

Language-guided Frequency Modulation for Large Vision-Language Models

Supplementary Material

This Appendix provides additional details to support the main paper and is organized as follows:

- **1:** More Experimental Results;
- **2:** Additional Ablation Study;
- **3:** More Qualitative Results;
- **4:** Limitation and Broader Impacts.

1. More Experimental Results

The additional experimental results are organized into two parts: (1) we evaluate the compatibility and scalability of the proposed method by integrating it into SoTA LVLMS of varying types and sizes, including the Qwen [2, 16] and InternVL [6, 24] families; and (2) we conduct a training-free evaluation on video understanding tasks.

1.1. General Visual Question Answering and Math-Related Benchmarks

To assess the effectiveness and generalizability of the proposed LFM, we integrate it into a variety of SoTA LVLMS and evaluate its performance across general visual question answering and math-related benchmarks. As shown in Table 1, the selected evaluation suite includes MME (sum), MMB (EN/CN), MMBv1.1, MMVet (turbo), MME-RealWorld, MMStar, and MathVista. The tested models cover a wide range of parameter scales (from 1B to 8B), and include leading architectures such as InternVL2.5 and Qwen2.5-VL. The results demonstrate that LFM consistently enhances the performance of various LVLMS across all benchmarks. These results confirm that LFM scales effectively across LVLMS of various sizes and yields significant improvements in both general and math-intensive multimodal reasoning tasks.

1.2. Comparison in Video Understanding

Building on the image-to-video paradigm introduced in FreeVA [18], we test our image-text pre-trained model directly on video-based dialogue tasks without any additional training. This innovative approach highlights the potential of leveraging image-text training for video comprehension tasks, bridging the gap between static and dynamic vision modalities.

The experimental results presented in Table 3 demonstrate the competitive performance of FreeVA-based methods. The LFM+FreeVA (34B) model achieves the highest accuracy on MSVD-QA, MSRVT-QA and ActivityNet-QA, outperforming all other methods. This indicates the model’s superior ability to generalize from image-text training to video question-answering tasks. The 13B variant

of LFM+FreeVA also demonstrates strong performance, achieving 77.6 on MSVD-QA, surpassing other larger models. This further validates the scalability of LFM across different model sizes. These results demonstrate the versatility and effectiveness of our pre-trained model in achieving outstanding performance across diverse video understanding benchmarks.

2. Additional Ablation Study

2.1. Hyperparameter Experiments

To investigate the impact of the modulation coefficient β in the fusion stage of LFM, we conducted an ablation study by varying β from 0 to 0.15 and evaluating performance on the MMB and MMStar benchmarks. As illustrated in Figure 1, performance consistently improves as β increases from 0 to 0.12, with the best results observed at $\beta = 0.12$. Beyond this point, performance begins to decline slightly, suggesting a diminishing return from further amplification of the modulated signal. We adopt $\beta = 0.12$ as the default setting in our experiments.

In addition to the strategy comparisons discussed earlier, we conducted ablation studies to examine the impact of varying the baseline and step size of the coefficient α_k . As shown in the Figure 2, we tested four baseline values (0.20, 0.25, 0.30, 0.35) across three step sizes (0.01, 0.02, 0.03). The results consistently demonstrate that moderate step sizes (particularly 0.02) tend to yield better performance across different group configurations. Among the baselines, $\alpha = 0.25$ with a step size of 0.02 achieves the highest performance, especially under the 3-group setting. These findings reinforce our choice of using a baseline α of 0.25 and a step size of 0.02 in the decreasing strategy as the default configuration.

2.2. MLP Design Choices

To assess the design choice of the visual-language projector described in Section 3.2.5, we conducted an ablation study comparing two alternative strategies for integrating the frequency-modulated features from the grouped vision encoder layers before feeding them into the LLM.

The default strategy in the main paper, referred to as “Concat & MLP”, concatenates the outputs from all K modulation groups along the channel dimension, resulting in a $K \cdot D$ -dimensional feature vector per patch. This vector is then projected to the LLM’s expected hidden size D_{llm} using a two-layer MLP. We compare this design against an alternative “Add & MLP” strategy, where the modulated outputs of all K groups are first element-wise summed, re-

Table 1. Performance of LFM on general visual question answering and math-related benchmarks when integrated with State-of-The-Art LVLMs.

Model Name	Model Size	MME _{sum}	MMB _{EN/CN}	MMBv1.1 _{EN}	MMVet _{turbo}	MME-RW _{EN}	MMStar	MathVista
LLaVA-OneVision [9]	0.5B	1438.0	61.6 / 55.5	59.6	32.2	–	37.7	34.8
InternVL2 [35]	1B	1794.4	65.4 / 60.7	61.6	32.7	40.2	45.7	
InternVL2.5 [4]	1B	1950.5	70.7 / 66.3	68.4	48.8	44.2	50.1	47.1
LFM (w/ InternVL2.5)	1B	2015.1	73.1 / 68.3	71.1	52.3	47.3	54.5	48.7
Qwen2-VL [16]	2B	1872.0	74.9 / 73.5	72.2	49.5	–	48.0	43.0
InternVL2 [5]	2B	1876.8	73.2 / 70.9	70.2	39.5	47.3	50.1	46.3
InternVL2.5 [4]	2B	2138.2	74.7 / 71.9	72.2	60.8	48.8	53.7	51.1
Qwen2.5-VL [3]	3B	2157.0	79.1 / 78.1	77.4	61.8	53.1	55.9	61.2
LFM (w/ InternVL2.5)	2B	2165.6	76.3 / 73.5	74.1	61.8	50.3	55.4	52.7
LFM (w/ Qwen2.5-VL)	3B	2188.5	81.4 / 80.7	79.5	62.7	54.9	58.3	62.9
Phi-3.5-Vision [1]	4B	–	76.0 / 66.1	72.1	43.2	–	47.5	43.9
InternVL2 [5]	4B	2059.8	78.6 / 73.9	75.8	51.0	52.1	54.3	58.6
InternVL2.5 [4]	4B	2337.5	81.1 / 79.3	79.3	60.6	55.3	58.3	60.5
LFM (w/ InternVL2.5)	4B	2359.7	83.2 / 81.6	81.8	63.5	57.8	60.4	62.5
Qwen2-VL [16]	7B	2326.8	83.0 / 80.5	80.7	62.0	56.5	60.7	58.2
MiniCPM-V2.6 [21]	7B	2348.4	81.5 / 79.3	78.0	60.4	–	59.6	60.6
InternVL2 [5]	8B	2210.3	81.7 / 81.2	81.2	52.3	53.5	57.6	58.3
InternVL2.5 [4]	8B	2344.1	84.6 / 82.6	83.2	62.8	59.1	60.2	64.4
Qwen2.5-VL [3]	7B	2347.0	83.5 / 83.4	82.6	67.1	68.5	63.9	68.2
LFM (w/ InternVL2.5)	8B	2376.6	85.9 / 84.1	85.0	64.3	63.1	62.7	66.1
LFM (w/ Qwen2.5-VL)	7B	2385.3	85.0 / 84.7	84.2	69.8	70.4	66.1	69.8

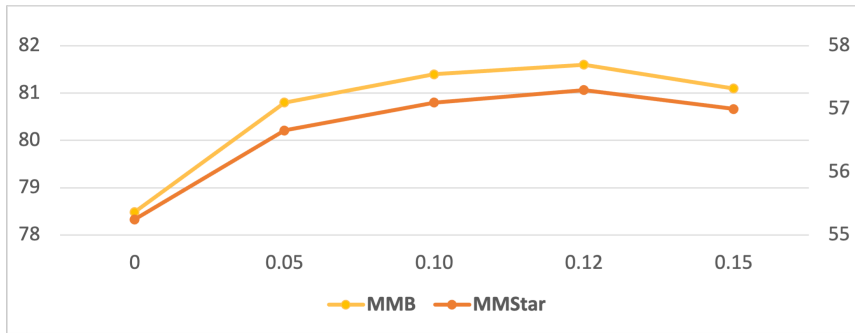


Figure 1. Performance under different coefficients β .

sulting in a D -dimensional feature vector per patch. The resulting feature is then passed through a smaller MLP to map from D to D_{lim} .

As shown in Table 4, both approaches are effective in enhancing language-aligned visual understanding. However, “Concat & MLP” consistently outperforms “Add & MLP” across benchmarks, owing to its capacity to preserve group-specific modulation information. In contrast, “Add

& MLP” introduces less computational overhead due to the reduced dimensionality of the projection input, offering a lightweight and viable alternative when efficiency is a concern.

Table 2. Performance on OCR, chart, and document understanding benchmarks.

Model Name	Model Size	ChartQA	TextVQA	DocVQA	InfoVQA	SEED-2-Plus	CharXiv _{RQ/DQ}
Qwen2-VL [16]	2B	73.5	79.7	90.1	65.5	62.4	–
Aquila-VL [8]	2B	76.5	76.4	85.0	58.3	63.0	–
InternVL2 [6]	2B	76.2	73.4	86.9	58.9	60.0	21.0 / 40.6
InternVL2.5 [4]	2B	79.2	74.3	88.7	60.9	60.9	21.3 / 49.7
LFM (w/ InternVL2.5)	2B	80.7	76.4	89.5	63.7	64.1	24.5 / 50.8
Phi-3.5-Vision [1]	4B	81.8	82.0	69.3	36.6	62.2	–
InternVL2 [5]	4B	81.5	80.4	89.2	67.0	63.9	24.5 / 48.0
InternVL2.5 [4]	4B	84.0	76.8	91.6	72.1	66.9	24.9 / 61.7
Qwen2.5-VL [3]	3B	84.0	79.3	93.9	77.1	67.7	31.3 / 58.6
LFM (w/ InternVL2.5)	4B	86.1	78.5	93.1	74.7	69.3	27.2 / 64.2
LFM (w/ Qwen2.5-VL)	3B	85.8	81.2	95.4	79.6	69.8	33.7 / 61.3
Ovis1.6-Gemma2 [14]	9B	–	–	–	–	–	–
MiniCPM-V2.6 [21]	6B	82.4	80.1	90.8	–	65.7	31.0 / 57.1
Molmo-7B-D [7]	7B	84.1	81.7	92.2	72.6	–	–
Qwen2-VL [16]	7B	83.4	81.9	94.4	75.6	69.0	–
InternVL2 [5]	8B	83.3	77.4	91.6	74.8	67.5	31.2 / 56.1
InternVL2.5 [4]	8B	84.8	79.1	93.0	77.6	69.7	32.9 / 68.6
Qwen2.5-VL [3]	7B	87.3	84.9	95.7	82.6	70.4	42.5 / 73.9
LFM (w/ InternVL2.5)	8B	85.9	81.4	94.8	79.6	70.3	35.4 / 70.5
LFM (w/ Qwen2.5-VL)	7B	89.2	87.0	96.1	83.5	72.9	44.7 / 75.8

Table 3. Performance comparison across various methods on multiple QA and benchmark tasks.

Method	LLM Size	GPT-3.5 Version	Train Free	MSVD-QA		MSRVTT-QA		ActivityNet-QA		Video-ChatGPT Benchmark				
				Acc	Score	Acc	Score	Acc	Score	CI	DO	CU	TU	CO
FrozenBiLM [19]	0.9B	MAR	✗	33.8	–	16.7	–	25.9	–	1.96	2.18	2.16	1.82	1.79
Video-LLAMA [22]	7B	MAR	✗	51.6	2.5	29.6	1.8	12.4	1.1	2.03	2.32	2.30	1.98	2.15
LLAMA-Adapter [23]	7B	MAR	✗	–	–	–	–	–	–	–	–	–	–	–
VideoChat [10]	7B	MAR	✗	56.3	2.8	45.0	2.5	26.5	2.2	2.23	2.50	2.53	1.94	2.24
Video-ChatGPT [15]	7B	MAR	✗	64.9	3.3	49.3	2.8	35.2	2.7	2.50	2.57	2.69	2.16	2.20
VaQuitA [17]	7B	MAR	✗	74.6	3.7	68.6	3.3	48.8	3.3	–	–	–	–	–
LLAVA+FreeVA [18]	7B	MAR	✓	81.5	4.0	72.9	3.5	58.3	3.5	2.88	2.52	3.25	2.32	3.07
BT-Adapter [13]	7B	JUN	✗	67.5	3.7	57.0	3.2	45.7	3.2	2.68	2.69	3.27	2.34	2.46
Video-LLAMA [12]	7B	JUN	✗	70.7	3.9	59.2	3.5	45.3	3.3	–	–	–	–	–
LLAMA-VID [11]	13B	JUN	✗	70.0	3.7	58.9	3.3	47.5	3.3	3.07	3.05	3.60	2.58	2.63
LLAVA+FreeVA [18]	13B	JUN	✓	71.8	3.8	59.2	3.5	54.5	3.5	2.90	2.52	3.26	2.32	3.07
LLAVA+FreeVA [18]	13B	JAN	✓	74.4	4.1	61.1	3.6	51.6	3.5	2.88	2.52	3.25	2.34	3.05
DC+FreeVA [20]	7B	JAN	✓	75.0	4.1	58.4	3.5	52.2	3.5	2.80	2.51	3.17	2.22	3.05
DC+FreeVA [20]	13B	JAN	✓	75.1	4.1	60.8	3.5	52.6	3.5	2.85	2.53	3.23	2.29	2.96
DC+FreeVA [20]	34B	JAN	✓	77.4	4.2	62.1	3.6	55.8	3.6	3.00	2.53	3.25	2.65	2.92
LFM+FreeVA	7B	JAN	✓	77.3	4.2	60.9	3.5	53.7	3.5	2.85	2.53	3.19	2.31	2.97
LFM+FreeVA	13B	JAN	✓	77.6	4.2	61.6	3.5	54.7	3.6	2.97	2.55	3.44	2.45	3.02
LFM+FreeVA	34B	JAN	✓	79.4	4.4	63.5	3.6	58.1	3.6	3.01	2.57	3.51	2.73	3.04

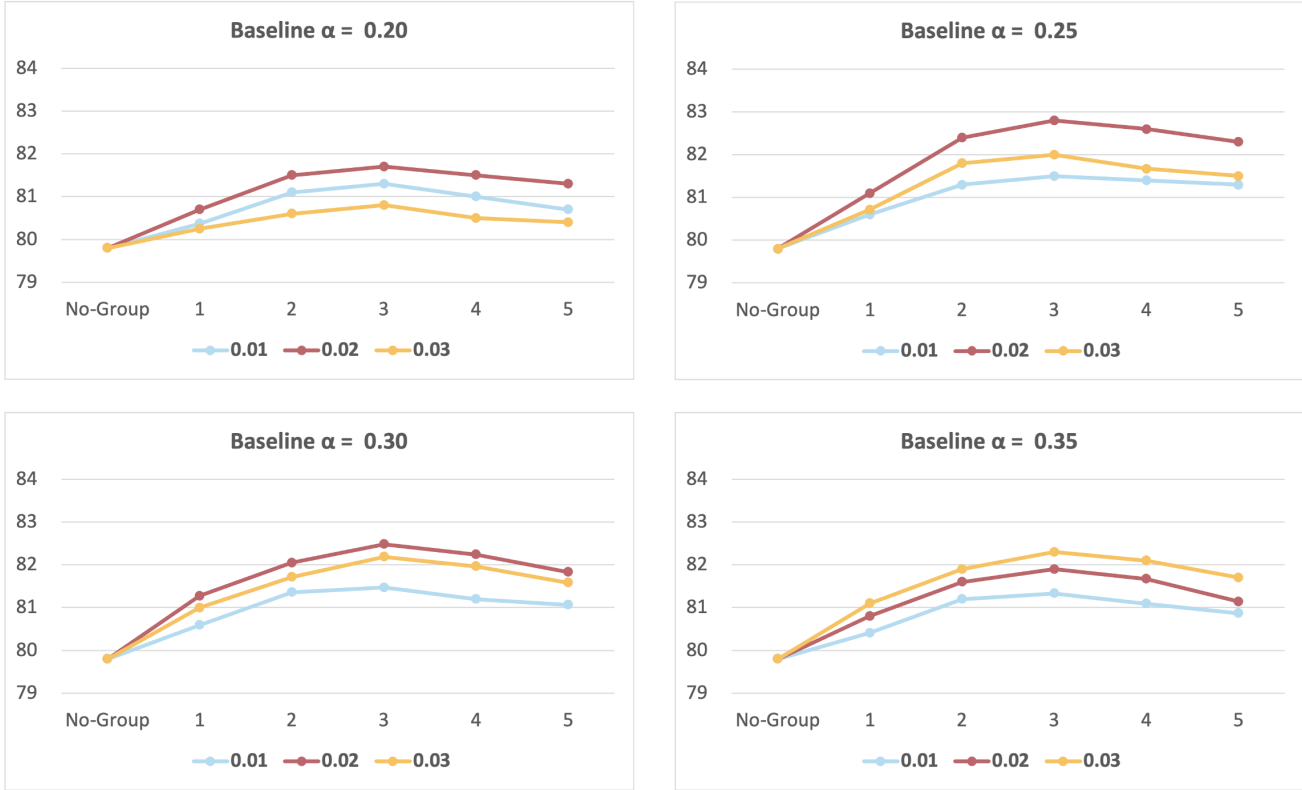


Figure 2. Performance under different coefficients α_k .

Table 4. Performance of LFM on general visual question answering and math-related benchmarks with different MLP design integrations.

Model	MLP Design	MME _{sum}	MMB _{EN/CN}	MMBv1.1 _{EN}	MMVet _{turbo}	MME-RW _{EN}	MMStar	MathVista
InternVL2.5-8B [4]	/	2344.1	84.6 / 82.6	83.2	62.8	59.1	60.2	64.4
LFM (w/ InternVL2.5)	Add & MLP	2368.3	85.3 / 83.5	84.7	64.0	61.5	61.9	65.5
LFM (w/ InternVL2.5)	Concat & MLP	2376.6	85.9 / 84.1	85.0	64.3	63.1	62.7	66.1
Qwen2.5-VL-7B [3]	/	2347.0	83.5 / 83.4	82.6	67.1	68.5	63.9	68.2
LFM (w/ Qwen2.5-VL)	Add & MLP	2370.6	84.3 / 83.8	83.7	69.0	69.8	65.3	69.1
LFM (w/ Qwen2.5-VL)	Concat & MLP	2385.3	85.0 / 84.7	84.2	69.8	70.4	66.1	69.8

3. More Qualitative Results

Figures 3, 4, 5, 6 and 7 present image-text multimodal dialogues covering a broader range of scenarios. The results demonstrate that LFM excels in multimodal understanding, showcasing strong capabilities in both visual comprehension and text generation. Qualitative evaluations highlight the model’s proficiency in addressing diverse challenges across various domains, including mathematics, charts, and everyday scenes.

Figure 8 extends these findings to video-text multimodal dialogues, where the model, trained on image-text data and enhanced with the FreeVA [18], achieves effective video un-

derstanding and demonstrates robust vision-language alignment in dynamic temporal scenarios.

4. Limitation and Broader Impacts

LFM demonstrates strong generalization across vision-language tasks and scales effectively across diverse model sizes and backbones, yet certain limitations warrant consideration. Although the frequency-domain modulation avoids introducing additional trainable parameters within the core mechanism, its reliance on pretrained encoders implies that the overall performance remains bounded by the representational capacity of the underlying vision-language backbone.

By introducing a lightweight mechanism for vision-language alignment, LFM enhances the controllability of LVLMs. The frequency-domain perspective enriches our understanding of modality interaction and offers a novel angle for optimizing multimodal representations. Looking ahead, frequency-domain techniques may offer a promising direction for further exploration in optimizing multimodal representations, potentially inspiring future research in this area.



What are the different components on the plate, and how are they arranged relative to each other?



The plate contains a sandwich, potato fries, and sliced pickles. The sandwich is cut into two halves and placed centrally on the plate. The sweet potato fries are positioned to the right of the sandwich, while the sliced pickles are located to the left. Additionally, there are two cups on the table, one close to the sandwich and another near the edge of the table. A wine glass is also present at the top left corner of the table. The arrangement creates a visually appealing presentation of the meal, with the different components easily accessible for the person enjoying the dish.

Figure 3. Qualitative results of the daily scenes understanding.



What are all the text on the card?

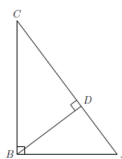


The text on the yellow card reads as follows:

- The authentic Absinth
- 55% and 10 mg Thujone p/kg
- 75% and 10 mg Thujone p/kg
- 80% and 30 mg Thujone p/kg

This text is likely to be a description of the alcohol content and the amount of Thujone, a compound found in absinthe, per kilogram of the beverage.

Figure 4. Qualitative results of OCR.



5. In the figure above, $AB = 6$ and $BC = 8$. What is the length of segment \overline{BD} ?

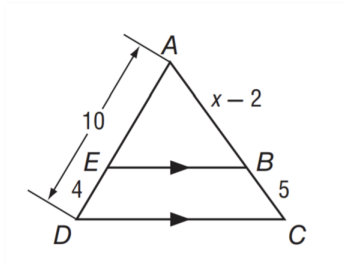
(A) 2
 (B) $\frac{12}{5}$
 (C) 4
 (D) $\frac{24}{5}$
 (E) 6



Answer the question in the figure and provide the options.



To find the length of segment BD , we can use the geometric mean theorem, which states that for right triangle ABC , the following relation holds: $AC / AB = BC / BD$
 Given $AB = 6$, $BC = 8$, and $AC = 10$ (calculated from the Pythagorean theorem), we substitute these values into the proportion: $10 / 6 = 8 / BD$
 Solving for BD , we get: $BD = 8 / 10 \times 6 = 24 / 5$
 Thus, the correct answer is (C).

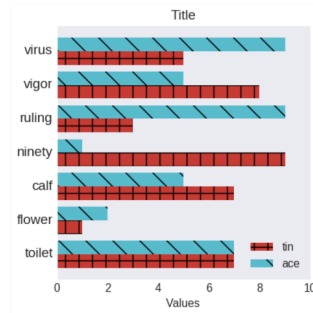


In the given triangle ABC , a smaller triangle ADE is formed with $EB \parallel DC$. What is the value of x if the triangles are similar?



Since triangle ADE and triangle ABC are similar, their corresponding sides are proportional. Using the property of similar triangles:
 $AD / AE = AC / BC$
 Substitute the given values: $10 / (10-4) = (AB + 5) / AB$
 Solving for AB , we get: $AB = 7.5$
 Given $AB = x - 2$
 Solve for x : $x = AB + 2 = 9.5$

Figure 5. Qualitative results of the math solving.



What does the bar chart represent?



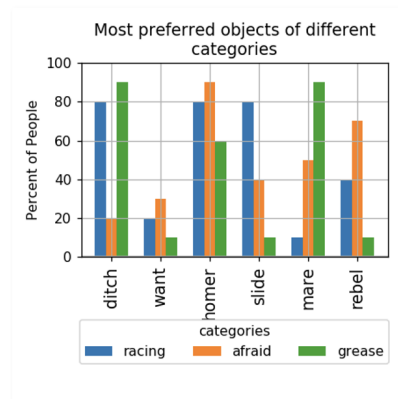
The bar chart shows a comparison of two categories, "tin" (red, grid-patterned) and "ace" (blue, diagonal-striped), across several labeled items (e.g., virus, vigor, ruling, etc.).

How do the categories "tin" and "ace" compare across the listed items?



For each item, the length of the bars indicates the values of "tin" and "ace." Generally, "ace" tends to have higher values compared to "tin" across most categories, with significant differences observed in items like "virus," "ruling," and "toilet." However, some items, such as "flower," show a smaller difference between the two categories.

Figure 6. Qualitative results of the bar chart understanding.



What is the difference in preference percentage between the "racing" and "afraid" categories for the "homer" object?



For the "homer" object, the "racing" category has a preference percentage of 80%, and the "afraid" category has a preference percentage of 90%. The difference in preference percentage between these two categories for the "homer" object is $90\% - 80\% = 10\%$.

Which category has the highest preference in the "mare" object according to the chart?



According to the chart, the "grease" category has the highest preference for the "mare" object, with a percentage close to 90%.

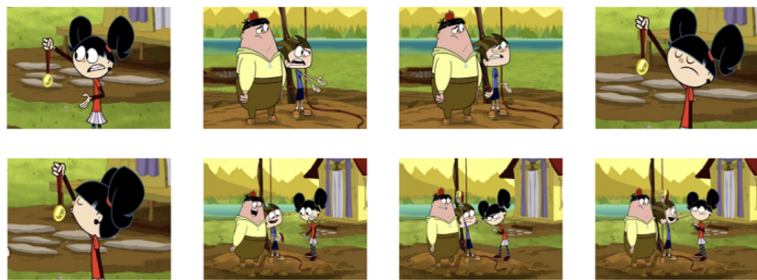
Figure 7. Qualitative results of the bar chart understanding.



What kind of activities or interactions are shown in the video?



The video depicts a soccer training session. The video shows a player dressed in black sportswear practicing on a soccer field. Initially, he is seen engaging in some basic ball control exercises. Later, he is joined by several teammates also dressed in black, suggesting a tactical drill. The presence of numbered markers on the field typically indicates a structured training exercise designed to guide players through specific positions and movements. Overall, the video illustrates various aspects of soccer training, including individual skills and team coordination exercises, aimed at improving the player's soccer abilities and tactical awareness.



What are the characters in the video discussing?



Based on the video content, the characters seem to be engaged in a discussion about a significant object, possibly the medal being held by one of them. The expressions and body language suggest a debate or exchange of opinions. The context appears to revolve around the medal's importance, ownership, or achievement, as indicated by the focused gestures and interactions among the characters.

Figure 8. Qualitative results of the video understanding.

References

- [1] Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2, 3
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4
- [4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 3, 4
- [5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 2, 3
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1, 3
- [7] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 3
- [8] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024. 3
- [9] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2
- [10] KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3
- [11] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 3
- [12] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. 2023. 3
- [13] Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H Li, and Ge Li. One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*, 2023. 3
- [14] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 3
- [15] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3
- [16] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 3
- [17] Yizhou Wang, Ruiyi Zhang, Haoliang Wang, Uttaran Bhattacharya, Yun Fu, and Gang Wu. Vaquita: Enhancing alignment in llm-assisted video understanding. *arXiv preprint arXiv:2312.02310*, 2023. 3
- [18] Wenhao Wu. Freeva: Offline mllm as training-free video assistant. *arXiv preprint arXiv:2405.07798*, 2024. 1, 3, 4
- [19] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. 3
- [20] Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. *arXiv preprint arXiv:2405.13800*, 2024. 3
- [21] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2, 3
- [22] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3
- [23] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 3
- [24] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1