Do Audio-Visual Large Language Models Really See and Hear?

Supplementary Material

Contents

Figure 8. **Factual Sample** A video where the visual content and audio content are highly correlated. The audio events in a factual sample can potentially be inferred by using visual cues

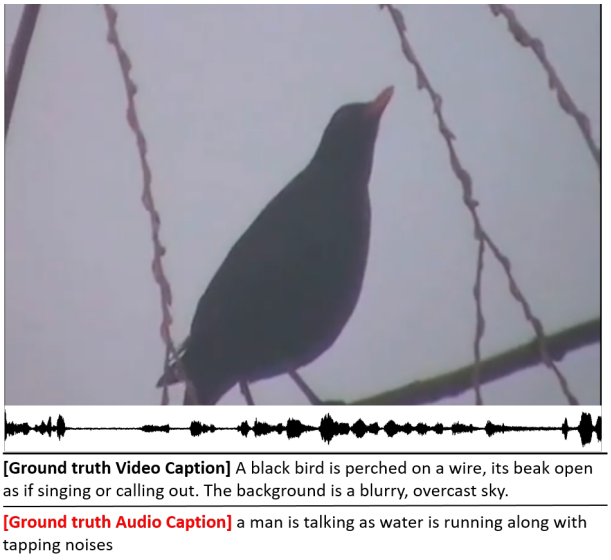


Figure 9. **Counterfactual Sample** A video where the visual content and audio content are in conflict. The audio events cannot be inferred from visual cues

A. Dataset

A.1. Data Source

For audio-visual captioning, while video captioning datasets are abundant, it is challenging to find datasets with human-annotated audio captions. Audio captioning is expensive and takes a lot of manual effort compared to visual captioning. We source our samples from AudioCaps [29], a standard benchmark for audio captioning derived from YouTube videos. Each clip is paired with human-written captions describing the sounds present, with 5 annotations per sample. AudioCaps has become a de facto standard for evaluating audio captioning alongside Clotho [15].

To obtain visual descriptions, we generate captions using GPT-4.1 [46]. We manually review and correct these generated captions where necessary. However, we find that videos in AudioCaps tend to feature relatively simple visual scenes, and GPT-4.1's outputs are generally accurate

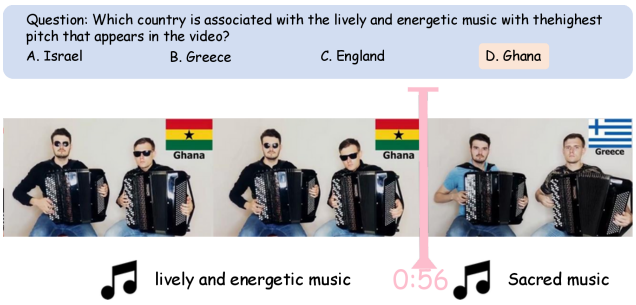


Figure 10. **World Sense Task** An example task from World Sense. These tasks couple perception and reasoning.

with minimal intervention required. Fig 8 depicts a sample with ground truth video and audio captions.

A.2. Counterfactual Sample Curation

To evaluate whether AVLLMs genuinely process audio independently of vision, we construct counterfactual samples where audio content conflicts with visible objects. While such cases occur naturally (*e.g.*, out-of-view sirens, background conversations), they are rare and difficult to scale for systematic evaluation.

We therefore create them synthetically. To construct a counterfactual sample, we take a video and pair it with an audio track that cannot plausibly be inferred from the visible objects. Fig 9 depicts a counterfactual sample with ground truth video and audio captions. We use audio and video captions as a proxy for semantic content: by finding audio-video pairs with dissimilar captions, we ensure their soundscapes are likely incompatible with the visual scene. Specifically, we embed all audio captions and GPT-4 generated video captions using the Qwen3-Embedding-8B [64] model. For audio, we compute embeddings for all 5 ground-truth captions per sample and average them. We then compute the cosine similarity matrix between all audio–video caption pairs and apply the Hungarian matching algorithm to find one-to-one assignments that minimize similarity. This ensures paired audio and video samples are semantically dissimilar.

Finally, we use FFmpeg to swap the original audio track of each video with its matched dissimilar audio. The complete procedure is detailed in Algorithm 1. From all generated pairs, we select 250 samples with lowest cosine similarity (≈ 0.498), ensuring strong counterfactual mismatches. This yields 250 factual samples (original audio-video pairs) and 250 counterfactual samples (mismatched pairs) for evaluation.

A.3. Existing Benchmarks

Existing audio-visual benchmarks are insufficient for our analysis. Benchmarks such as World Sense [23] (example in Fig 10) couple perception with reasoning, requiring AVLLMs to first perceive audio-visual events and then apply world knowledge to answer questions. Since we aim to isolate perceptual capabilities and identify modality biases, such reasoning-dependent tasks conflate multiple AVLLM capabilities.

Other benchmarks like AVHBench [52] (example in Fig 11) focus on perception through targeted questions about specific audio or visual events, while the corresponding video or audio cues attempt to mislead the AVLLM and induce hallucinations. However, we observe that these misleading cues are not sufficiently adversarial. For instance, in AVHBench's video-induced audio hallucination category, models achieve 75.6% accuracy with both modalities present. Removing the video component yields 73.0% accuracy, indicating that the visual modality fails to mislead the model. Such tasks do not create sufficient modality con-

Algorithm 1 Counterfactual Dataset Construction

Require: Videos $V = \{v_i\}$, Audios $A = \{a_i\}$, Captions $C = \{c_{ij}\}$, Encoder E
Ensure: Counterfactual pairs S

Step 1: Compute Semantic Embeddings
for $i \leftarrow 1$ to N **do**
  $e_{a_i} \leftarrow \frac{1}{5}\sum_{j=1}^{5} E(c_{ij})$ ▷ Centroid of 5 audio captions
  $e_{v_i} \leftarrow E(\text{GPT-4.1}(v_i))$ ▷ Vision-only caption embedding
end for
Step 2: Global Assignment
$M \in \mathbb{R}^{N\times N} \leftarrow$ CosineSimilarity matrix between $\{e_a\}$ and $\{e_v\}$
$\pi \leftarrow \text{HungarianMatching}(M)$ ▷ Optimal 1-to-1 mapping indices
Step 3: Filter and Construct
$S_{candidates} \leftarrow \{(A[\pi(i)], V[i]) \mid i \in 1..N\}$
Sort $S_{candidates}$ by similarity score $M_{i,\pi(i)}$
$S \leftarrow$ Select 250 pairs from $S_{candidates}$ with score ≈ 0.498
return Synthesize videos for S (swap audio tracks)

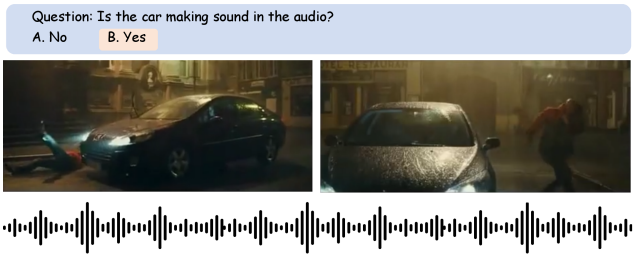


Figure 11. **AVHBench Task** An example task from AVHBench. These tasks evaluate perception capabilities.

flict to stress-test whether models genuinely process and integrate both modalities independently or exhibit bias toward one modality over the other.

B. Evaluation

B.1. Human Evaluation Study

To validate our LLM-as-a-judge approach, we conduct a human evaluation study on 200 stratified samples (100 factual, 100 counterfactual). Two graduate students familiar with audio-visual content independently rate each generated caption on a 0-1 scale for audio and visual fidelity separately. Annotators are briefly introduced to the same rubric used by the LLM judge and given the opportunity to clarify any ambiguities. Following the LLM judge setup, annotators evaluate only the text captions (generated description, ground-truth audio captions, and ground-truth video captions) without access to the actual videos, ensuring scalability and consistency with the automated evaluation.

We compute the correlation between human ratings and LLM judge scores using Spearman's ρ, obtaining $\rho = 0.816$ for audio caption fidelity and $\rho = 0.732$ for video caption fidelity, demonstrating strong alignment between the LLM judge and human judgment.

B.2. LLM Judge Prompts

We design prompts with in-context examples to calibrate LLM-as-judge rating with human judgement. Fig 13 details the prompt used to measure audio caption fidelity and Fig 14 details the prompt used to measure video caption fidelity.

C. Qualitative Analysis

In this section, we include additional qualitative results (Fig 15 and Fig 16 of the existing AVLLMs in different scenarios, such as counterfactual samples and under the effect of attention knockouts.

D. Additional Results

To demonstrate the generalizability of our findings beyond Qwen-Omni [56] series of AVLLMs, we extend our analysis to other representative AVLLMs, specifically MiniCPM-o2.6 [24], VideoLLaMA 2.1 [63], and InternOmni [1].

D.1. Probing Audio Representations

We probe the intermediate audio representations of MiniCPM-o2.6 and VideoLLaMA 2.1 to verify if the "competence gap", where latent audio information exists but fails to manifest in generation is generalizable to more AVLLMs. Consistent with our main results, we observe a significant difference between latent capabilities and generated output. For MiniCPM-o2.6, in counterfactual samples we measure a latent audio recall of **75.4**% compared to a generated caption fidelity of **22.1**%. Similarly, VideoLLaMA 2.1 achieves a latent recall of **59.9**% against a generated fidelity of **34.1**%.

Qualitatively, we observe that the decoded audio tokens in these models accurately describe sound events (e.g., "siren", "barking"). However, unlike Qwen2.5-Omni, we do not observe multilingual token representations in the intermediate layers of MiniCPM-o2.6 or VideoLLaMA 2.1. This suggests that the multilingual audio representations observed in Qwen are likely specific to the model's training data.

D.2. Investigating Information Flow

We replicate the attention knockout experiments on MiniCPM-o2.6 and VideoLLaMA 2.1 to trace the cross-modal information flow. The results are visualized in Figure 12b and Figure 12a. We observe integration patterns very similar to those reported in the main paper: both audio and visual information are processed primarily in the deeper transformer layers. Crucially, we confirm the phenomenon of visual interference: blocking the visual pathways ($G \nrightarrow V$) in these deep layers results in a recovery of audio understanding performance, further validating that visual representations actively interfere with audio cues during the final stages of generation.

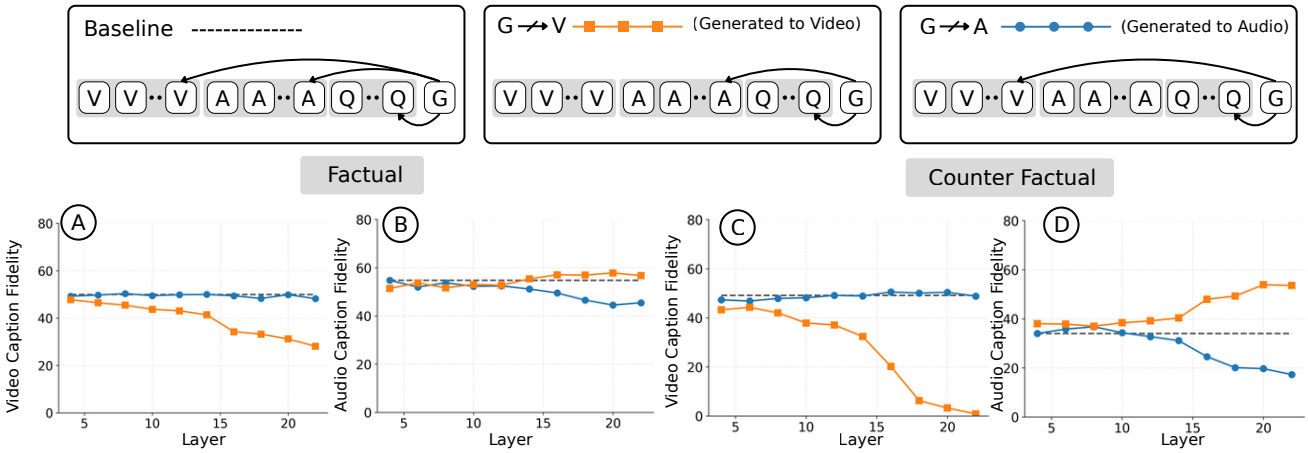
VideoLlama 2.1: Figure 12a shows the results of attention knockout experiments for VideoLlama 2.1. In video understanding for factual samples, we observe that blocking audio has no observable impact. However, blocking video does have minor impact, which is largely recovered by compensating using audio cues. We see a similar trend with audio understanding in factual samples. In video understanding for counterfactual samples, we observe almost complete loss of video understanding upon blocking video. For audio understanding in counter-factual, blocking video in fact drastically improves audio understanding.

MiniCPM-o2.6: Figure 12b shows the results of attention knockout experiments for MiniCPM-o2.6. In video understanding for factual samples, we observe a similar pattern as before. Interestingly, even in factual samples, blocking video leads to improvement in audio understanding performance. Surprisingly, the drop in video understanding performance in counterfactual suggests that cross-modal transfer might not be restricted in deeper layers and the window for this transfer stretches beyond the window size of 9 that we use for knockouts. We observe a similar pattern in audio understanding for counterfactual scenarios, where audio understanding does improve, but not as drastically as in previous models.

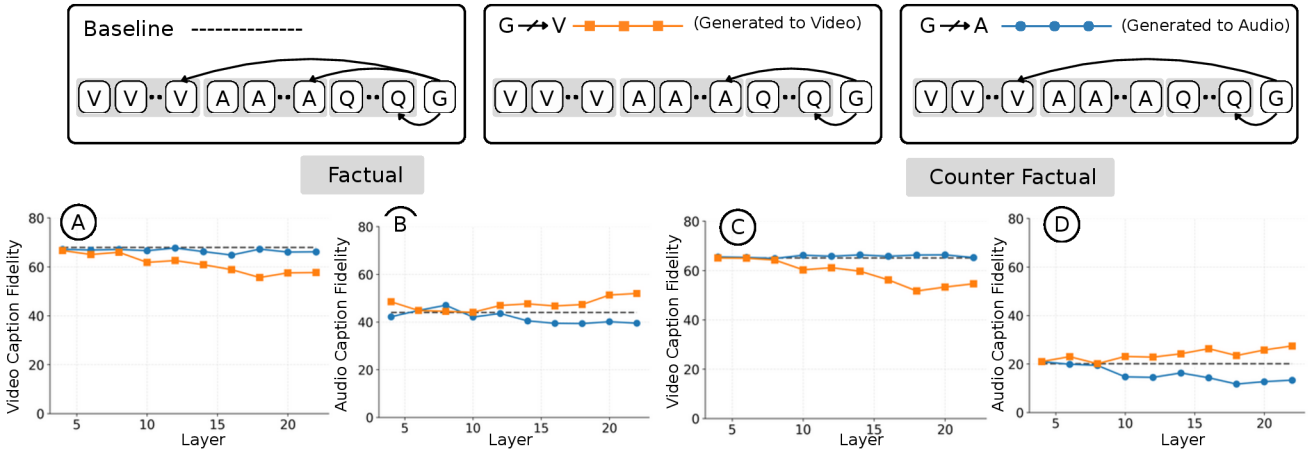
D.3. Investigating Origins of Visual Bias

Finally, we investigate the origin of this visual bias by analyzing InternOmni and its base model, InternVL [13]. InternOmni is particularly relevant for this analysis as it is explicitly initialized from the InternVL checkpoint. We compare the output token distributions of InternOmni (given audio-visual input) against InternVL (given vision-only input) using the metrics defined in Section 8.

We observe a low KL divergence of **0.46** between the distributions. Furthermore, regarding audio-related tokens, we find that **70.62**% are unshifted ($\eta = 1$) and only **10.07**% are shifted ($\eta > 3$). This strong alignment indicates that InternOmni's generation remains dominated by the priors of its vision-language base model. These findings reinforce our conclusion that the modality imbalance in AVLLMs could potentially be an inherited trait from the initialization and alignment phases.



(a) **VideoLlama 2.1:** Blocking audio-visual information flow, for both factual and counterfactual samples



(b) **MiniCPM-o2.6:** Blocking audio-visual information flow, for both factual and counterfactual samples

Figure 12. **Comparison of Information Flow Blocking.** Blocking audio-visual information flow, for both factual and counterfactual samples in VideoLlama 2.1 (top) versus MiniCPM-o2.6 (bottom).

Audio Caption Fidelity Prompt

Evaluate how accurately the candidate caption describes the audio compared to the reference caption.
The caption is an audio-visual caption and may include visual descriptions. This is fine and should be ignored
Specifically, you should neither reward nor penalize candidate caption for any visual description and focus only on audio-related content
On the other hand, you should penalize model if it hallucinates some audio events

Candidate Caption: I see a grassy area with a few pigeons. There are some rocks and a wooden fence in the background. I hear some birds chirping and a bit of rustling. There's also a sound of a drill in the background, and then a sigh. What do you think about this scene?
Reference Caption: drilling with music playing and a man speaks
score: 0.25
reason: The candidate correctly identifies the drilling sound but misses two major audio events (music playing and man speaking). More critically, it hallucinates three audio events that don't exist in the reference (birds chirping, rustling, and a sigh), suggesting significant audio misperception or confabulation.

Candidate Caption: Well, in the video, I see a baby. The baby is being held by someone. The baby looks a bit chubby and has a small head. The baby's skin is smooth and soft. The baby's eyes are open, and it seems to be looking at the person holding it. The baby's mouth is open, and it looks like it might be making a sound or crying. Oh, and there's a bit of a smile on the person's face who's holding the baby. There's also some background noise, but it's not really clear what it is. It could be some kind of ambient sound in the room. So, what do you think about the video? Do you have any other questions?
Reference Caption: a man laughs followed by a baby crying
score: 0.5
reason: The candidate correctly identifies baby sounds (mentions baby "making a sound or crying"), capturing one of the two main audio events. However, it completely misses the man laughing, which is a distinct and salient audio event in the reference. The vague mention of "background noise" and "ambient sound" shows audio awareness but lacks the specificity of identifying laughter.

Candidate Caption: Well, in the video, you can see the inside of a car. There's a red dashboard with some stickers on it. You can also see the steering wheel and a few other bits of the car's interior. As for what I hear, there's a constant hum of the engine, which is probably from the car being in motion. And there's also the sound of the wind outside, like it's a nice day. So, what do you think about the video? Do you like the sound of the car engine?
Reference Caption: a vehicle engine running smoothly
score: 0.80
reason: The candidate accurately identifies the primary audio event - the vehicle engine running (described as "constant hum of the engine"). However, it hallucinates an additional audio event by mentioning "the sound of the wind outside," which is not present in the reference caption. The core audio description is correct, but the added wind sound warrants a moderate penalty.

Candidate Caption: Well, in the video, you can see a close - up of a car engine. There are all kinds of parts like the air filter, the belts, and the hoses. The engine looks pretty complex with all these different components. And, um, there's a sound of the engine starting up. It's a bit of a loud, rumbling noise. What do you think about the engine? Do you have any questions about it?
Reference Caption: an engine revs repeatedly
score: 0.70
reason: The candidate correctly identifies the engine sound as the primary audio event with appropriate descriptors ("loud, rumbling noise"). While it describes this as "starting up" rather than "revs repeatedly," it still captures the essential audio content - a loud engine sound. The missing detail about the repetitive revving pattern is a relatively minor omission compared to getting the core audio event correct.

Candidate Caption: <CANDIDATE_CAPTION>
Reference Caption: <REFERENCE_CAPTION>

Rate the accuracy precisely in 0-1 scale.
You need to strictly return your output as a json with the following keys
- **score:** A precise value between 0 and 1
- **reason:** Reasoning for why you provided that score

Figure 13. The prompt used to measure audio caption fidelity.

Video Caption Fidelity Prompt

Evaluate how accurately the candidate caption describes the video compared to the reference caption.
The caption is an audio-visual caption and may include audio descriptions. This is fine and should be ignored
Specifically, you should neither reward nor penalize candidate caption for any audio description and focus only on vision-related content
On the other hand, you should penalize model if it hallucinates some visual events
Lastly, we are not expecting very detailed caption, so even an accurate but high-level description of key details is acceptable.

Candidate Caption: Well, in the video, I see a bunch of pigeons. There's one pigeon walking on the grass first. Then, there are more pigeons in the background, some are flying around. The grass is green and a bit dry in some places. There are also some trees and a building in the background. Oh, and there's a bird cooing in the background too. It's a pretty peaceful scene. So, what do you think about it? Do you have any other questions about the video?
Reference Caption: A pigeon is walking on a grassy field. It suddenly takes off and flies into the air. The camera follows the pigeon as it soars through the sky.
score: 0.9
reason: The candidate accurately captures the high-level visual scene (multiple pigeons, grass, outdoor space with trees/building, some flying). The "pigeon walking on the grass" likely refers to the spinning/rolling behavior. Minor details like snow patches and pigeon color variation are omitted but don't significantly impact accuracy. Audio description (cooing) is appropriately ignored per instructions.

Candidate Caption: Well, in the video, I see a baby. The baby is being held by someone. The baby looks a bit chubby and has a small head. The baby's skin is smooth and soft. The baby's eyes are open, and it seems to be looking at the person holding it. The baby's mouth is open, and it looks like it might be making a sound or crying. Oh, and there's a bit of a smile on the person's face who's holding the baby. There's also some background noise, but it's not really clear what it is. It could be some kind of ambient sound in the room. So, what do you think about the video? Do you have any other questions?
Reference Caption: A baby is being held upright and their initial expression appears neutral or happy while looking at an adult facing them. Over the sequence, the baby's facial expression gradually shifts to one of distress, and by the end, the baby looks like they are crying. The adults face is visible near the baby throughout, possibly interacting with them.
score: 0.8
reason: The candidate accurately captures the main visual elements: baby being held, looking at the person holding them, mouth open (suggesting crying), and the adult's visible face/expression. While it doesn't explicitly describe the temporal progression from neutral to distressed, it does capture the crying state which is the endpoint. The core visual content is largely correct.

Candidate Caption: Well, in the video, I see a man in a black shirt and beige pants walking through a room. There's a computer on a desk, a clock on the wall, and a bulletin board with some papers on it. The room has a checkered floor. As for the sound, there's a baby crying in the background. It's a bit loud and really makes you feel the atmosphere of the room. So, what do you think about the video? Do you have any other questions?
Reference Caption: A man enters an office-like room, walks towards the camera, then suddenly appears to get startled or react in surprise near the foreground. He lifts his arms above his head and then quickly turns and looks around, suggesting he may have seen or experienced something unexpected in the room.
score: 0.6
reason: The candidate accurately describes the man, his clothing, and the room setting (desk, computer, clock, bulletin board, checkered floor), capturing the environmental context well. It identifies him walking through the room. However, it misses the startled reaction, arm-raising, and surprised behavior - the key dramatic action. Still, it gets roughly half the visual content (person, setting, basic movement) correct.

Candidate Caption: In the image, I see a person standing in a room with a white wall and a window in the background. The person is wearing a white shirt and appears to be looking at something off to the side. The room is well-lit, and there is a piece of furniture, possibly a desk or a table, in the foreground. The overall atmosphere of the room is calm and quiet.
Reference Caption: A man enters an office-like room, walks towards the camera, then suddenly appears to get startled or react in surprise near the foreground. He lifts his arms above his head and then quickly turns and looks around, suggesting he may have seen or experienced something unexpected in the room.
score: 0.0
reason: The candidate describes a static, calm scene (person standing, looking to side, well-lit room) while the reference describes dynamic action (man entering, walking, getting startled, lifting arms, turning). The candidate completely misses all the key motion and dramatic reaction events. It captures only the most basic element (a person in a room) but fails to convey any of the actual visual narrative or action.

Candidate Caption: <CANDIDATE_CAPTION>
Reference Caption: <REFERENCE_CAPTION>

Rate the accuracy precisely in 0-1 scale.
You need to strictly return your output as a json with the following keys
- **score:** A precise value between 0 and 1
- **reason:** Reasoning for why you provided that score

Figure 14. The prompt used to measure video caption fidelity.

A) Factual with out-of-view audio

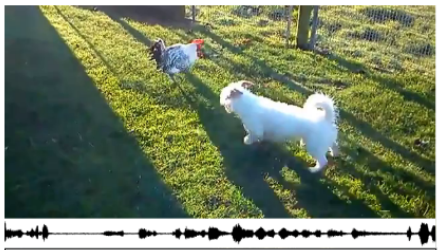
[Ground truth video caption] A small white dog and a chicken are running around in a grassy yard. The dog is chasing the chicken, and they both run back and forth across the yard

[Ground truth audio caption] A rooster cackles and a dog yelps and whimpers followed by a man laughing and then speaking then the dog barks

[Describe what you see and hear] The video shows a dog and a chicken running in the yard. The dog is chasing the chicken, barking and growling. The chicken is flapping its wings and making clucking sounds. The sun is shining brightly, casting a warm glow over the scene.

[Judge Response (Audio Fidelity)] Sounds related to the dog and chicken are correctly captured – barking / growling, clucking. However the man laughing is not captured.
[Audio Fidelity Score:] 0.8

B) Counterfactual



[Ground truth video caption] A small white dog and a chicken are running around in a grassy yard. The dog is chasing the chicken, and they both run back and forth across the yard

[Ground truth audio caption] Typing on a keyboard is ongoing in a quiet environment

[Describe what you see and hear] The video shows a dog and a chicken running in the grass. The dog is chasing the chicken, and the chicken is trying to escape from the dog. The main source of sound in the video is barking of the dog

[Judge Response (Audio Fidelity)] None of the actual sounds are captured
[Audio Fidelity Score:] 0.0

C) Counterfactual with video knockout

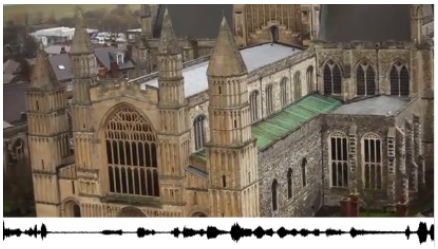
[Ground truth video caption] A small white dog and a chicken are running around in a grassy yard. The dog is chasing the chicken, and they both run back and forth across the yard

[Ground truth audio caption] Typing on a keyboard is ongoing in a quiet environment

Describe what you hear] The video shows a person typing on the keyboard. The main audio sound is the keyboard typing

[Judge Response (Audio Fidelity)] Sound of keyboard typing is captured correctly
[Audio Fidelity Score:] 1.0

A) Factual



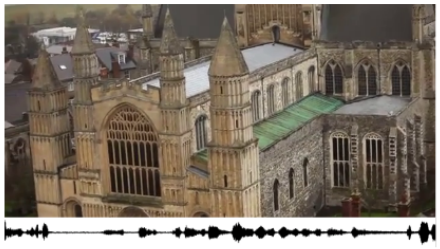
[Ground truth video caption] The video shows a series of aerial views of a large, historic building with Gothic architectural features. The building has multiple towers and spires, with intricate stone carvings and pointed arches. The camera pans around the structure, revealing its grand scale and detailed design. The surrounding area includes other buildings and greenery, suggesting an urban setting. The overall atmosphere is one of historical significance and architectural beauty.

[Ground truth audio caption] church bells ringing as rain pours and wind blows into a microphone

[Describe what you see and hear] The video shows an aerial view of the cathedral of of London, England. The large stone building with many windows and towers is the main focus of the video. The sound of change ringing fills the air, creating a peaceful atmosphere.

[Judge Response (Video Fidelity)] Visual elements like historic building (cathedral), towers, windows stone carvings are correctly captured in the video, although a few minor elements are hallucinated but have similar meaning
[Video Fidelity Score:] 0.8

B) Counterfactual



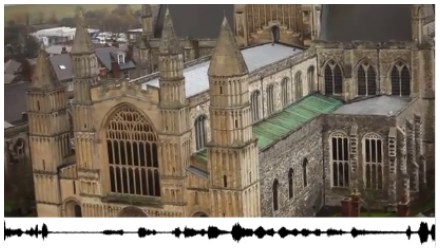
[Ground truth video caption] The video shows a series of aerial views of a large, historic building with Gothic architectural features. The building has multiple towers and spires, with intricate stone carvings and pointed arches. The camera pans around the structure, revealing its grand scale and detailed design. The surrounding area includes other buildings and greenery, suggesting an urban setting. The overall atmosphere is one of historical significance and architectural beauty.

[Ground truth audio caption] a man speaks then a second man speaks followed by a woman speaking

[Describe what you see] The video shows an aerial view of of the Palace of Westminster, also known as the Houses of Parliament, located in London, England. The building is a large, Gothic-style structure with multiple towers and spires, and a green roof. The video also shows a large stone wall surrounding the building.

[Judge Response (Video Fidelity)] Almost all visual elements are captured correctly, although some minor elements in the background are missed
[Video Fidelity Score:] 0.85

C) Counterfactual with audio knockout



[Ground truth video caption] The video shows a series of aerial views of a large, historic building with Gothic architectural features. The building has multiple towers and spires, with intricate stone carvings and pointed arches. The camera pans around the structure, revealing its grand scale and detailed design. The surrounding area includes other buildings and greenery, suggesting an urban setting. The overall atmosphere is one of historical significance and architectural beauty.

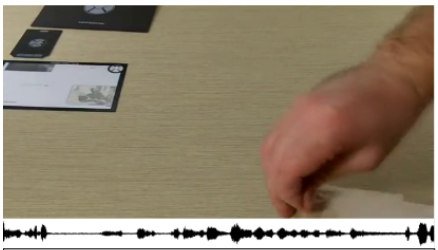
[Ground truth audio caption] a man speaks then a second man speaks followed by a woman speaking

[Describe what you see] The video shows an aerial view of the Palace of Westminster, also known as the Houses of Parliament, in London, England. The large Gothic style building with multiple towers and spires can be seen from above. The video also shows a large cathedral with a similar architectural style.

[Judge Response (Video Fidelity)] Almost all elements are captured correctly with minor misses of visuals in the background and minor hallucinations
[Video Fidelity Score:] 0.88

Figure 15. **Qualitative examples**. We illustrate some outputs generated by VideoLlama along with its LLM-judge score and reasoning, for factual and counterfactal samples and under attention knockout. The examples in the first row capture audio fidelity and the ones below capture video fidelity

A) Factual with out-of-view audio



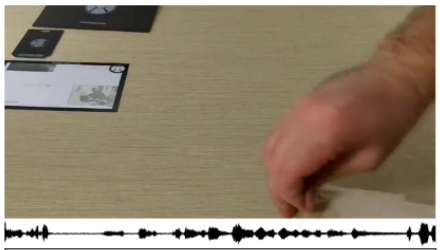
[Ground truth video caption] A hand reaches into the frame and picks up a small, rectangular object from a table. The object is wrapped in plastic and appears to be a small electronic device or gadget. The hand then lifts the object up to show it to the camera before placing it back on the table.

[Ground truth audio caption] a man is speaking as paper is crumpling

[Describe what you hear] The sound of crinkling plastic wrap is heard, suggesting that the person might be opening a package or handling some kind of material.

[Judge Response (Audio Fidelity)] Sounds of paper crinkling are correctly identified. However, the man speaking is completely missed which is the main source.
[Audio Fidelity Score:] 0.5

B) Factual with video knockout (OOV audio)



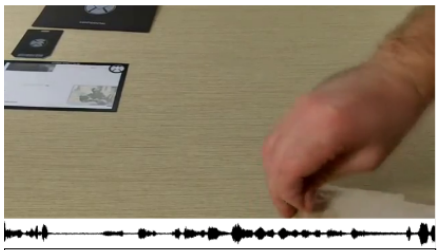
[Ground truth video caption] A hand reaches into the frame and picks up a small, rectangular object from a table. The object is wrapped in plastic and appears to be a small electronic device or gadget. The hand then lifts the object up to show it to the camera before placing it back on the table.

[Ground truth audio caption] a man is speaking as paper is crumpling

[Describe what you hear] The sound of the crumpling and tearing paper is accompanied by a male voice speaking in an excited manner.

[Judge Response (Audio Fidelity)] Almost all the audio elements are captured correctly - paper crumpling and man speaking
[Audio Fidelity Score:] 1.0

C) Counterfactual with video knockout



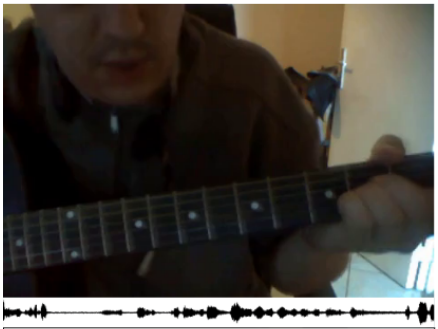
[Ground truth video caption] A hand reaches into the frame and picks up a small, rectangular object from a table. The object is wrapped in plastic and appears to be a small electronic device or gadget. The hand then lifts the object up to show it to the camera before placing it back on the table.

[Ground truth audio caption] Horses neighing then snorting and trotting on a dirt surface

Describe what you hear] The audio contains a sequence of distinct sounds. Initially, there is a sound that resembles the snapping or cracking of of wood, followed by a series of deep growls and grunts.

[Judge Response (Audio Fidelity)] The sound of horse neighing (grunting) is captured but trotting is missed. Also it hallucinated wood breaking sound.
[Audio Fidelity Score:] 0.5

A) Factual



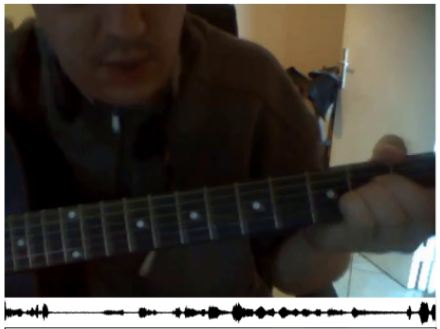
[Ground truth video caption] A person is playing an acoustic guitar. The camera focuses on the person's hand moving along the neck of the guitar, pressing down on the strings to form chords. The background is slightly blurred but appears to be an indoor setting with a door and some furniture visible.

[Ground truth audio caption] a man talking then whistling as guitar music plays

[Describe what you see and hear] The video features a person in an indoor setting, focusing on their playing a stringed musical. The're wearing a dark jacket and holding the guitar with one hand while using strumming or picking it with the other. Throughout the sequence of frames, there's minimal movement from the person, emphasizing their engagement with the instrument. In the background, you can see some household items like shoes hanging on hooks on the wall, adding to the domestic atmosphere.

[Judge Response (Video Fidelity)] identifies the core visual elements: a person playing a guitar (stringed instrument), using their hands for strumming/picking, and the domestic indoor setting.
[Video Fidelity Score:] 0.7

B) Counterfactual



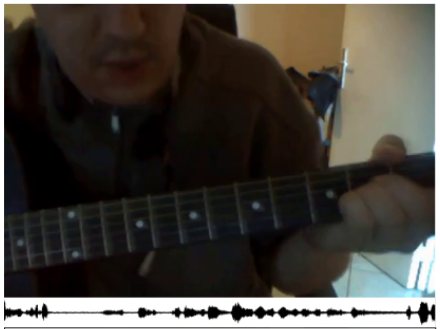
[Ground truth video caption] A person is playing an acoustic guitar. The camera focuses on the person's hand moving along the neck of the guitar, pressing down on the strings to form chords. The background is slightly blurred but appears to be an indoor setting with a door and some furniture visible.

[Ground truth audio caption] glass doors rattling and sliding shut followed by a plastic clack

[Describe what you see] The video features a person playing an acoustic guitar in a room. The the individual is dressed in a dark jacket with a zipper and has headphones around their neck, indicating they might be focused on the music or recording session. Throughout the sequence, there are slight movements of fingers along the fretboard, suggesting the performance of various chords or notes.The setting includes a door wall, some hanging clothes in the background, and tiled flooring visible at the bottom right corner

[Judge Response (Video Fidelity)] Almost all visual elements are captured correctly, although some minor elements in the background are missed
[Video Fidelity Score:] 0.8

C) Counterfactual with audio knockout



[Ground truth video caption] A person is playing an acoustic guitar. The camera focuses on the person's hand moving along the neck of the guitar, pressing down on the strings to form chords. The background is slightly blurred but appears to be an indoor setting with a door and some furniture visible.

[Ground truth audio caption] glass doors rattling and sliding shut followed by a plastic clack

[Describe what you see] The video features a person in an indoor setting, wearing a dark-colored jacket with a zipper. The focus is on the guitar's neck and fretboard, which has white dot in markers along its length. Throughout the sequence, the individual adjusts their fingers on the fretboard, suggesting they be playing or demonstrating something on the guitar. In the background, there are some hanging items, possibly jackets or bags, against a wall, adding to the indoor ambiance of the scene.

[Judge Response (Video Fidelity)] captures the core visual elements: a person interacting with a guitar, focusing on the neck/fretboard, and adjusting fingers to suggest playing/demonstration, indoor setting. Some background details and the face visibility is missing
[Video Fidelity Score:] 0.75

Figure 16. **Qualitative examples**. We illustrate some outputs generated by MiniCPM-2.6-o along with its LLM-judge score and reasoning, for factual and counterfactal samples and under attention knockout. The examples in the first row capture audio fidelity and the ones below capture video fidelity