **Correctness.** Annotators verify the accuracy of the forged details identified in the annotation. This includes verifying whether the detected forgery type is consistent with the given forgery type, and verifying whether the answer provided in L3 is correct.
- **Plausibility.** Ensure that the LMM's explanation of the forgery makes sense given the provided video and audio evidence. This includes confirming that the model's rationale is consistent with the actual content and that no logical inconsistencies are present.
- **Clarity.** Check that the generated explanation is clear, concise, and easily understandable. This includes ensuring that the LMM's rationale is well-structured and communicates the key forgery details effectively.

Fig. 12 shows an example of a manually revised annotation. The explanation originally generated by the LMM is rather vague, while the manually revised version provides a more precise description of the forged details.

B. Evaluation

B.1. Evaluation Metrics

For all objective questions (binary judgment, forgery type classification, and forgery detail selection), we report two metrics: *Accuracy* and *macro-F1*. Accuracy measures the overall proportion of correctly answered questions, while macro-F1 evaluates the balance of performance across different classes by giving each class equal weight, which is particularly important under class imbalance.

Normalized Bias Index (NBI): To assess whether the evaluated models exhibit unintended bias when performing *forgery-type classification*, we adopt the Normalized Bias Index (NBI) to quantify performance asymmetry across different forgery categories. Specifically, for each forgery type, we measure the model's recall on the *correct* option and on the *other* options, and compute the NBI as

$$\mathrm{NBI} = \frac{R_{\text{correct}} - R_{\text{other}}}{R_{\text{correct}} + R_{\text{other}}} \in [-1, 1], \tag{1}$$

where R_{correct} and R_{other} denote the recall rates for the correct option and other incorrect options. A positive and large NBI value indicates that the model tends to predict samples of that category as the correct option, while a negative and small NBI value indicates that the model tends to predict them incorrectly. By normalizing the recall difference, NBI provides a stable and comparable measure of type-specific prediction bias, enabling us to analyze whether the model systematically favors or suppresses particular forgery types during classification.

GPT-Score: For open-ended questions requiring explanatory analysis of audio–video forgeries, we employ

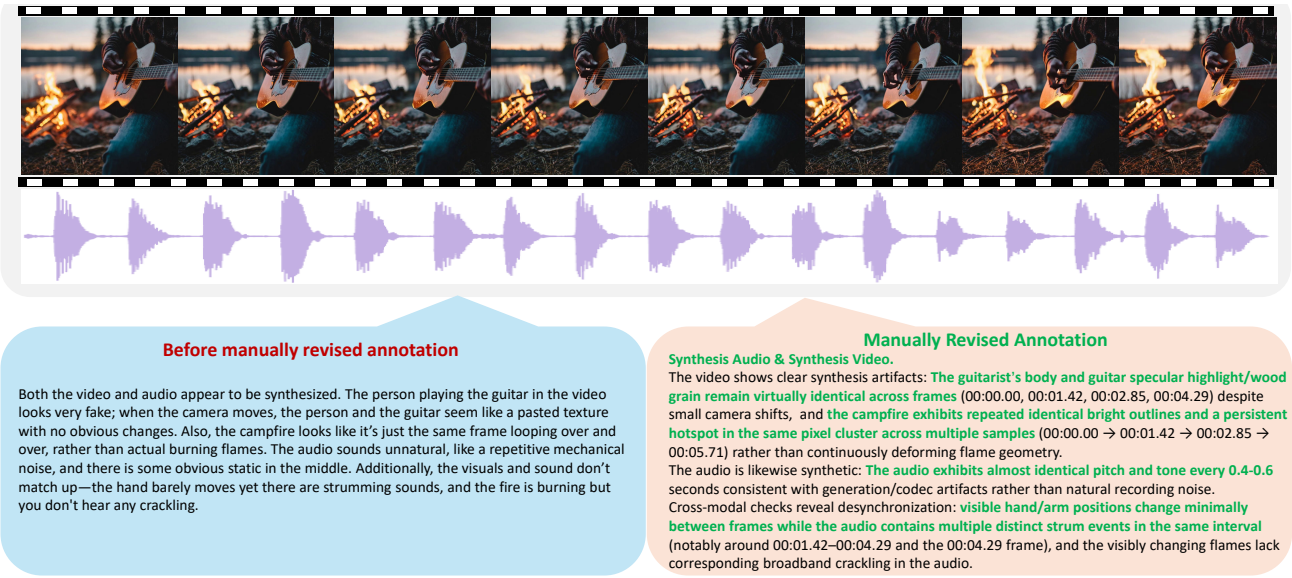


Figure 12. An example of human verification and revision.

a GPT-based evaluator to assess the quality of model-generated responses. The scoring process adheres to three criteria that together capture the correctness, relevance, and reliability of the model's explanation.

1) Classification Accuracy. The evaluator first determines whether the predicted forgery combination matches the ground-truth label. For binary classification, any "Edit" or "Synthesis" state in either audio or video modality is mapped to the "Fake" category, while samples with both modalities labeled as "Real" are mapped to "Real." For multi-class classification, partial correctness is considered, and the evaluator assigns a score based on the degree of alignment with the ground-truth forgery type.

2) Explainability Content Similarity. The evaluator compares the explanatory content in the model response against a human-crafted reference rationale, assessing how well the model captures the core evidence relevant to the forgery. This criterion rewards responses that accurately reproduce the essential manipulation details, the affected regions, and the observable artifacts described in the reference answer.

3) Reasonableness of Explanatory Content. Beyond similarity, the evaluator examines whether the explanation is *reasonable*—that is, whether the model avoids incorrectly treating authentic regions as forged while still providing logically coherent justifications for the detected anomalies. If the overall identified region and the reasoning are plausible, even when formulated differently from the reference rationale, the explanation can be credited as correct.

B.2. Main results in each scenario.

We compute all evaluation metrics for each scenario. Fig **??** reports representative results for all scenarios, and reveals a performance degradation in scenarios where human-related signals are weaker or absent, compared to human speech scenarios.

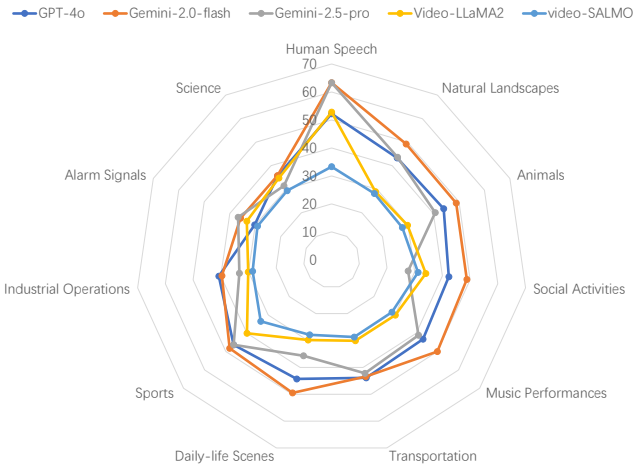


Figure 13. Performance of the evaluated AV-LMMs on Binary Authenticity Judgment.

B.3. Case Study

In this section, we present an analysis of AV-LLMs' behavior on audio-video forgery detection tasks, with an emphasis on the instances where the model produced incorrect answers. Examining these failure cases is essential for revealing the model's practical strengths and weaknesses. The insights gained from this analysis not only highlight current performance limitations but also inform future model development and training strategies.

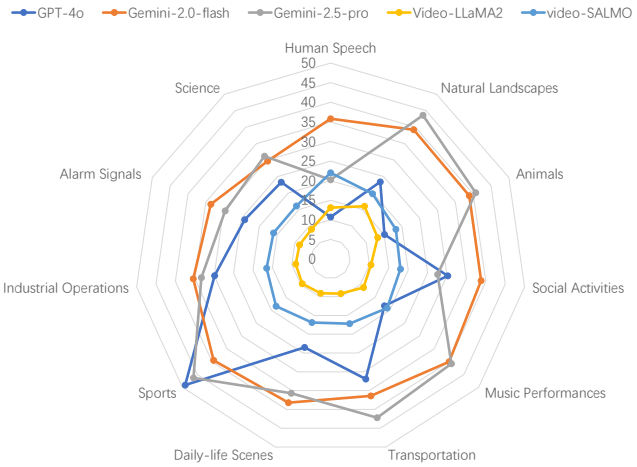


Figure 14. Performance of the evaluated AV-LMMs on Multiple-Choice Forgery Classification.

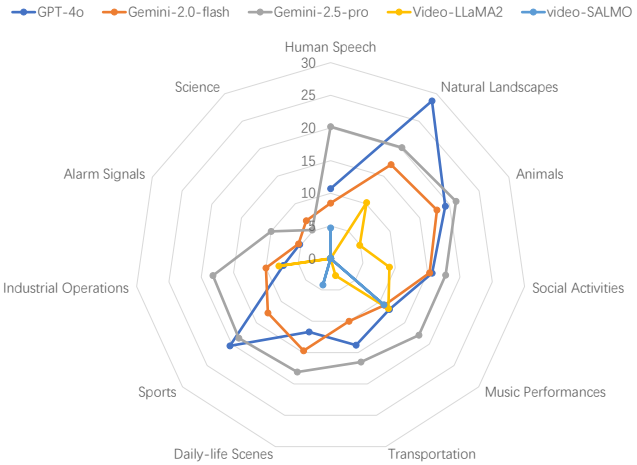


Figure 15. Performance of the evaluated AV-LMMs on Forgery Detail Selection.

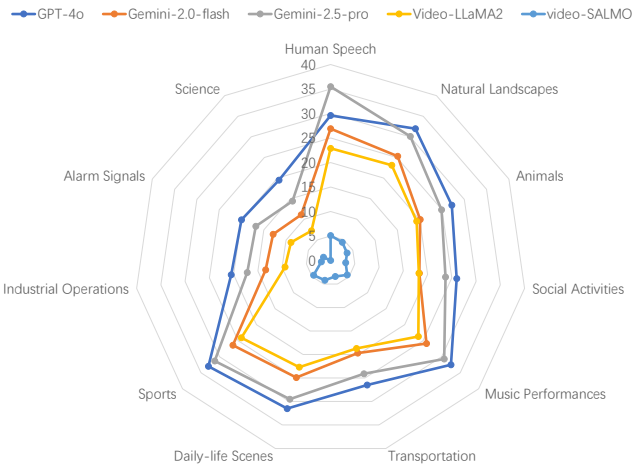


Figure 16. Performance of the evaluated AV-LMMs on Open-Ended Forgery Explanation.

Binary Judgment

Question:

You have been shown an audio-video clip.
Is this clip entirely captured from the real physical world without any AI involvement?

A. Yes
B. No

(Video-LLaMA2) Response: (A)

Ground Truth: (A)

Figure 17. An example of Binary Judgment.

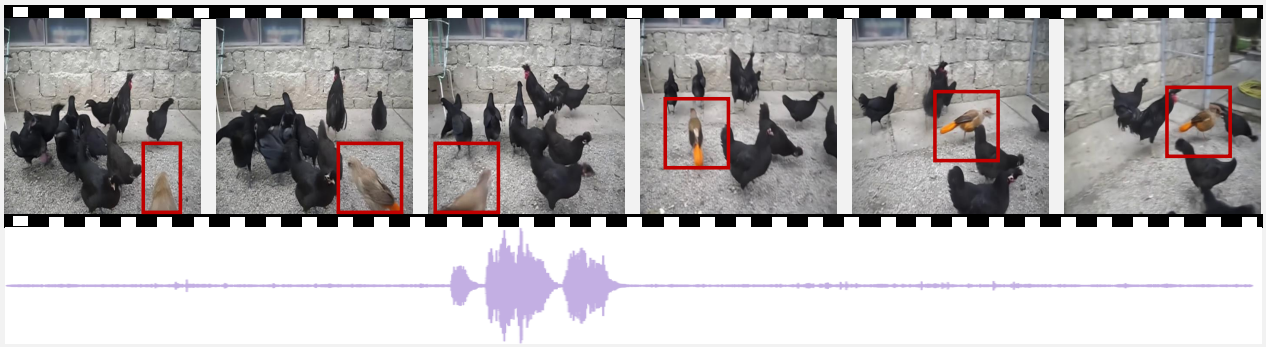
Forgery Types Classification

Question:

You have been shown an audio-video clip. This audio-video clip may have been captured from the real world, or it may have been generated or edited by an advanced AI model.

Which category does this audio-video belong?

A. Real Audio & Real Video
B. Real Audio & Edit Video
C. Real Audio & Synthesis Video
D. Edit Audio & Real Video
E. Edit Audio & Edit Video
F. Synthesis Audio & Real Video
G. Synthesis Audio & Synthesis Video



(video-SALMONN) Response: (C)

Ground Truth: (B)

Analysis:

The model detects signs of forgery in the video but **overlooks the fact that only the bird in the video exhibits unusual features**, such as a blurred head and high tail saturation, while **the rest of the video appears real**. Therefore, **it incorrectly identifies the edited video as a synthesized video**.

Figure 18. An example of Forgery Types Classification.

Forgery Detail Selection

Question:

You have been shown an audio-video clip. This audio-video clip may have been captured from the real world, or it may have been generated or edited by an advanced AI model.

Please determine which of the following options is correct.

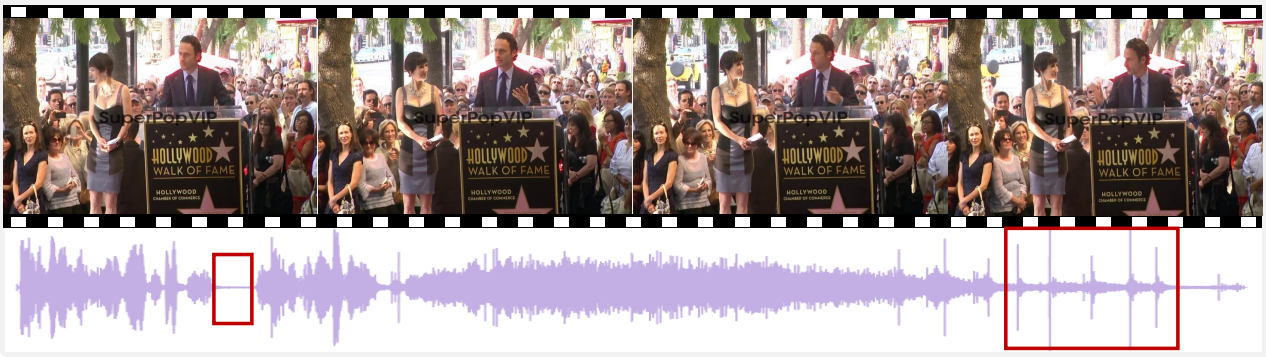
A. At 00:01.4–00:02.8 the speaker's mouth is closed while loud speech is audible.

B. The left background briefly contains a person who disappears then reappears one frame later.

C. The podium plaque contains duplicated lettering tiles.

D. Between 00:02.0 and 00:02.2, the audio suddenly stops, and between approximately 00:03.5 and 00:08.5, a highly stable, repetitive, sharp tone appears.

E. There is nothing inauthentic in this clip.



(GPT-4o) Response: (E)

Ground Truth: (D)

Analysis:

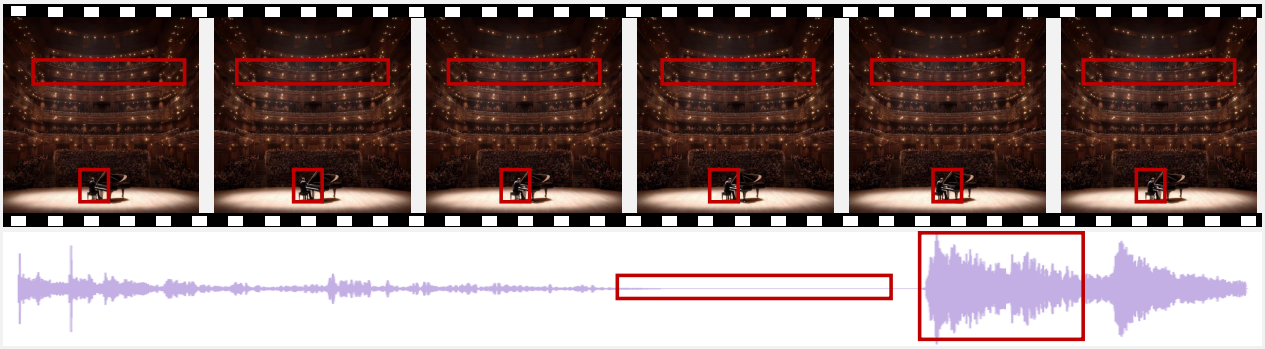
The model **lacks sufficient ability to perceive audio modalities** and is **completely unable to detect abnormal silences and repetitive pitches** in the audio.

Figure 19. An example of Forgery Detail Selection.

Explanatory Reasoning

Question:

You have been shown an audio-video clip. This audio-video clip may have been captured from the real world, or it may have been generated or edited by an advanced AI model.
Decide whether this clip is fully real, partially edited, or fully AI-generated. Then provide grounded forensic evidence about why.



(Gemini-2.5-Pro) Response: Real audio and real video.
Video: Lighting/shadows: **The lighting and shadows are consistent and realistic.** For instance, in the frame at 00:07, the shadows cast by the pianist's fingers on the piano keys accurately reflect the direction of the overhead stage lighting. Geometry/depth: The perspective and depth of the scene are coherent. **The receding lines of the stage and the arrangement of the audience in the background (visible at 00:01) adhere to proper geometric perspective.** Objects in the background are naturally less focused, indicating a realistic depth of field.
Audio: The sequence of sounds—applause for the performer, a pause, and then the start of the music—is semantically correct for a concert performance. **The sound of applause at the beginning directly corresponds to the visual of the audience clapping as the pianist walks on stage. The piano music begins shortly after the pianist is shown sitting at the keyboard with hands in position (00:03-00:05), which is a perfectly logical and natural timing.**

Ground Truth: Synthesized Audio & Synthesized Video.
Video: Individual balcony point-lights that **keep the same shapes/positions** while their **local contrast and edge definition jump abruptly between frames** (notably between ~00:05.71 and ~00:07.12), producing a stuttered pattern rather than natural lighting variation. **The pianist's head is too smooth, and his profile lacked normal facial texture.**
Audio: A strong broadband music/ambience is present from the start until about 00:05–00:06, **then a near-total silence occurs from ~00:05.5–00:07.5**, after which the signal resumes (~00:07.5–00:10)
Critically, **the pianist remains visually in performance posture at the piano across 00:05.71–00:07.12 while the soundtrack is silent.**

Analysis:

In the video modality, the model **fails to detect abrupt changes in background lighting** and **ignores issues such as overly smooth facial surfaces and a lack of detailed facial texture in the human**.
In the audio modality, the model **fails to detect unusually long pauses and sudden audio fluctuations**.

Figure 20. An example of Explanatory Reasoning.