Multi-speaker Attention Alignment for Multimodal Social Interaction

Supplementary Material

Appendix

Tasks Instructions

We evaluate our method on four types of social interaction tasks. Each task takes as input the video frames, transcripts, and speaker bounding boxes. We describe the tasks and example prompts used for MLLMs below:

- **Video Question Answering.** This task requires answering questions grounded in multi-party dialogue videos. In TVQA+, each question is accompanied by five candidate choices. Instruction prompt:
  `<video>\nWatch this video of speakers social interaction, be aware of their non-verbal behaviours. Read their conversation, question and choose the correct answer. {Conversation}. Q: How does Sheldon feel? a0: tired, ..., a4: angry.`
- **Speaking Target Identification.** This task identifies the addressee (speaking target) of the current speaker in the dialogue. Instruction prompt:
  `<video>\nWatch this video of {N} speakers social interaction, be aware of their non-verbal behaviours. Read the conversation. {Conversation}. Predict the speaking target (speaking to whom) of this sentence: Speaker 0: Who were you?`
- **Pronoun Coreference Resolution.** This task aims to resolve pronouns in the dialogue transcripts to their corresponding speakers. Instruction prompt:
  `<video>\nWatch this video of {N} speakers social interaction, be aware of their non-verbal behaviours. Read the conversation. {Conversation}. Predict which speaker should be the 'he' in this sentence: Speaker 1: Did he not say that?`
- **Mentioned Player Prediction.** This task requires linking a dialogue mentioned name to the correct participant appearing in the video. Instruction prompt:
  `<video>\nWatch this video of {N} speakers social interaction, be aware of their non-verbal behaviours. Read the conversation. {Conversation}. Predict which speaker should be the 'Mitchell' in this sentence: Speaker 3: I think it's Mitchell.`

Baseline Methods

We compare our method against several baseline approaches in Table 2, described as follows:

- **Random.** For the VQA task, the model randomly selects one answer from five candidates. For MMSI tasks, it randomly selects one speaker among N candidates.
- **ST-VLM-7B [27] and TLNet [37]** are highly competitive models on TVQA+. However, since our setting requires at least one speaker with bounding box annotations, our training and test sets differ from theirs. Despite this difference, our method substantially outperforms these baselines.
- **MMSI [31]** employs a transformer to align and fuse text features from language models with visual interaction features derived from bounding boxes and keypoints. Classification tasks are then performed via a masked modeling objective to solve three social interaction tasks. We report their best-performing result, i.e., the RoBERTa-based baseline, for comparison.
- **OnlineMMSI [35]** leverages bounding boxes and keypoints as visual prompts to align multiple speakers for MLLMs. Since the implementation was not publicly released, we re-implemented it based on Qwen2.5-VL, using only bounding box annotations for a fair comparison. Notably, our reproduced results are even higher than original reported numbers, which may be due to preprocessing differences, but this does not affect the comparison fairness.
- **Qwen2.5-Text [5]** is a text-only baseline where Qwen2.5-VL is given only the dialogue transcripts without any video or location inputs.
- **Qwen2.5-VL w/ VL [5]** is a strong MLLM baseline. We use the instruction prompt described above, which is the same as our method's visual and language inputs.
- **Qwen2.5-VL w/ VLB [5]** takes additional bounding box coordinates as text prompts, e.g., `Speaker 1, t=0, [100,100,300,400].`
- **InternVL3 [87] / LLaVA-NeXT-Video [77] w/ VLB** are also competitive transformer-based MLLMs. They take the same input prompts as **Qwen2.5-VL w/ VLB**, including both visual content and box coordinates in text.

Experiment with Detected Boxes

To evaluate the generalization of our method in real application settings, we conduct an additional evaluation experiment where detected boxes replace the ground-truth boxes provided by the datasets. Specifically, for each video, we use SAM2 to track speaker boxes across all frames, initialized from the ground-truth boxes on the first frame. The model is trained using ground-truth annotations.

Table 6. Effect of using detected bounding boxes.

Method	Box	TVQA+	MMSI		
		VideoQA	*T*	*P*	*M*
Qwen2.5-VL	GT	86.1	64.8	76.6	62.4
Qwen2.5-VL	SAM2	85.1	63.3	75.5	61.7
Qwen2.5-VL+Ours	GT	**87.3**	**68.5**	**78.6**	**66.0**
Qwen2.5-VL+Ours	SAM2	85.9	67.6	77.8	65.6

As shown in Tab. 6, our method still maintains advantages compared to baseline approaches. This is because baseline methods rely on injecting box coordinates into the text input, forcing the model to indirectly infer spatial speaker alignment through language. This mechanism does not enable the model to truly learn multimodal, multi-speaker alignment, and therefore it cannot meaningfully utilize or correspond to the spatial information encoded by the boxes. In contrast, our method injects bias directly at the attention computation level, enabling the model to form explicit and modality-grounded associations between speakers' visual and textual representations.

We also observe that our method's accuracy drops less on MMSI compared to TVQA+. This is expected, as videos in MMSI depict board-game interactions with largely *fixed participant seating*, resulting in much more stable detected boxes than the dynamic camera viewpoints and frequent occlusions in TVQA+. This finding highlights a promising application scenario of our method: in social environments with relatively stable spatial layouts, such as board-game interactions or video-meeting settings, our method can reliably leverage high-quality detected boxes to achieve strong and generalizable social understanding performance.

Computation Cost

We evaluate the computational overhead introduced by our method in terms of inference time and FLOPs. As shown in Tab. 7, our method increases inference time by approximately 20%, while having a minimal impact on the model's forward FLOPs. This additional computation mainly arises from reading the bounding boxes and speaker-related text tokens to determine the positions of the social-aware bias. Once these positions are computed, during attention calculation only a small subset of attention heads are activated, and the bias is applied via one matrix multiplication, a single column-wise max operation, and one matrix addition, resulting in a negligible impact on FLOPs. Overall, the computational cost remains lightweight and fully acceptable for MLLMs performing social interaction tasks.

Table 7. Inference time and FLOPs on TVQA+. Inference time is measured on 2,211 samples, and FLOPs are computed on the sample with the largest input length in the dataset.

Method	Time(s)	FLOPs($\times 10^{12}$)
Qwen2.5-VL	455	4.01
Qwen2.5-VL+Ours	574	4.05
InternVL3	409	8.84
InternVL3+Ours	516	8.97

Zero-shot Comparisons

To further validate the robustness of our approach, we conduct zero-shot evaluations on InternVL3 and Qwen2.5-VL in Tab. 8. We also compare our method against two recent training-free baselines: ControlMLLM++ [67], which utilizes current speaker boxes as input, and ViCrop-rel-att [75]. Since these implementations are designed for single-image VQA and we feed only one speaking video frame as input, they face an inherent disadvantage in this setting.

Despite this, our method consistently outperforms them across all tasks. This is because our design explicitly targets multi-person alignment by dynamically guiding different speaker tokens to attend to distinct regions, while [67] and [75] mainly focus on question-relevant regions and lack generalization to multi-speaker conversations. These results strongly validate our core idea: the limited understanding of multi-speaker dynamic interactions is a key bottleneck for social reasoning tasks in current MLLMs. Enhancing multi-person attention alignment yields consistent gains regardless of whether the model is fine-tuned or not.

Table 8. Zero-shot evaluation and comparison with training-free methods. *: image-based.

Method	TVQA+	MMSI		
	VideoQA	*T*	*P*	*M*
InternVL3	76.5	58.7	38.8	50.8
InternVL3+Ours	76.5	**60.7**	**40.2**	**51.2**
Qwen2.5-VL	**75.9**	52.4	24.8	41.4
Qwen2.5-VL+[67]*	75.4	53.9	**29.7**	39.4
Qwen2.5-VL+[75]*	73.7	53.5	29.5	42.2
Qwen2.5-VL+Ours	**75.9**	**57.3**	29.5	**43.7**

Analysis of Active Heads and Transformer Layers

To identify which transformer layers are most responsible for capturing social dynamics, we investigate the distribution of active heads and cross-attention weights across layers 10–19 for Qwen2.5-VL on the MMSI dataset in Fig. 5. Our analysis reveals that layers 13, 14, and 16 consistently exhibit the highest activation levels and strongest cross-modal attention. Since these layers demonstrate similar behavioral patterns, we select layer 16 as a representative for our visualizations in the main text. This suggests that multi-speaker alignment is most effectively processed in the middle-to-late stages of the transformer architecture.

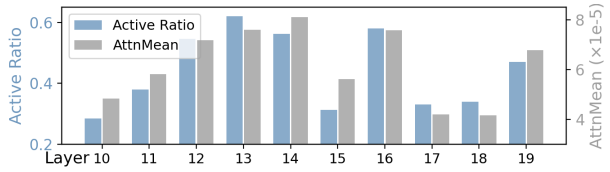

Figure 5. Active head ratio and mean attention weight in Qwen2.5-VL.

Ablations on InternVL3

We extend our ablation studies to the InternVL3 model to evaluate the sensitivity of the threshold λ. As shown in Tab. 9, while the optimal numerical value for λ may vary between architectures, both Qwen2.5-VL and InternVL3 achieve peak performance when the active head ratio is approximately 20%. This indicates that the active ratio serves as a reliable and generalizable heuristic for hyperparameter selection across different MLLMs.

Table 9. Ablation study of threshold λ on InternVL3. We report the percentage of active heads alongside accuracy on TVQA+ and MMSI.

λ	**Active Heads (%)**	**TVQA+**	**MMSI**		
		VideoQA	*T*	*P*	*M*
0	35.7	88.2	67.7	78.5	64.7
2e-5	23.7	**89.3**	<u>68.8</u>	**81.2**	64.1
5e-5	14.9	<u>89.1</u>	**69.7**	<u>80.5</u>	**65.7**
8e-5	9.1	89.0	67.3	77.8	63.0
inf	0.0	85.6	65.0	76.9	63.0