# SVHalluc: Benchmarking Speech–Vision Hallucination in Audio-Visual Large Language Models

Supplementary Material

## A. Additional Qualitative Results

In this section, we visualize video–question pairs from SVHalluc, together with predictions from the latest audio–visual large language models (LLMs). Our SVHalluc shows diversity in both task design and video scenarios, providing a comprehensive benchmark for evaluating speech-vision hallucination in audio-visual LLMs.

- **Speaker variability**. SVHalluc includes videos where multiple people appear visually in the scene (Figure S1(a)) or only present in the audio (Figure S1(b)).
- **Illumination variability**. Our dataset includes videos under diverse lighting conditions, such as bright environments (Figure S2(a)) or dusky settings (Figure S2(b)).
- **Camera viewpoint diversity**. SVHalluc includes videos with different viewpoints, such as fixed top-down views (Figure S3(a)), moving side-view (Figure S3(b)), and transitions between object-centric and human-centric views (Figure S3(c)).

## B. GPT Prompts for Dataset Creation

During dataset creation, we adopt GPT [2] models to extract the objects and actions for task formulation. To find objects and actions that uniquely appear in video or speech, we prompt the following instruction based on the caption of the video and speech. We find that GPT could return good results in most cases, and we perform a human verification based on GPT results.

> Given caption1 and caption2, extract action and objects. Action refers to verbs, e.g., put in the bowl.
> First, extract all actions and objects from each caption. Second, find unique actions and objects that only appear in one caption but not the other, termed as $cap1\_only_{act}$, $cap1\_only_{obj}$, $cap2\_only_{act}$, $cap2\_only_{obj}$, respectively.
> Following is one example. Output in the json format.
> Input:
> caption1: put the eggs in the bowl caption2: cut the eggs and cucumber
> Output:
> { $cap1\_only_{act}$: ["put in the bowl"],
> $cap1\_only_{obj}$: [],
> $cap2\_only_{act}$: ["cut"],
> $cap\_only_{obj}$: ["cucumber"],
> }

## C. Dataset Quality Control

To ensure the dataset quality, we design diverse strategies to filter out unqualified samples, including rule-based strategy and GPT-based strategy. For example, we filter out short videos and speeches to avoid the case that the events are not clearly shown. We also filter out samples if the speech only includes general information, e.g., "this looks great", instead of describing the events. After dataset creation, we perform a final human verification to ensure the dataset quality. The high performance of Gemini-2.5 Pro also indicates the high quality of our dataset, while highlighting the severe hallucinations of open-source models.

## References


[1] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3

[2] openAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025. 1

[3] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 2, 3

[4] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 2, 3


Figure S1. **Speaker variability in SVHalluc.** We show predictions of Qwen3-Omni [4], Qwen2.5-Omni [3], and Video-LLaMA 2 [1] in diverse speaker settings.

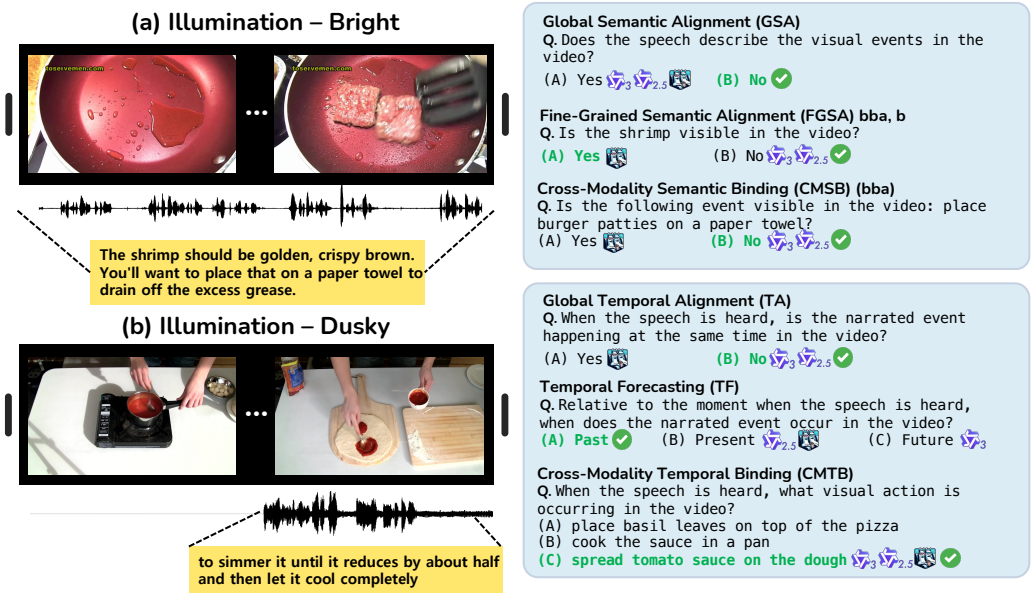


Figure S2. **Illumination variability in SVHalluc.** We show predictions of Qwen3-Omni [4], Qwen2.5-Omni [3], and Video-LLaMA 2 [1] in diverse illumination settings.

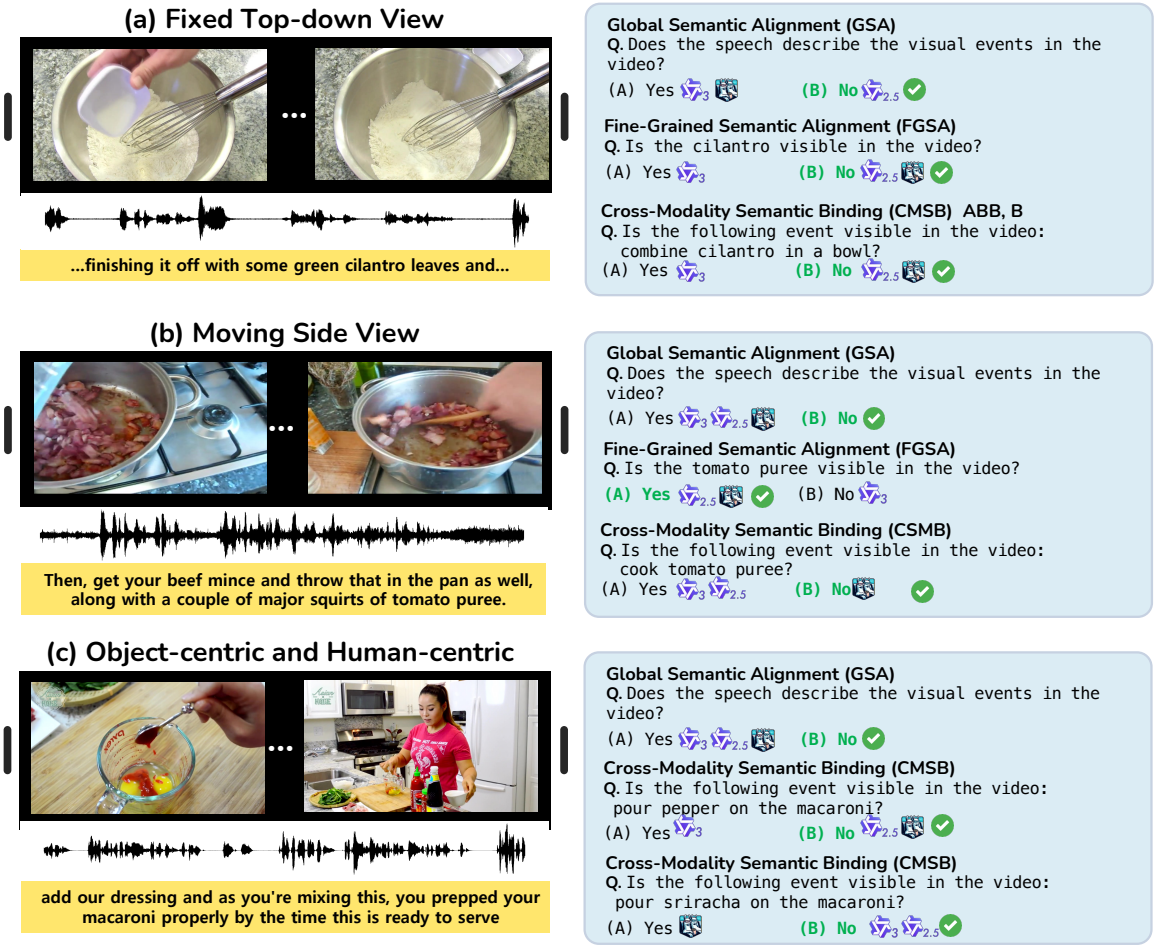


Figure S3. **Camera viewpoint variability in SVHalluc.** We show predictions of Qwen3-Omni [4], Qwen2.5-Omni [3], and Video-LLaMA 2 [1] in diverse camera viewpoint settings.