A. Supplementary Experiments on the Generalizability Analysis of MLLM4WTAL

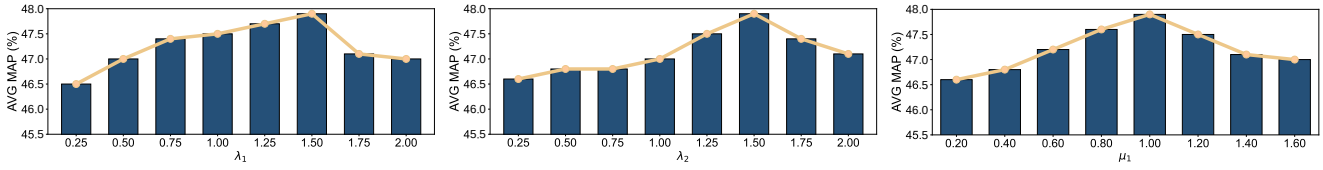
The main text employs the Video-LLaMA 7B[2] as the Multimodal Language Learning Model (MLLM) and DELU[1] as the WTAL Base Model. It is important to note that our MLLM4WTAL paradigm is adaptable to various MLLMs and WTAL Base Models. To further validate the generality of our approach, we conducted experiments using a stronger MLLM as indicated in table1. The results demonstrate that for DELU, upgrading to the more powerful Video-LLaMA 13B[2] enhances the accuracy of key and complete semantics, thus achieving significant performance improvements over Video-LLaMA 7B. This indicates that our approach can indeed benefit from stronger MLLM models. In the experiment conducted by Zhou et al.[3], we employed the more robust Baseline method. Using this approach, both Video-LLaMA 7B and Video-LLaMA 13B models outperformed the current state-of-the-art. This demonstrates that our approach and the advanced WTAL Base Model can deliver exceptional performance.

Baseline	MLLM	THUMOS14							AVG		
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.5	0.3:0.7	0.1:0.7
DELU[1]	-	71.5	66.2	56.5	47.7	40.5	27.2	15.3	56.48	37.44	46.40
	Video-LLaMA 7B	72.3	67.0	57.3	48.2	42.4	29.7	18.4	57.44	39.20	47.90
		+0.8	+0.8	+0.8	+0.5	+1.9	+2.5	+3.1	+0.96	+1.76	+1.50
	Video-LLaMA 13B	73.4	67.2	58.6	49.5	43.7	31.1	19.1	58.48	40.40	48.94
		+1.9	+1.0	+2.1	+1.8	+3.2	+3.9	+3.8	+2.00	+2.96	+2.54
Zhou et al.[3]	-	74.0	69.4	60.7	51.8	42.7	26.2	13.1	59.70	38.90	48.30
	Video-LLaMA 7B	74.3	69.8	61.8	52.3	43.0	30.8	16.6	60.24	40.90	49.80
		+0.3	+0.4	+1.1	+0.5	+0.3	+4.6	+3.5	+0.54	+2.00	+1.50
	Video-LLaMA 13B	75.1	71.6	63.1	54.8	44.9	31.6	18.1	61.90	42.50	51.31
		+1.1	+2.2	+2.4	+3.0	+2.2	+5.4	+5.0	+2.20	+3.60	+3.01

Table 1. Ablation Studies of Different MLLMs

B. hyperparametric Analysis

We perform ablation experiments to investigate the effects of the hyperparameters λ_1,λ_2 and μ_1 in the equilibrium loss L_{KSM} and L_{CSR}, respectively. The results of these experiments are summarized in the figure above. Various values are tested to determine the optimal weights for each hyperparameter. Based on the experimental results, we empirically set λ_1 = 1.50,λ_2 = 1.50 and μ_1=1.00 as they yield the best performance.



(a) λ_1 (b) λ_2 (c) μ_1

Figure 1. Hyperparameter Sensitivity Analysis Line Chart.

C. More visualizations

References

[1] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European conference on computer vision*, pages 192–208. Springer, 2022. 1

[2] 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. 1

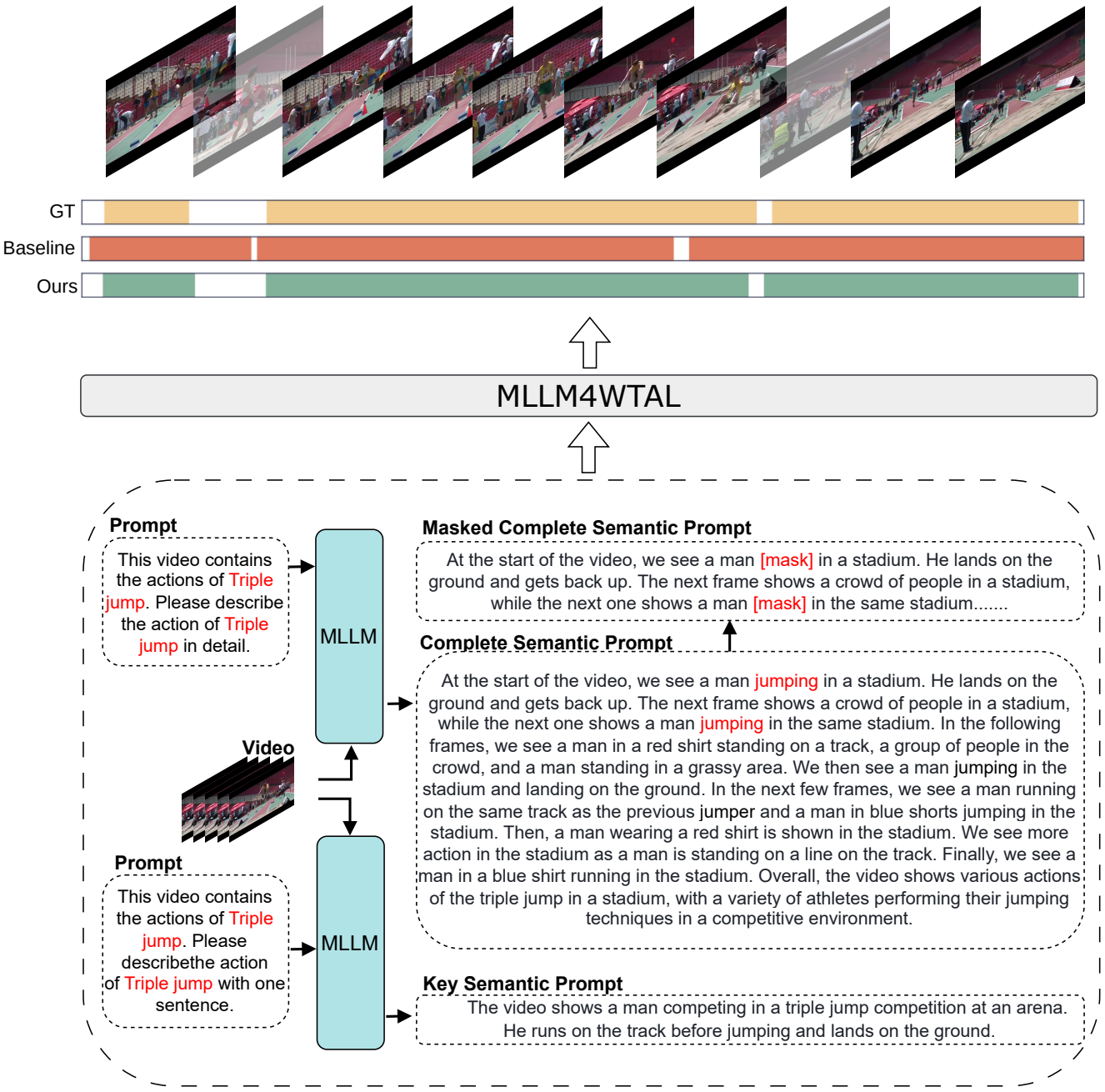


Figure 2. The MLLM takes key semantic prompt $\mathcal{P}_{key}$ to generate key semantic description $\mathcal{D}_{key}$, which is then fed into a semantic matching module to match key semantics and locate the critical intervals of temporal actions. The MLLM also takes complete semantic prompt $\mathcal{P}_{complete}$ to produce complete semantic description $\mathcal{D}_{complete}$, which is then input into a semantic reconstruction module to achieve the reconstruction of complete semantics, identifying the complete intervals of temporal actions. The two branches engage in interactive distillation, effectively solving the common issues of over-completeness and in-completeness seen in previous methods.

[3] Jingqiu Zhou, Linjiang Huang, Liang Wang, Si Liu, and Hongsheng Li. Improving weakly supervised temporal action localization by bridging train-test gap in pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23003–23012, 2023. 1

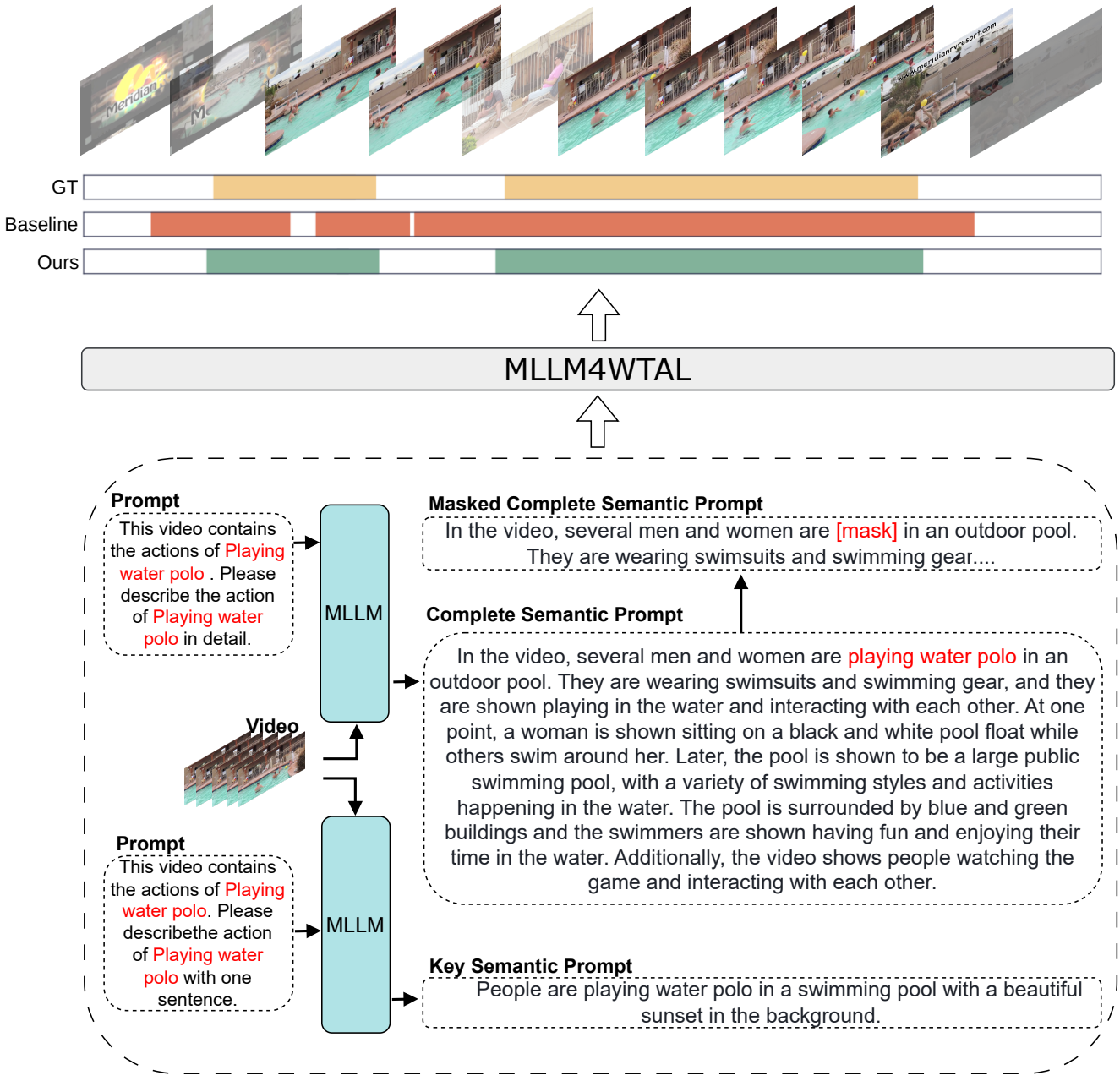


Figure 3. The MLLM takes key semantic prompt $\mathcal{P}_{key}$ to generate key semantic description $\mathcal{D}_{key}$, which is then fed into a semantic matching module to match key semantics and locate the critical intervals of temporal actions. The MLLM also takes complete semantic prompt $\mathcal{P}_{complete}$ to produce complete semantic description $\mathcal{D}_{complete}$, which is then input into a semantic reconstruction module to achieve the reconstruction of complete semantics, identifying the complete intervals of temporal actions. The two branches engage in interactive distillation, effectively solving the common issues of over-completeness and in-completeness seen in previous methods.

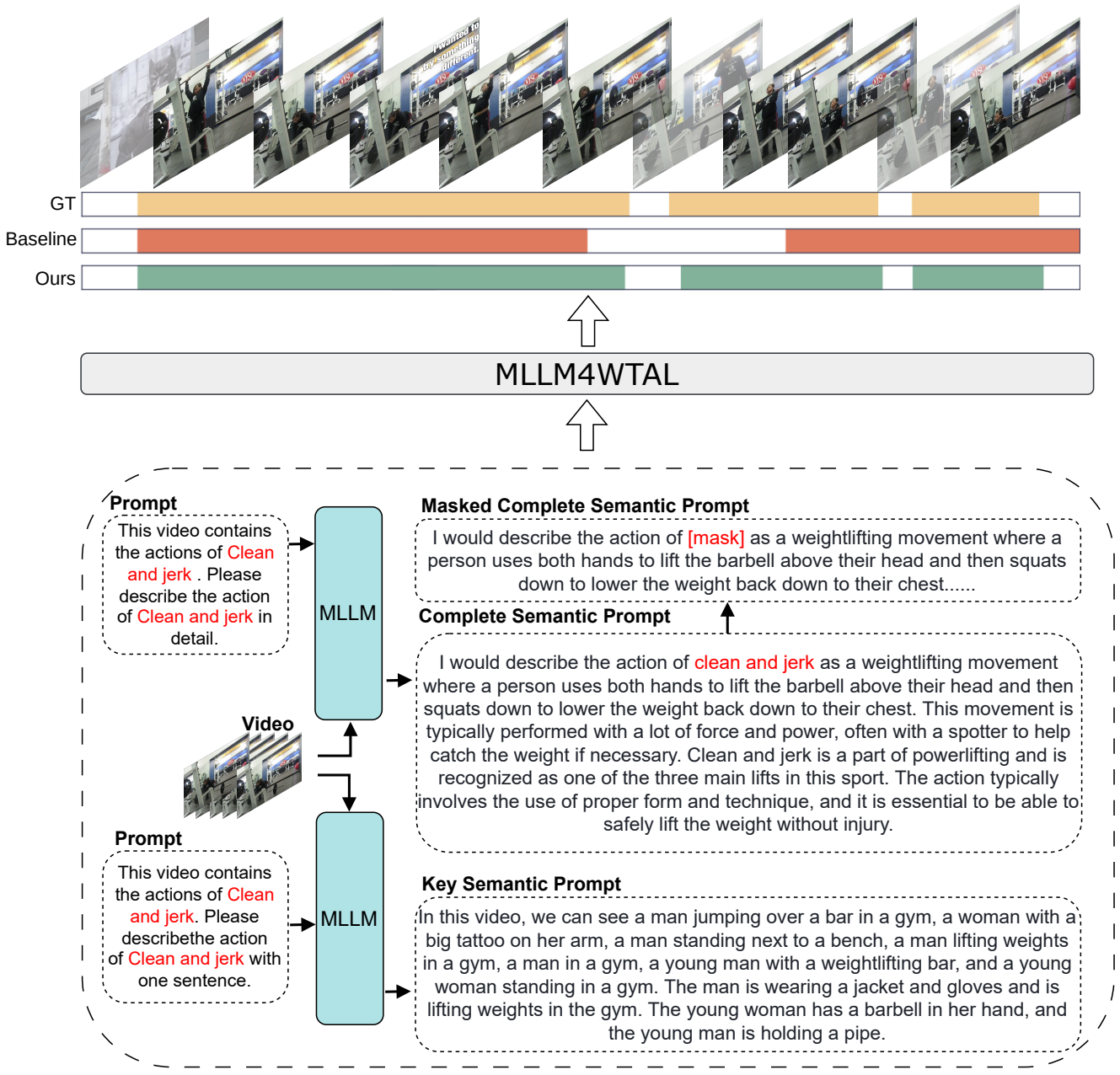


Figure 4. The MLLM takes key semantic prompt $\mathcal{P}_{key}$ to generate key semantic description $\mathcal{D}_{key}$, which is then fed into a semantic matching module to match key semantics and locate the critical intervals of temporal actions. The MLLM also takes complete semantic prompt $\mathcal{P}_{complete}$ to produce complete semantic description $\mathcal{D}_{complete}$, which is then input into a semantic reconstruction module to achieve the reconstruction of complete semantics, identifying the complete intervals of temporal actions. The two branches engage in interactive distillation, effectively solving the common issues of over-completeness and in-completeness seen in previous methods.

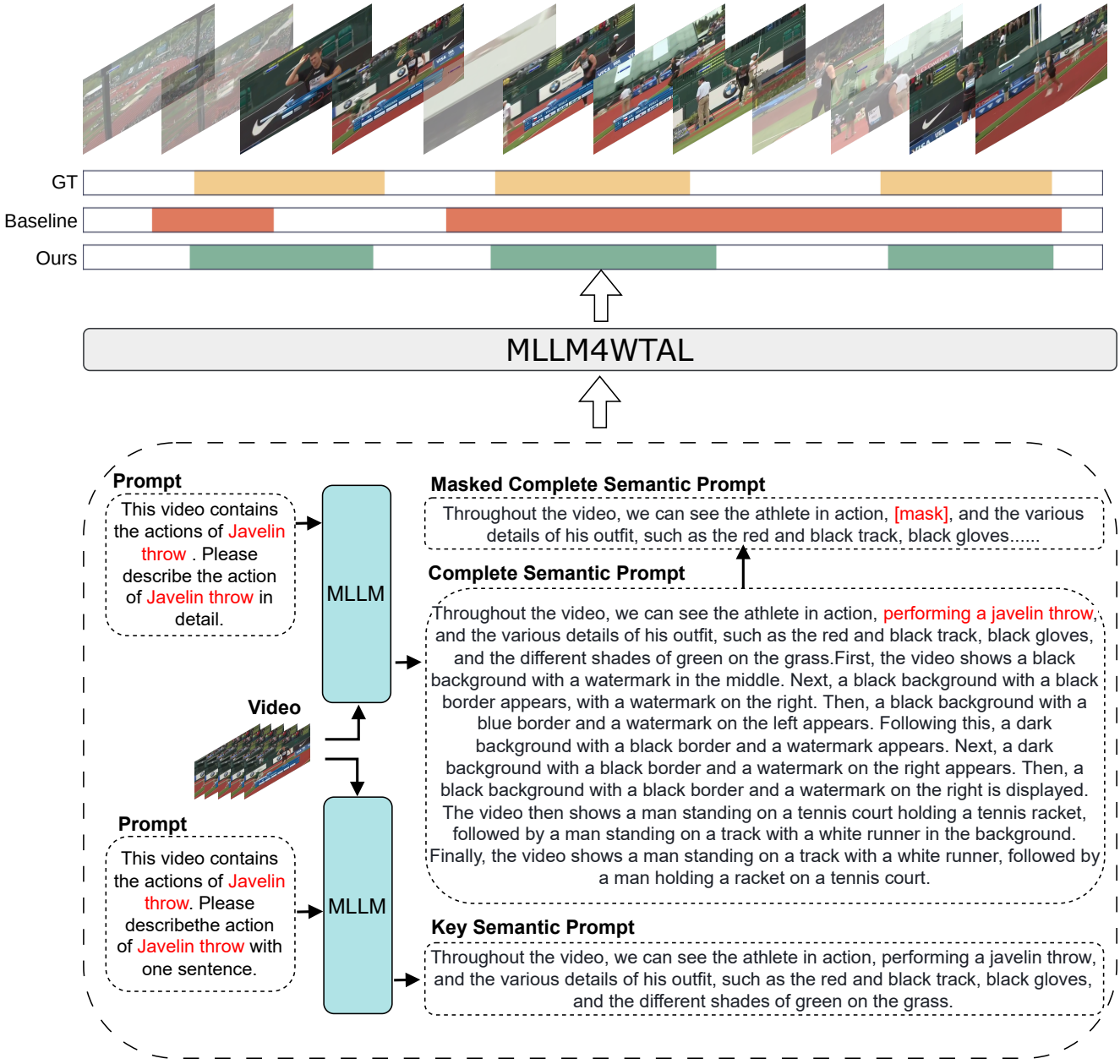


Figure 5. The MLLM takes key semantic prompt $\mathcal{P}_{key}$ to generate key semantic description $\mathcal{D}_{key}$, which is then fed into a semantic matching module to match key semantics and locate the critical intervals of temporal actions. The MLLM also takes complete semantic prompt $\mathcal{P}_{complete}$ to produce complete semantic description $\mathcal{D}_{complete}$, which is then input into a semantic reconstruction module to achieve the reconstruction of complete semantics, identifying the complete intervals of temporal actions. The two branches engage in interactive distillation, effectively solving the common issues of over-completeness and in-completeness seen in previous methods.

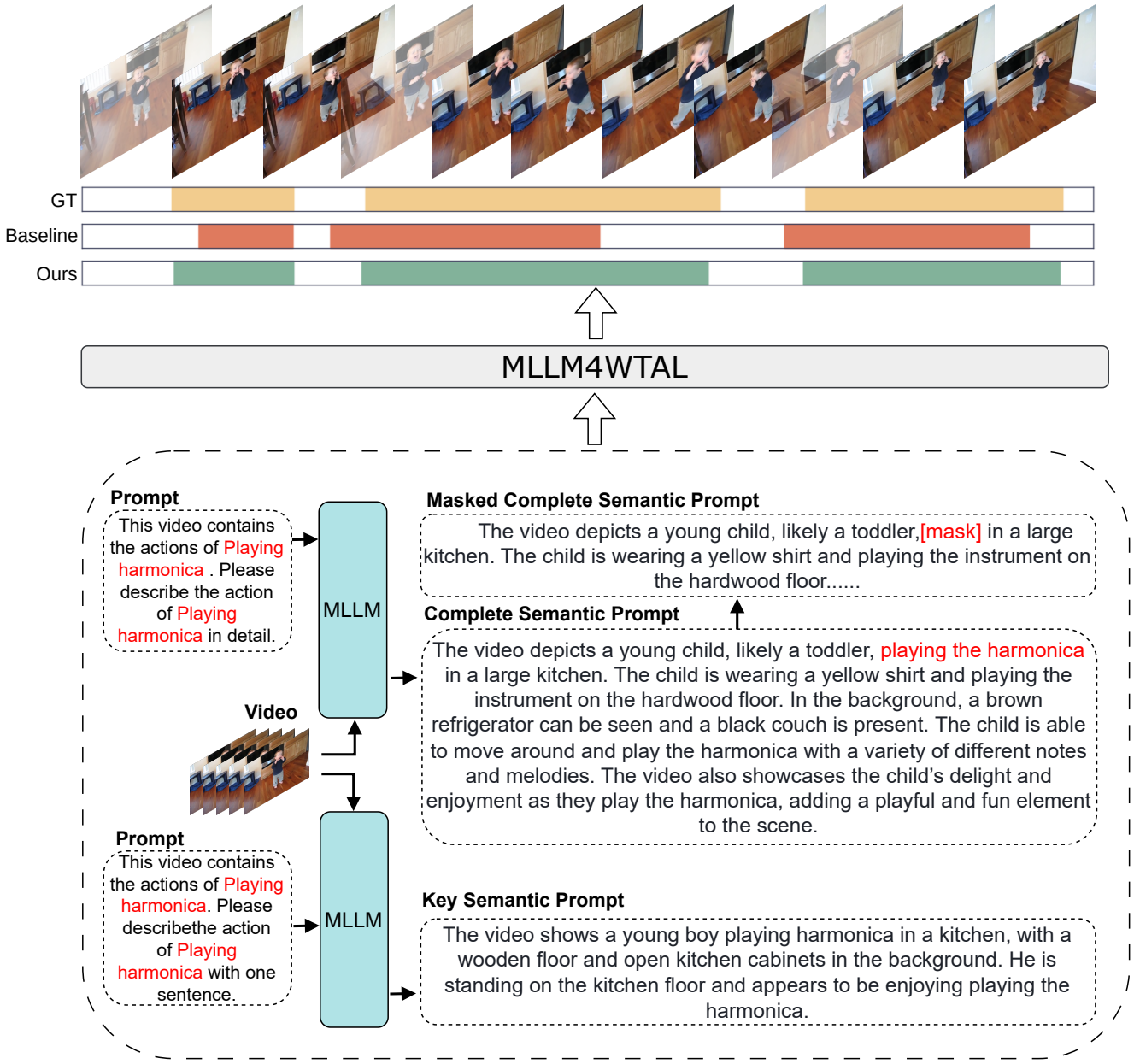


Figure 6. The MLLM takes key semantic prompt $\mathcal{P}_{key}$ to generate key semantic description $\mathcal{D}_{key}$, which is then fed into a semantic matching module to match key semantics and locate the critical intervals of temporal actions. The MLLM also takes complete semantic prompt $\mathcal{P}_{complete}$ to produce complete semantic description $\mathcal{D}_{complete}$, which is then input into a semantic reconstruction module to achieve the reconstruction of complete semantics, identifying the complete intervals of temporal actions. The two branches engage in interactive distillation, effectively solving the common issues of over-completeness and in-completeness seen in previous methods.

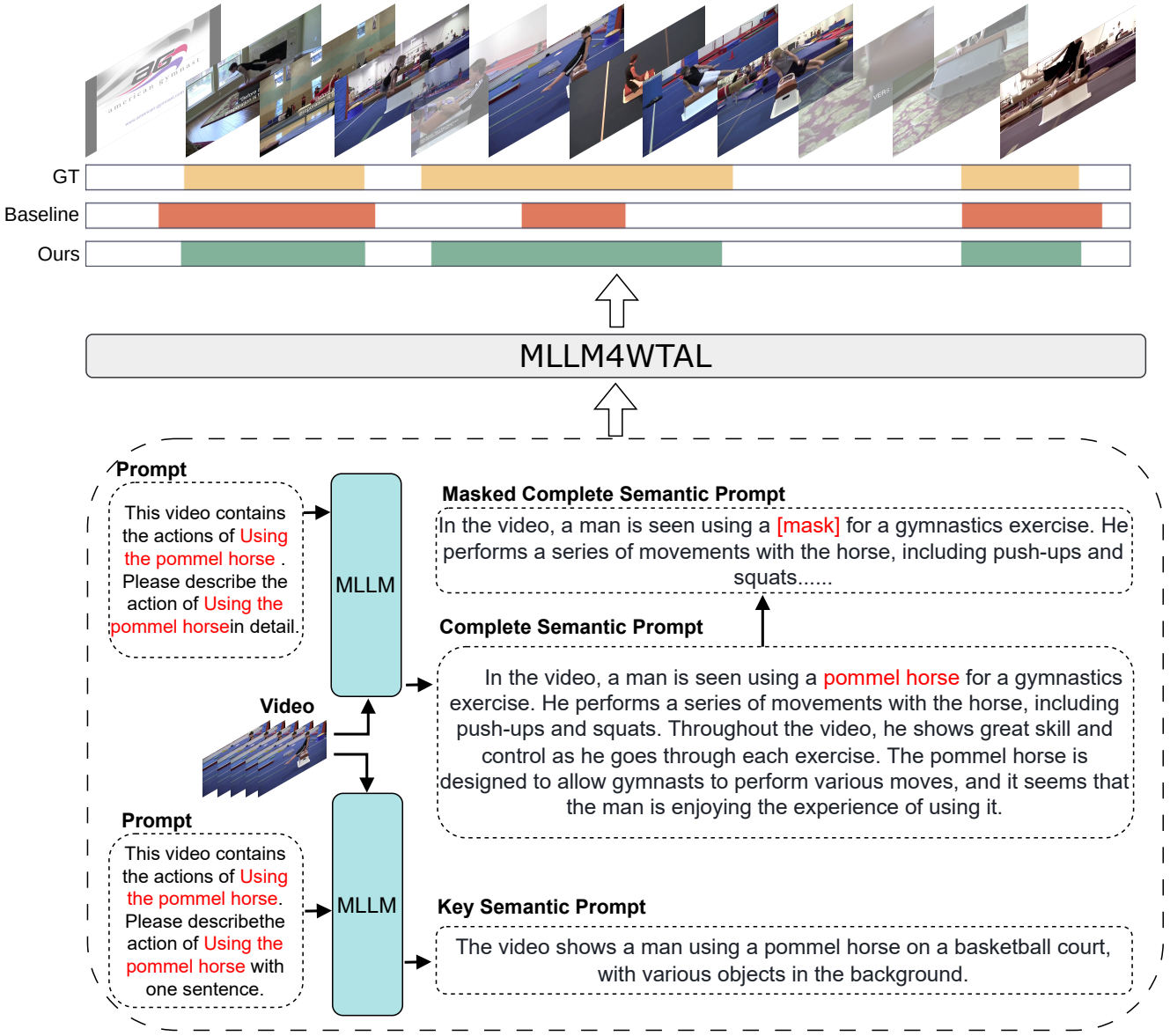


Figure 7. The MLLM takes key semantic prompt $\mathcal{P}_{key}$ to generate key semantic description $\mathcal{D}_{key}$, which is then fed into a semantic matching module to match key semantics and locate the critical intervals of temporal actions. The MLLM also takes complete semantic prompt $\mathcal{P}_{complete}$ to produce complete semantic description $\mathcal{D}_{complete}$, which is then input into a semantic reconstruction module to achieve the reconstruction of complete semantics, identifying the complete intervals of temporal actions. The two branches engage in interactive distillation, effectively solving the common issues of over-completeness and in-completeness seen in previous methods.