Do State-of-the-art Audio-visual VLMs Understand Audio-video Temporal Misalignment

Motonobu Kimura [1,2], Ren Ohkubo[1,2], Yue Qiu[1], Yutaka Satoh[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST), [2]University of Tsukuba
{kimura.0726, ookubo.0720, qiu.yue, yu.satou}@aist.go.jp,

Abstract

Audio-visual vision–language models (VLMs) have recently leapt forward, excelling at recognizing and localizing audio-visual events from videos, generating videos with audio from text prompt. Yet whether these models truly understand the temporal synchrony between what is seen and what is heard remains unanswered. Existing systems (i) mostly sparsely sample video frames, making accurate alignment challenging, or (ii) inherit M-RoPE/TM-RoPE encodings that are only reliable within a two-second window; all are trained and evaluated exclusively on perfectly aligned audio–video pairs. Understanding misalignment is critical: safety-critical applications require millisecond-level localisation of events, and temporal desynchronisation is an emerging attack surface. We introduce a compact evaluation set that injects controlled audio–video time shifts into real-world clips and use it to test two leading audio-visual VLMs, Gemini 2.0 Flash and Qwen-2.5 Omni. Both models obtained accuracies below chance rate in recognizing mis-alignment between audio and visual information, exposing a clear gap in current audio-visual understanding and motivating alignment-aware model development.

1. Introduction

Humans recognize events in daily life by integrating visual and auditory information. They infer information about areas outside their visual field based on sounds, and understand the state transitions of objects—such as breaking or collapsing—by combining what they see and hear.

In the field of computer vision, audio-visual recognition is also a crucial technology and has been applied to a variety of tasks such as Audio-Visual Segmentation (AVS) [1], Audio-Visual Target Speaker Extraction (AV-TSE) [2], and Audio-Visual Event Localization (AVEL) [3]. In these tasks, audio information is used to complement visual input, enabling more detailed event recognition compared to using visual information alone.

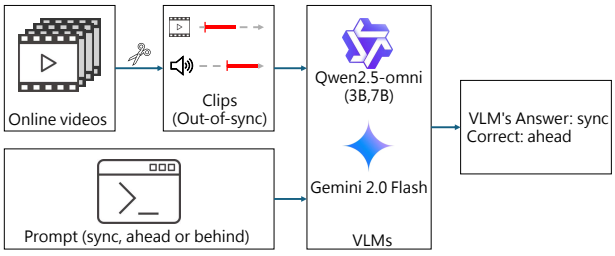

Figure 1. Illustration of Audio-video Temporal Misalignment evaluation.

Furthermore, research on audio-visual VLMs [4] is currently being actively pursued. These models aim to comprehensively understand events from both visual and auditory modalities, enabling functionalities such as caption generation and the localization of sound-emitting objects.

One of the essential capabilities required for audio-visual vision-language models is the ability to accurately understand the temporal relationships between sound and video. This ability is critical for accurately recognizing and localizing audio-visual events. In applications where precise synchronization between visual and auditory input is necessary, desynchronizing the audio and video can serve as a novel form of adversarial attack. Therefore, accurate alignment between audio and visual information is of great importance. However, current methods under consideration do not yet fully achieve perfect synchronization between video and audio.

One of the reasons for this limitation is the inherent difference in modality: video and audio are discretely sampled with different sampling manners. Many VLMs [5–8] extract video features frame by frame in a discrete manner, making it difficult to achieve perfect alignment with audio. While methods such as MRoPE/TMRoPE [5] attempt to synchronize video and audio by segmenting video into 2-second intervals and aligning them via timestamps, it may be challenging to handle misalignments that are less than

2 seconds. Additionally, the datasets used for training and evaluation are typically composed of audio and video that are perfectly synchronized. As a result, these models tend to be vulnerable to even minor misalignments between the modalities.

To address this issue, we conducted an experimental investigation to determine whether current VLMs are capable of recognizing misalignments between video and audio. The VLMs examined in this study were Qwen2.5-Omni [5] and Gemini 2.0 Flash [9]. We constructed a dataset consisting of 3–10 second clips generated from 51 videos collected from online resources. For each clip, the audio track was shifted forward and backward relative to the video in 0.5-second increments, within a range of 0 to 5 seconds. Overall, the dataset has 4,284 video clips. These modified clips were then input to the models to examine whether they could recognize the temporal misalignment between audio and video. In detail, the models were prompted to select the most appropriate description from the following three options: the audio and video are synchronized; the audio lags behind the video; or the audio precedes the video (Figure 1).

As a result, Qwen2.5-Omni achieved an overall accuracy of 4.7%, and Gemini 2.0 Flash achieved 26.0%, both of which are below the random chance level. Furthermore, Qwen2.5-Omni consistently responded with "sync" for all clips, while Gemini 2.0 Flash responded with "sync" in more than 50% of the cases, indicating a need for further investigation into the underlying causes. These findings suggest that current VLMs are not capable of recognizing misalignment between audio and video. However, the ability to detect such misalignment is crucial in many real-world applications and forms a core aspect of human audio-visual understanding. Therefore, improving this capability is essential for developing VLMs that more closely align with human perception.

2. Related Work

2.1. Audio-Visual Recognition Evaluation

Audio Recognition Evaluation for VLMs: Benchmarks for evaluating the performance of audio-visual recognition in VLMs include Massive Multi-task Audio (MMAU) [10], MMAR [11]. MMAU is a benchmark designed to assess the inference and information extraction capabilities of models on sounds, including speech, environmental sounds, and music. The dataset consists of approximately 10,000 QA pairs, enabling evaluation across 27 types of cognitive skills. It comprehensively covers the domains of speech, music, and environmental sounds, allowing for a broad assessment of a model's ability to extract and infer information from various auditory sources. Similarly, MMAR is a benchmark that supports evaluation on speech, environmental sounds, music, and their mixtures. Unlike MMAU, which primarily focuses on event understanding without addressing deep-level reasoning, MMAR introduces a hierarchical inference framework—comprising Signal, Perception, Semantic, and Cultural layers—to assess the model's ability for deeper reasoning.

Audio-Visual Recognition Evaluation for VLMs: LongVALE [8] is constructed using an efficient and scalable pipeline that includes high-quality multimodal long-video filtering, omni-modal event boundary detection, and omni-modal event caption generation based on audiovisual correlation reasoning. While MMAU and MMAR serve as evaluation benchmarks for auditory capabilities, LongVALE is specifically designed to evaluate event recognition in long-form videos by integrating vision, audio, and speech modalities. VAST-27M [12] is a dataset constructed from an open-domain video corpus using a two-stage automated pipeline. It consists of video clips that contain five vision-based captions, five audio-based captions, and one omni-modal caption. Due to the diversity of caption types, the dataset supports various learning tasks such as vision-to-text and audio-to-text. Furthermore, an omni-modal foundation model called VAST has been proposed, trained on the VAST-27M dataset. VAST is capable of understanding and processing all four modalities—vision, audio, subtitles, and text—and can handle a wide range of tasks across these modalities.

These benchmarks and models conduct evaluation and training exclusively on videos in which the audio and visual streams are synchronized. Therefore, it is unclear whether they are capable of recognizing misalignment between audio and video. In audio-visual event recognition, auditory information serves as a crucial complement to the visual content, enabling a more detailed understanding of events. Consequently, the inability to detect audio-visual misalignment poses a significant problem. In response, we conduct an experiment to investigate whether current VLMs are capable of recognizing such misalignment.

2.2. Audio-Visual Recognition Method

VideoLLaMA2 [7] is based on its predecessor, VideoLLaMA [6], and is composed of two branches: a visual-language branch and an audio-language branch, each operating independently. These two branches are connected via a large language model, which performs the cross-modal processing. The training datasets for video-text sources, such as Panda-70M [13]. Additionally, WavCaps [14] is used during the initial stage of training for the audio-language branch. Qwen2.5-Omni is composed of four components: a Vision Encoder, an Audio Encoder, a Thinker, and a Talker. The Thinker accepts multiple types of modalities as input, including text, images, audio, and video, and is responsible for text generation. The Talker receives the rep-

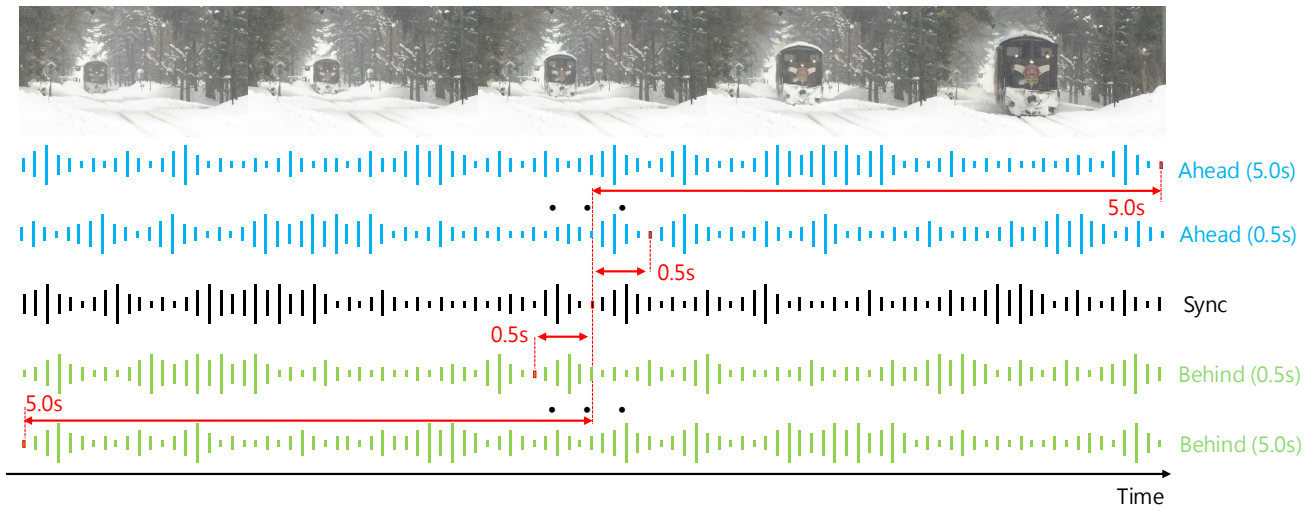

Figure 2. Illustration of the process of creating video clips.

resentations and text generated by the Thinker and outputs individual speech tokens. Pretraining is conducted in three stages: in the first stage, training is focused specifically on the vision and audio encoders; in the second stage, training is performed using a wide range of multimodal data; and in the final stage, the model is trained on long sequence data with a sequence length of 32k to handle complex, extended input sequences. Meanwhile, Gemini 2.0 Flash and Gemini 2.5 Pro [9] are closed-source models capable of processing videos with audio, and they have achieved the highest accuracies in various audio-visual tasks. However, the architecture and training data of the Gemini models remain undisclosed.

All these methods are trained on aligned audio-visual datasets, which might influence their ability to recognize misalignment between audio and visual data. However, understanding misalignment is a critical ability in various applications, yet it is underexplored. Therefore, in our research, we selected an open-sourced model, Qwen2.5-Omni, and a closed-sourced model, Gemini 2.0 Flash, and evaluated how they recognize misalignment between audio and visual data.

3. Experiment

3.1. Evaluation Dataset

To cover different video genres and examine how model performance varies across them, we collected 51 online videos from five distinct genres. Video lengths ranged from one minute to 15 minutes. Each video was manually verified to ensure proper synchronization between audio and video. The genres and the number of videos used are shown in Table 1.

Table 1. The genre and number of videos used.

Video genre	**Number**
cooking (daily cooking videos)	11
instrument (violin, guitar, drum set, piano, concert)	10
sports (tennis, table tennis, football, basketball, running match)	10
autonomous sensory meridian response (ASMR)	10
speech (English, Chinese, French, Italian, Japanese, Arabic)	10

As shown in Figure 2, from each video, clips of 3 seconds, 5 seconds, 7 seconds, and 10 seconds in length were randomly extracted. For each of these clips, versions were prepared in which the audio was shifted both forward and backward relative to the video in 0.5-second increments within a range of 0 to 5 seconds, resulting in 20 video clips with jittered audio and one original aligned video for each clip. Overall, the dataset contains 4,284 video clips. These clips were then input to the models to examine whether they could detect the misalignment between audio and video.

3.2. Methods

The models examined in this study are the open-source model Qwen2.5-Omni and close-source Gemini 2.0 Flash. For Qwen2.5-Omni, both the 3B and 7B variants were evaluated. These three models were compared against a random baseline.

The prompt used for the evaluation is shown in Figure 3. The models were instructed to respond with “sync” if they judged the audio and video to be synchronized, “ahead” if

```
f"You will watch a {args.clip_len}-second video
    with audio.\n"
"Reply with ONLY ONE word:\n"
"sync = audio matches video\n"
"ahead = audio earlier than video\n"
"behind = audio later than video"
```

Figure 3. Illustration of the input prompt.

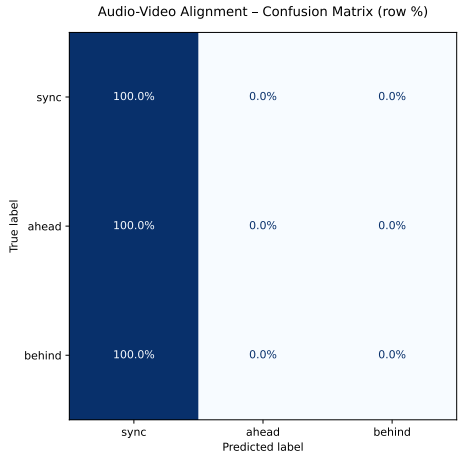

Figure 4. Recognition results of audio-visual misalignment of Qwen2.5-Omni.

the audio was perceived to be ahead of the video, and "behind" if the audio was perceived to be delayed relative to the video. In addition, the variable "args.clip_len" is designed to store the duration, in seconds, of the input clip.

3.3. Result Analysis

As a result of the experiment, Qwen2.5-Omni responded with "sync" for all clips across all genres in both the 3B and 7B versions, resulting in an overall accuracy of 4.7%. Notably, adjusting the input prompt did not produce any significant change in performance. A graph illustrating the results is shown in Figure 4.

Gemini 2.0 Flash was evaluated using only 10-second clips. As a result, the overall accuracy was 26.0%, and the comparison between the ground truth and the model's predictions is illustrated in the graph shown in Figure 5. The accuracy for each video genre is shown in Table 2.

When examining the accuracy by video genre, Gemini 2.0 Flash demonstrated lower accuracy for videos involving musical instrument performances and cooking. This is considered to be due to the relatively limited number of musical performance videos used during training, and in the case of cooking videos, the presence of diverse actions and temporally irregular sounds compared to other genres, which made temporal alignment of audio more difficult.

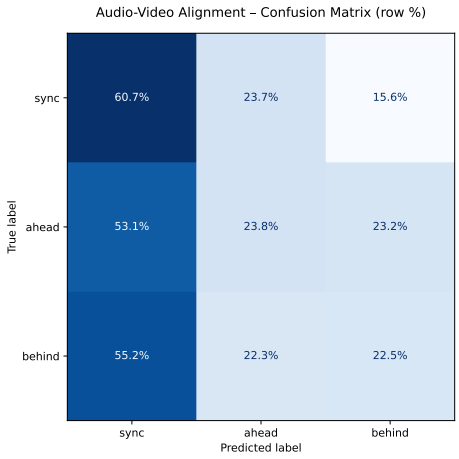

Figure 5. Recognition results of audio-visual misalignment of Gemini 2.0 Flash.

Table 2. Per-video-genre accuracy of Gemini 2.0 Flash.

Video genre	**Accuracy (%)**
cooking	13.5
instrument	20.3
sports	28.2
ASMR	33.7
speech	34.0

Overall, in our experiment, both models performed worse than the random baseline (33.3% accuracy) and fell far short of human performance, revealing a critical limitation of existing VLMs. This highlights a potential vulnerability to adversarial attacks and raises concerns about the models' robustness and trustworthiness.

4. Discussion and Future Work

Our experiments reveal that two state-of-the-art VLMs perform below chance at detecting audio–video misalignment.

As future work, we plan to investigate the underlying reasons why Qwen2.5-Omni responded with "sync" in all cases, and why Gemini 2.0 Flash also selected "sync" with a frequency exceeding 50%. Additionally, while the present experiment involved shifting the audio relative to the video, future studies will explore other experimental conditions, such as exploring the spatial and semantic misalignment recognition of current models, replacing parts of the audio with audio from different videos, using video genres not included in the current dataset, and evaluating the effect of adding noise to the audio.

References

[1] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, and Bhiksha Raj. Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3402–3413. IEEE Computer Society, 2024. 1

[2] Ruijie Tao, Xinyuan Qian, Yidi Jiang, Junjie Li, Jiadong Wang, and Haizhou Li. Audio-visual target speaker extraction with reverse selective auditory attention. *arXiv preprint arXiv:2404.18501*, 2024. 1

[3] Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards open-vocabulary audio-visual event localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8362–8371, 2025. 1

[4] Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. *arXiv preprint arXiv:2504.02061*, 2025. 1

[5] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 1, 2

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, 2024. 2

[7] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2

[8] Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18959–18969, 2025. 1, 2

[9] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 3

[10] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024. 2

[11] Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025. 2

[12] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023. 2

[13] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 2

[14] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 2