Vision-aided Unsupervised Constituency Parsing with Multi-MLLM Debating

Dong Zhang[1], **Haiyan Tian**[1], **Qingying Sun**[2], **Shoushan Li**[1*]
[1] School of Computer Science & Technology, NLP Lab, Soochow University, China
[2] School of Computer Science and Technology, Huaiyin Normal University, Huaian, China
dzhang@suda.edu.cn

Abstract

This paper presents a novel framework for vision-aided unsupervised constituency parsing (VUCP), leveraging multimodal large language models (MLLMs) pre-trained on diverse image-text or video-text data. Unlike previous methods requiring explicit cross-modal alignment, our approach eliminates this need by using pre-trained models like Qwen-VL and VideoLLaVA, which seamlessly handle multimodal inputs. We introduce two multi-agent debating mechanisms–consensus-driven (CD) and round-driven (RD)–to enable cooperation between models with complementary strengths. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on both image-text and video-text datasets for VUCP, improving robustness and accuracy.

1 Introduction

Unsupervised constituency parsing (UCP), which extracts syntactic structures from unannotated text, has long relied on textual context and semantic cues for sequence labeling (Yang and Tu, 2022, 2023). By assigning syntactic tags to tokens, these models aim to uncover hierarchical structures within sentences (Gu et al., 2022; Tseng et al., 2023). However, recent advances in natural language processing (NLP) have revealed the limitations of a purely textual approach, especially in domains like social media or instructional content, where visual information such as images or videos often appears. Studies have explored how incorporating visual information can improve constituency parsing accuracy, highlighting that visual context provides crucial cues absent in text alone, thereby aiding the resolution of syntactic ambiguities. This has led to the development of vision-aided unsupervised constituency parsing (VUCP) (Zhao and Titov, 2020).

However, leveraging multimodal data for unsupervised parsing presents several challenges. Two key issues persist: First, prior work (Zhang et al., 2021, 2022) necessitates explicit alignment between textual and visual data during training, requiring labor-intensive image-text or video-text correspondences, as shown in Figure 1 (a). This approach limits scalability due to the need for large, curated, and synchronized datasets. By using multimodal large language models (MLLMs) pre-trained on diverse cross-modal datasets, we eliminate the need for additional alignment, offering significant resource savings for VUCP. Second, existing methods (Shayegh et al., 2024c,a; Hou and Li, 2024; Zhang et al., 2025) often focus on optimizing a single model, overlooking the complementary strengths of multiple models (Li et al., 2023a, 2024). For example, Model A may struggle with adjective phrase identification but perform well on noun phrases, while Model B excels at the reverse. By combining their strengths (Tran et al., 2025) as shown in Figure 1 (b), we believe that VUCP could be significantly improved.

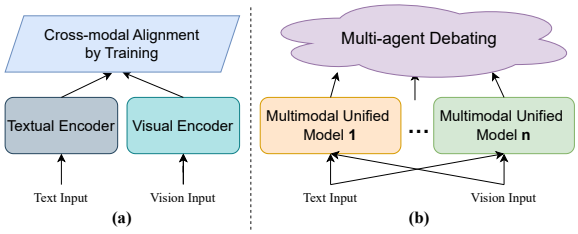

Figure 1: (a) Previous studies need image-text or video-text aligned training with single model. (b) Our approach employs multi-modal aligned large language model (MLLMs) with multiple MLLMs' cooperation.

To address these challenges, we propose a novel multi-MLLM debating (MMD) framework for VUCP. The MLLMs we use, such as Qwen-VL (Bai et al., 2023) and VideoLLaVA (Lin et al., 2023), have been pre-trained on vast image-text pairs and can naturally process both image-text and video-text inputs, bypassing the need for ex-

*Corresponding Author: Shoushan Li

Findings of the Association for Computational Linguistics: ACL 2025, pages 6800–6810

pensive cross-modal alignment. Furthermore, we introduce two multi-agent debating mechanisms within our framework, where models, each with different strengths, collaborate in either a consensus-driven (CD) mode or round-driven (RD) mode. In CD mode, the judge checks for consensus at each round, while in RD mode, a judge selects the best answer among the outputs of both models across all rounds. These debating modes facilitate collaborative decision-making, improving accuracy and robustness. The key contributions of this work are:

• We propose a framework that eliminates the need for additional cross-modal alignment, leveraging pre-trained MLLMs to process both image-text and video-text inputs for VUCP.

• We introduce two multi-agent debating mechanisms that combine the strengths of different models, enhancing VUCP's accuracy and robustness.

• Through extensive experiments and detailed analysis, we demonstrate that our MMD approach achieves state-of-the-art performance on both image-text and video-text datasets for VUCP.

2 Related Work

Unsupervised Constituency Parsing. Supervised constituency parsing (CP) is limited by the high cost and time required for dataset annotation (Cui et al., 2022). Unsupervised CP addresses this by leveraging large unannotated corpora (Cao et al., 2020; Shayegh et al., 2024b). Liu et al. (2023) introduce a simple probabilistic context-free grammar (PCFG) with independent left and right productions for unsupervised parsing. In vision-aided scenarios, Zhao and Titov (2020) propose image-aware unsupervised CP through text-image matching, while Zhang et al. (2022) align text spans with video using pre-trained PCFGs for video-aided parsing. However, these methods focus on single-model optimization and ignore the potential of other homologous models for unsupervised constituency parsing (VUCP).

In contrast, our work utilizes two unsupervised debate mechanisms, involving interactions among multiple homologous multimodal large language models (MLLMs) to enhance VUCP performance.

Multi-agent Collaboration. A limited number of multi-agent approaches have been applied to NLP tasks (Du et al., 2024; Liang et al., 2024), and none to VUCP. Existing methods largely focus on large-scale supervised learning with multiple agents (Estornell and Liu, 2024; Wang et al., 2024b) and knowledge extraction (Liu et al., 2024b). Our approach introduces a novel multi-MLLM debate framework with two argument selection schemas, representing the first such attempt in VUCP.

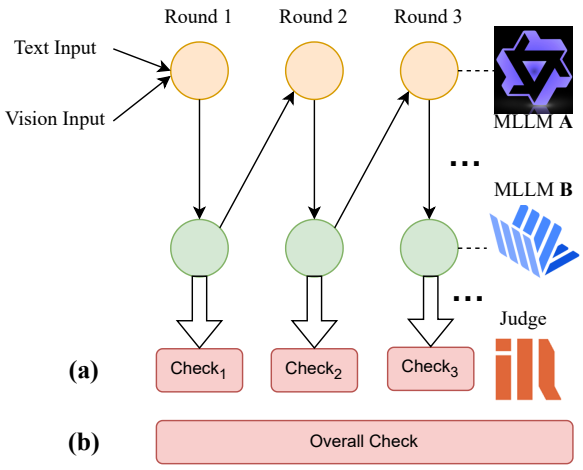

Figure 2: The overall architecture of our MMD framework with two kinds of argument selection schemes: (a) consensus-driven scheme checks the consensus of both outputs at each round. (b) round-driven scheme selects the best argument among the outputs.

3 Methodology

Figure 2 illustrates the simplified framework of our approach.

Task Definition: In the context of large language models (LLMs), constituency parsing is redefined as sequence generation tasks. Given an input sentence $X = \{X_1, X_2, \ldots, X_n\}$ nd a pre-trained model $\mathcal{M}$, the sentence structure is analyzed and predicted as follows:

$$Y = \mathcal{M}(X) \sim P_{\mathcal{M}}(Y|X) = \prod_{i=1}^{V} P_{\mathcal{M}}(y_i \mid y_{<i}; X) \quad (1)$$

where V denotes the vocabulary set.

3.1 Warm-Start of MLLMs for VUCP

LLMs often struggle with constituency parsing tasks due to their limited expertise in accurately predicting constituent spans. To address this, we adopt a bracket-based serialization structure, which avoids issues caused by missing symbols and provides a more intuitive format, e.g.,

(I (am ((a big fan) (of (american football)))))

To warm up the backbone MLLMs for constituency parsing (CP), we fine-tune them using a subset of general, widely-used corpora (Marcus

Dataset	NP	VP	PP	SBAR	ADJP	ADVP
Coco-val	16936	6170	8580	484	469	350
Coco-test	16930	5949	8668	453	449	364
YouCook2	7452	2471	3231	99	65	20

Table 1: The number statistics of each phrase on different datasets.

et al., 1993) from diverse domains (unseen in test sets), optimizing the following loss function:

$$\text{Loss} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} y_{t,i}\log(\hat{y}_{t,i}) \tag{2}$$

where T is the sequence length, N is the vocabulary size, $y_{t,i}$ is the true label of the i-th word at the t-th time step (one-hot encoded vector), and $\hat{y}_{t,i}$ is the predicted probability of the i-th word at the t-th time step.

3.2 Multi-MLLM Debating Mechanism

Without losing its generality, we use two MLLMs to design our approach, as shown in Figure 2. After the above warm up, we obtain the affirmative debater $\mathcal{L}$ and negative debater $\mathcal{V}$, we leverage multi-round interactions between the models to encourage dialectical thinking on constituency parsing (CP) results. To create a debating atmosphere for them, we first set up system instructions then:

For the **affirmative party** $\mathcal{L}$, it generates what it believes to be the correct CP result on the input sentence X, denoted as:

$$\begin{aligned} \mathit{Aff}_1 &= \mathcal{L}.ask(X) \\ &= \arg\max p(Y_{aff_1}|Prompt, X; \theta_{\mathcal{L}}) \end{aligned} \tag{3}$$

where the function $.ask()$ represents the probabilistic sequence generation by auto-regressive schema $p(w_i|w_{<i})$ parameterized on $\theta_{\mathcal{L}}$ of the tuned model $\mathcal{L}$. In the first round, the prompt is simply given: *Please perform constituency parsing on the following sentence:* X.

For **negative party**, we feed the above primary result *Aff*, along with the debating prompt, into $\mathcal{V}$. It evaluates the potential errors on *Aff*, reasons through them, raises objections, and proposes its result:

$$\begin{aligned} \mathit{Neg}_1 &= \mathcal{V}.ask(\mathit{Aff}_1) \\ &= \arg\max p(Y_{neg_1}|Prompt, \mathit{Aff}; \theta_{\mathcal{V}}) \end{aligned} \tag{4}$$

where $Prompt$ denotes the instruction to make each debater express their arguments based on the previous debate history, which can refer to appendix.

Upon providing the instructions outlined above, the model generates the parsing result it deems correct, which is presented after the final token: *Your_result*.

For the **judge party** $\mathcal{J}$, its role is to oversee and regulate the debate process through two primary mechanisms: 1) As depicted in Figure 2 (a), it determines whether the outputs of the two models are consistent in the current round by invoking the function: $\mathcal{J}.ask(\mathit{Aff}_1, \mathit{Neg}_1)$. If the result is True, the debate concludes; otherwise, it proceeds to the next round. 2) As illustrated in Figure 2 (b), at the conclusion of the maximum number of iterations, it selects the optimal argument among all generated results: $\mathcal{J}.ask(\mathit{Aff}_1, \mathit{Neg}_1, \ldots, \mathit{Aff}_n, \mathit{Neg}_n)$, where n denotes the total number of rounds.

4 Experimentation

For a thorough evaluation of our MMD, we conduct extensive experiments and detailed analysis.

4.1 Experimental Setting

Datasets. For the image-text scenario, the validation and test sets from **MSCOCO (Lin et al., 2014)** are selected following (Zhao and Titov, 2020). For the video-text scenario, the test set from **YouCook2** (no val) are selected following (Zhang et al., 2021, 2022). The phrase number of each label on all datasets can refer to Table 1.

For warm up, we randomly select 450 different samples from the PTB (Marcus et al., 1993) training set, which never appear in the above testing data.

Evaluation Metrics. Our study adheres to the evaluation frameworks established by (Zhang et al., 2021) and (Zhao and Titov, 2020). During the testing phase, we exclude single-word and sentence-level spans from consideration and instead focus on assessing two key metrics: the corpus-level average F1 (C-F1) and the sentence-level average F1 (S-F1). The calculation of C-F1 is given by the

COCO-Val			COCO-Test			YouCook2		
Model	**C-F1**	**S-F1**	**Model**	**C-F1**	**S-F1**	**Model**	**C-F1**	**S-F1**
VC-PCFG	58.34	58.26	VC-PCFG	59.30	59.40	PTC-PCFG	58.90	63.20
GPT-4o-mini	74.30	74.91	GPT-4o-mini	73.85	74.66	GPT-4o-mini	69.91	70.72
BLIVA	76.26	75.45	BLIVA	76.54	75.82	Qwen2-VL	78.21	77.12
LLM-enhanced ST	86.80	86.22	LLM-enhanced ST	86.48	85.99	LLM-enhanced ST	82.41	82.34
LLaVA-1.5	76.72	71.36	LLaVA-1.5	77.25	72.13	VideoLLaVA	75.34	77.20
Qwen-VL	74.30	73.55	Qwen-VL	74.62	73.93	InternVL2	82.53	82.49
MMD (CD)	85.62	86.26	MMD (CD)	86.29	86.30	MMD (CD)	82.59	83.03
MMD (RD)	**88.01**	**88.20**	MMD (RD)	**87.89**	**88.08**	MMD (RD)	**83.37**	**84.20**

Table 2: Performance comparison of different models on COCO-Val, COCO-Test, and YouCook2 datasets.

formula:

$$C\text{-}F1 = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{5}$$

Here, TP denotes the total number of constituents that are correctly predicted, FP represents the total number of constituents predicted incorrectly, and FN signifies the total number of correct constituents that are overlooked.

For S-F1, its formula is expressed as:

$$S\text{-}F1 = \frac{1}{n}\sum_{i=1}^{n} F1_i \tag{6}$$

where n refers to the total number of sentences, and $F1_i$ is the F1 score corresponding to the i-th sentence.

Baselines. 1) **VC-PCFG** (Zhao and Titov, 2020): SOTA for image-text scenario. 2) **GPT-4o-mini**: outperforming GPT-3.5 Turbo while reducing costs by over 60%. 3) **BLIVA** (Hu et al., 2024): SOTA on various vision-language tasks. 4) **LLM-enhanced ST** (Li et al., 2023b): SOTA on cross-domain constituency parsing using self-training with LLMs. 5) **LLaVA-1.5** (Liu et al., 2024a): SOTA across 11 benchmarks with just 1.2M publicly available data. 6) **Qwen-VL** (Bai et al., 2023): SOTA in various real-world dialog tasks with image-text context. 7) **PTC-PCFG** (Zhang et al., 2022): SOTA for video-text scenario. 8) **Qwen2-VL** (Wang et al., 2024a): competitive performance on various video-language tasks. 9) **VideoLLaVA** (Lin et al., 2023): outperforming Video-ChatGPT and achieveing superior results across multiple video-text benchmarks. 10) **InternVL2** (Chen et al., 2024): SOTA on various video-language tasks.

Implementation Details. Due to the resource constraint, we employ the 7B or 8B version for MLLMs. Specifically, for the text + image dataset, we use LoRA fine-tuning with llava-v1.5-7b and Qwen-VL-Chat as *aff* and *neg*. For the text + video dataset, we use LoRA fine-tuning with VideoLLaVA-7B and InternVL2-8B as *aff* and *neg*. For judge, we leverage the Baichuan2-13B (Baichuan, 2023) to unify the processing for both scenarios.

4.2 Main Experimental Results

We compare several state-of-the-art baselines with our proposed Multi-MLLM Debate (MMD) approach across three datasets, as shown in Table 2. The key findings are as follows:

Both closed-source models (e.g., GPT-4o-mini with 35B parameters) and open-source models (e.g., BLIVA and Qwen2-VL) significantly outperform traditional state-of-the-art approaches for Visual-Universal Constituency Parsing (VUCP) without the use of large language models (LLMs), such as VC-PCFG and PTC-PCFG. This highlights the effectiveness of LLMs for UCP tasks, motivating our adoption of LLMs for parsing. Our MMD with CD (Constraint Debating) performs worse than MMD with RD (Relaxed Debating) and slightly lags behind the best-performing baseline LLM-enhanced ST. This may be attributed to the overly stringent consistency constraints of the CD approach, underscoring the importance of selecting an appropriate debating strategy for VUCP. Overall, MMD (RD) outperforms all other methods, demonstrating the efficacy of our proposed multi-MLLM debating framework.

4.3 Case Study

Figure 3 illustrates an example of constituency parsing for the sentence "A girl and a woman holding umbrellas". In the first round, both models produce errors or incomplete results. However, following the second round of debate, Qwen successfully

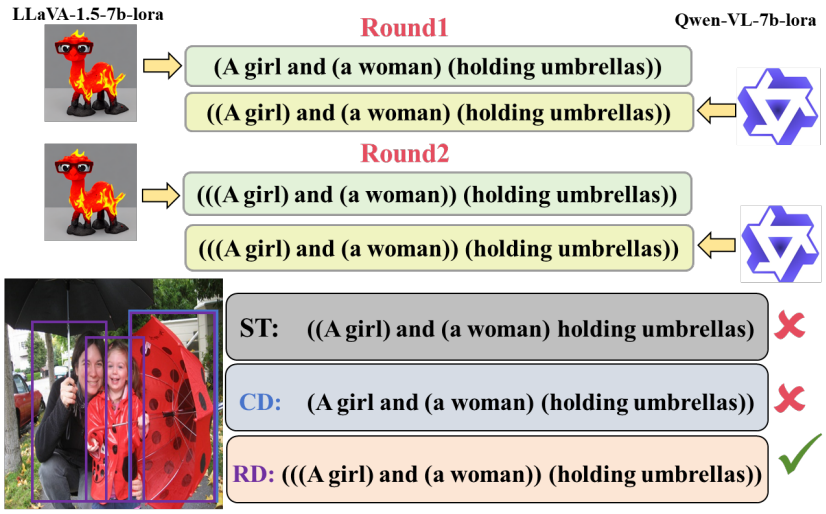

Figure 3: A real example of two-round process in our MMD and UCP results comparison of the best-performed baseline (LLM-enhanced ST) and our MMD with consensus-driven (CD) and round-driven (RD) modes.

generates the correct parsing. This highlights the potential of our approach to iteratively refine predictions through collaborative model interactions. Additionally, the top-performing baseline, ST, fails to correctly parse the final constituent due to its inability to access visual modality. Similarly, MMD (CD) does not fully label "A girl", potentially due to incorrect detection of the number of girls. These cases reveal areas for improvement in both existing models and our own approach, emphasizing the need for careful design and further refinement to yield the best MMD (RD).

4.4 Discussion

Our in-depth investigations show that consensus is typically achieved after 2-3 rounds of discussion. Beyond this, additional discussions tend to diminish performance, as observed in Figures 4 and 5 of the appendix. This is primarily due to the model generating more debate content with longer contextual input, which may induce hallucinations.

In practice, the occurrence of reaching consensus is much fewer than that of forced termination. Similar to real-life debate competitions, consensus is often hard to attain, and termination via time limits or a judging party is common. While consensus-reaching is an expected scenario, a perfect solution remains elusive, which will be a focus of our future exploration.

Currently, no correlation between debate length and input structural properties has been found. We will further investigate this aspect in future research.

5 Conclusion

We present MMD, a novel framework advancing vision-aided unsupervised constituency parsing through pre-trained MLLMs and multi-agent debates. By eliminating cross-modal alignment costs and synergizing model strengths via RD/CD protocols, MMD achieves SOTA performance (about 10% F1 gains vs. single MLLM) on image/video-text benchmarks. The debate mechanisms effectively resolve syntactic ambiguities by integrating complementary predictions, while pre-trained MLLMs ensure scalable multimodal processing. This work demonstrates the viability of resource-efficient, collaboration-driven parsing in complex multi-modal contexts, offering new directions for unsupervised syntactic analysis.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China grants (NSFC No. 62206193 and No.62006093), and the Humanities and Social Sciences Planning Foundation from the Ministry of Education of China(Grant No.24YJAZH113).

7 Limitations

While MMD demonstrates strong performance, it inherits constraints from pretrained MLLMs, such as biases in training data and limited generalization to low-resource languages. The debate mechanisms, though effective, increase inference time

compared to single-model approaches. Additionally, performance may degrade in domains with sparse visual-textual correlations. Future work should address these issues by enhancing model robustness, optimizing computational efficiency, and exploring lightweight alignment strategies for broader applicability.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Unsupervised parsing via constituency tests. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4798–4808.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. 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.

Leyang Cui, Sen Yang, and Yue Zhang. 2022. Investigating non-local features for neural constituency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2065–2075.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *ICML 2024*. OpenReview.net.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. In *NIPS 2024*.

Xiaotao Gu, Yikang Shen, Jiaming Shen, Jingbo Shang, and Jiawei Han. 2022. Phrase-aware unsupervised constituency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6406–6415. Association for Computational Linguistics.

Yang Hou and Zhenghua Li. 2024. Character-level chinese dependency parsing via modeling latent intra-word structure. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2943–2956. Association for Computational Linguistics.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Florence, Italy. Association for Computational Linguistics.

Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023a. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192.

Jianling Li, Meishan Zhang, Peiming Guo, Min Zhang, and Yue Zhang. 2023b. Llm-enhanced self-training for cross-domain constituency parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8174–8185.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 7281–7294. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Liang Liu, Dong Zhang, Shoushan Li, Guodong Zhou, and Erik Cambria. 2024b. Two heads are better than one: Zero-shot cognitive reasoning via multi-llm knowledge fusion. In *CIKM 2024*, pages 1462–1472.

Wei Liu, Songlin Yang, Yoon Kim, and Kewei Tu. 2023. Simple hardware-efficient pcfgs with independent left and right productions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1662–1669.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Behzad Shayegh, Yanshuai Cao, Xiaodan Zhu, Jackie C. K. Cheung, and Lili Mou. 2024a. Ensemble distillation for unsupervised constituency parsing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Behzad Shayegh, Yanshuai Cao, Xiaodan Zhu, Jackie CK Cheung, and Lili Mou. 2024b. Ensemble distillation for unsupervised constituency parsing. In *Proceedings of the 12th International Conference on Learning Representations*, pages 1–18.

Behzad Shayegh, Yuqiao Wen, and Lili Mou. 2024c. Tree-averaging algorithms for ensemble-based unsupervised discontinuous constituency parsing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15135–15156. Association for Computational Linguistics.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *Preprint*, arXiv:2501.06322.

Yuan Tseng, Cheng-I Jeff Lai, and Hung-Yi Lee. 2023. Cascading and direct approaches to unsupervised constituency parsing on spoken sentences. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024b. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *ACL 2024*, pages 6106–6131.

Songlin Yang and Kewei Tu. 2022. Bottom-up constituency parsing and nested named entity recognition with pointer networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2403–2416. Association for Computational Linguistics.

Songlin Yang and Kewei Tu. 2023. Don't parse, choose spans! continuous and discontinuous constituency parsing via autoregressive span selection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8420–8433. Association for Computational Linguistics.

Songyang Zhang, Linfeng Song, Lifeng Jin, Haitao Mi, Kun Xu, Dong Yu, and Jiebo Luo. 2022. Learning a grammar inducer from massive uncurated instructional videos. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 233–247.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021. Video-aided unsupervised grammar induction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1513–1524.

Ziyan Zhang, Yang Hou, Chen Gong, and Zhenghua Li. 2025. Data augmentation for cross-domain parsing via lightweight LLM generation and tree hybridization. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 11235–11247. Association for Computational Linguistics.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound pcfgs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4369–4379.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7590–7598.

A Details of Our Approach

For *negative party*, we feed the above primary result *Aff*, along with the debating prompt, into $\mathcal{V}$. It evaluates the potential errors on *Aff*, reasons through them, raises objections, and proposes its result:

$$\begin{aligned} Neg_1 &= \mathcal{V}.ask(Aff_1) \\ &= \arg\max p(Y_{neg_1}|Prompt, Aff; \theta_{\mathcal{V}}) \end{aligned} \tag{7}$$

where $Prompt$ denotes the instruction to make each debater express their arguments based on the previous debate history:

> $\cdots$ Please evaluate the potential errors based on prior result of constituent parsing and provide the answer you think is right. The output must be a nested bracketing structure without any extra content, containing all the words of the sentence, and the number of "(" must match the number of ")". Here are some examples for your reference and learning. Please format the output according to the examples.
> Sentence: "The children ate the cake with a spoon"
> Prior_result: (The (children ((ate (the cake)) (with (a spoon)))))
> Your_result: ((The children) ((ate (the cake)) (with (a spoon))))
> $\cdots$
> Sentence: "The little boy likes red tomatoes"
> Prior_result:(The (little boy) (likes (red tomatoes)))
> Your_result: ((The little boy) (likes (red tomatoes)))
> Sentence: X
> Prior_result: *Aff*
> Your_result:

B Details of Experimental Setting

B.1 Datasets

MSCOCO (Lin et al., 2014) provides five descriptive sentences for each image. Its validation set and test set each contain 1,000 images and 5,000 corresponding sentences. The data preprocessing method follows (Shen et al., 2019; Kim et al., 2019; Zhao and Titov, 2020), where all punctuation is removed.

YouCook2 (Zhou et al., 2018) consists of 89 cooking recipes. Each video includes an average of six procedural steps, annotated with temporal boundaries and described using imperative English sentences. Its test set, used for experimental evaluation, contains 3,310 video-sentence pairs.

B.2 Implementation Details

For the text + image dataset, LoRA fine-tuning uses llava-v1.5-7b and Qwen-VL-Chat, with learning rates set to 2e-4 and 1e-5, and the maximum output length of the LLM is set to 2048. For the text + video dataset, LoRA fine-tuning uses VideoLLaVA-7B and InternVL2-8B, with learning rates both set to 2e-4, and the maximum output length of the LLM is set to 2048. The temperature is set to 0.2 to increase model determinism. The models in this study are implemented using PyTorch and the Huggingface Transformers library.

B.3 Baselines

1) **VC-PCFG** (Zhao and Titov, 2020): A 2020 model for visually grounded constituency parsing, which extends probabilistic context-free grammars (PCFG) with end-to-end differentiable learning and integrates image-text alignment with a language modeling objective.

2) **GPT-4o-mini**: A compact multimodal model launched by OpenAI in 2024, supporting both text and image inputs, featuring a 128K context window. It outperforms GPT-3.5 Turbo while reducing costs by over 60%.

3) **BLIVA** (Hu et al., 2024): An enhanced version of InstructBLIP with a Visual Assistant, improving performance on text-rich and general VQA tasks.

4) **LLM-enhanced ST** (Li et al., 2023b): A method for cross-domain constituency parsing that enhances traditional self-training by using large language models (LLMs) to generate domain-specific raw corpora iteratively. The approach incorporates grammar rules to guide the LLM and criteria for selecting pseudo instances, outperforming traditional self-training methods.

5) **Qwen-VL** (Bai et al., 2023): A vision-language model series based on Qwen-LM, designed for text and image understanding, excelling in image captioning, question answering, and visual grounding, with a strong performance in real-world dialog tasks.

6) **LLaVA-1.5** (Liu et al., 2024a): A large multimodal model that improves LLaVA's vision-language cross-modal connector with CLIP-ViT-L-336px and MLP projection, achieving state-of-the-art performance across 11 benchmarks with just

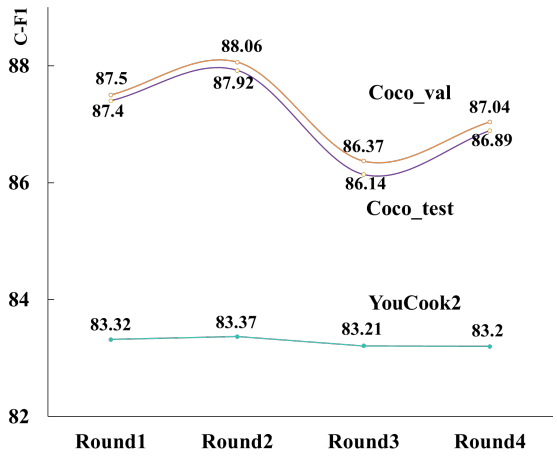

Figure 4: The performance trend of C-F1 with different maximum iteration rounds for debating.

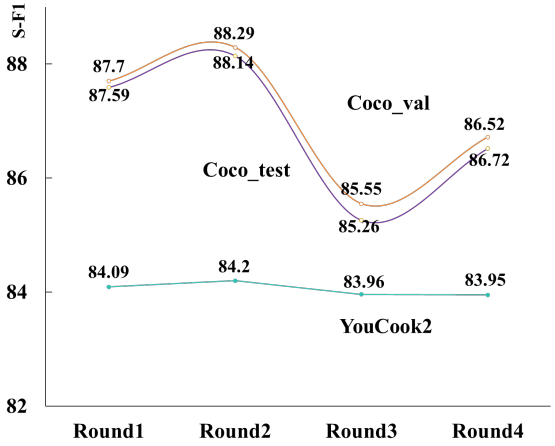

Figure 5: The performance trend of S-F1 with different maximum iteration rounds for debating.

1.2M publicly available data.

7) **PTC-PCFG** (Zhang et al., 2022): A model that leverages loosely correlated video-text data to induce syntactic grammars without relying on manually designed features. Trained on large-scale YouTube data with no direct text-video alignment, it outperforms previous methods on three unseen datasets, achieving higher F1 scores despite domain shifts and noisy labels.

8) **Qwen2_VL** (Wang et al., 2024a): An upgraded version of Qwen-VL, featuring dynamic image resolution and Multimodal Rotary Position Embedding (M-RoPE) for enhanced multimodal processing.

9) **InternVL2** (Chen et al., 2024): A multimodal large language model in the InternVL series. It outperforms most open-source models and is competitive with commercial models across tasks like document comprehension, chart analysis, OCR, scientific problem-solving, and cultural understanding.

10) **VideoLLaVA** (Lin et al., 2023): A unified vision-language model that integrates images and videos into the same language feature space, addressing misalignment in previous approaches. It outperforms Video-ChatGPT and achieves superior results across multiple image and video benchmarks, showing mutual enhancement between images and videos.

C Analysis and Discussion

To investigate the performance trend with the iteration number increasing, we report the performance of our **MMD** when the maximum iterations n is set as 1, 2, 3 and 4 as shown in Figure 4 and 5. From these figures, we can see that on all datasets, as the maximum number of iterations increases, the performance of our **MLD** first improves and then decreases. Among them, the performance is the best in the second max iterations, significantly exceeding all baselines. This indicates that our method requires multiple iterations to make the two models help each other, but not necessarily the more iterations the better. Additionally, on the dataset YouCook2, the performance impact of different max iteration times is not significant. This may be due to the characteristics of this video-text dataset, in which video contains a large amount of information. By effectively capturing helpful information for the text from the beginning, it can already make accurate predictions. So the following discussion of our MMD may not be very necessary.

D Details of Experiments

We compare many advanced baselines with our proposed Multi-MLLM Debating (MLD) approach on three datasets of different domains as follows:

D.1 Overall Comparison

We first report the overall performance on all kinds of constituency labels as shown in Table 2.

The following observations can be drawn from this table:

(1) The MMD approach, which leverages debates between multiple LLMs (Qwen-VL and LLaVA-1.5 for COCO datasets; VideoLLaVA and InternVL2 for YouCook2), consistently achieves the highest scores across all metrics and datasets. This highlights the strength of the multi-MLLM debate strategy in improving parsing accuracy.

(2) Traditional parsing methods like VC-PCFG and PTC-PCFG show significantly lower performance, especially on metrics like C-F1 and S-F1, when compared to modern LLM-based approaches. This indicates the limitations of traditional models in handling complex multimodal data.

(3) LLM-enhanced ST, which involves self-training using traditional parsers and LLM-generated corpora, achieves competitive results across all datasets. This demonstrates the importance of leveraging both traditional and modern methodologies for boosting model performance.

(4) While GPT-4o-mini performs better than traditional methods, it falls behind fine-tuned or debated models like BLIVA and InternVL2. This suggests that general-purpose large models require fine-tuning or debate mechanisms to excel in constituency parsing tasks.

D.2 Fine-grained Comparison

We also report the fine-grained performance on each type of constituent, such as verb phrases (VP), prepositional phrase (PP), and subordinated clause (SBAR), which is shown in Table 3, 4 and 5 . From these tables, we can observe that most baselines cannot maintain stable performance on different datasets or constituent categories, while our **MMD** outperforms other methods significantly in all scenarios. This indicates that our approach can adapt to various scenarios through multi-MLLM debating strategy.

Table 3: Different Models' Performance on the COCO Validation Dataset

Model	NP	VP	PP	SBAR	ADJP	ADVP	C-F1	S-F1
VC-PCFG	50.97	83.32	84.14	88.02	75.48	88.57	58.34	58.26
GPT-4o-mini	65.68	62.72	75.17	84.30	18.55	87.14	74.30	74.91
BLIVA	71.46	68.64	63.17	65.70	14.29	48.86	76.26	75.45
LLM-enhanced ST*	89.42	74.88	74.98	55.79	7.25	44.00	86.80	86.22
Qwen-VL	67.50	60.47	75.26	84.71	21.54	64.00	74.30	73.55
LLaVA-1.5	56.32	76.94	75.86	84.92	28.78	65.71	76.72	71.36
MMD (CD)	80.34	88.28	83.20	91.11	18.97	80.54	85.62	86.26
MMD (RD)	82.82	90.24	86.91	95.04	28.36	83.71	88.01	88.20

Table 4: Different Models' Performance on the COCO Test Dataset.

Model	NP	VP	PP	SBAR	ADJP	ADVP	C-F1	S-F1
VC-PCFG	54.90	83.20	80.90	89.00	38.80	86.30	59.30	59.40
GPT-4o-mini	65.26	63.10	73.58	77.48	17.37	84.07	73.85	74.66
BLIVA	71.52	69.93	63.00	68.87	14.25	52.75	76.54	75.82
LLM-enhanced ST*	88.89	74.65	73.96	61.59	9.35	43.68	86.48	85.99
Qwen-VL	67.04	61.22	76.26	87.64	19.60	66.76	74.62	73.93
LLaVA-1.5	57.04	77.58	77.46	90.29	28.51	65.38	77.25	72.13
MMD (CD)	80.71	86.15	80.26	88.21	19.22	78.25	86.29	86.30
MMD (RD)	82.64	90.27	86.88	96.91	26.50	81.87	87.89	88.08

Table 5: Different Models' Performance on the YouCook2 Dataset

Model	NP	VP	PP	SBAR	ADJP	ADVP	C-F1	S-F1
PTC-PCFG	78.70	69.90	80.50	58.90	43.20	65.00	58.90	63.20
GPT-4o-mini	66.56	68.11	85.11	83.84	27.69	65.00	69.91	70.72
Qwen2_VL	76.70	61.51	85.36	85.86	53.85	40.00	78.21	77.12
LLM-enhanced ST*	84.22	40.83	79.85	8.08	24.62	20.00	82.41	82.34
InternVL2	79.70	64.39	92.42	89.90	46.15	60.00	82.53	82.49
VideoLLaVA	73.94	43.42	82.20	80.81	36.92	55.00	75.34	77.20
MMD (CD)	81.62	65.51	89.90	90.35	49.38	61.21	82.59	83.03
MMD (RD)	86.51	69.04	94.15	92.93	56.92	65.00	83.37	84.20