Beyond Unimodal Shortcuts: MLLMs as Cross-Modal Reasoners for Grounded Named Entity Recognition

Jinlong Ma[1], Yu Zhang[1], Xuefeng Bai[1]*, Kehai Chen[1], Yuwei Wang[2]*, Zeming Liu[3], Jun Yu[1], Min Zhang[1]

[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China,
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China,
[3]School of Computer Science and Engineering, Beihang University, Beijing, China
24s151046@stu.hit.edu.cn, yuzhang2717@gmail.com, zmliu@buaa.edu.cn,
ywwang@ict.ac.cn, {baixuefeng, chenkehai, yujun, zhangmin2021}@hit.edu.cn

Abstract

Grounded Multimodal Named Entity Recognition (GMNER) aims to extract text-based entities, assign them semantic categories, and ground them to corresponding visual regions. In this work, we explore the potential of Multimodal Large Language Models (MLLMs) to perform GMNER in an end-to-end manner, moving beyond their typical role as auxiliary tools within cascaded pipelines. Crucially, our investigation reveals a fundamental challenge: MLLMs exhibit *modality bias*, including visual bias and textual bias, which stems from their tendency to take unimodal shortcuts rather than rigorous cross-modal verification. To address this, we propose Modality-aware Consistency Reasoning (**MCR**), which enforces structured cross-modal reasoning through Multi-style Reasoning Schema Injection (MRSI) and Constraint-guided Verifiable Optimization (CVO). MRSI transforms abstract constraints into executable reasoning chains, while CVO empowers the model to dynamically align its reasoning trajectories with Group Relative Policy Optimization. Experiments on GMNER and visual grounding tasks demonstrate that MCR effectively mitigates modality bias and achieves superior performance compared to existing baselines. The code and data are released at https://github.com/aaaalonga/MCR.

1 Introduction

Grounded Multimodal Named Entity Recognition (GMNER, Yu et al., 2023) aims to organize key multimodal information into structured representations, which simultaneously identifies named entities in the text and grounds them to their corresponding visual bounding boxes. As a foundational task, GMNER facilitates various downstream applications, such as recommendation systems (Acharya et al., 2023) and knowledge-based question answering (Zhang et al., 2024).

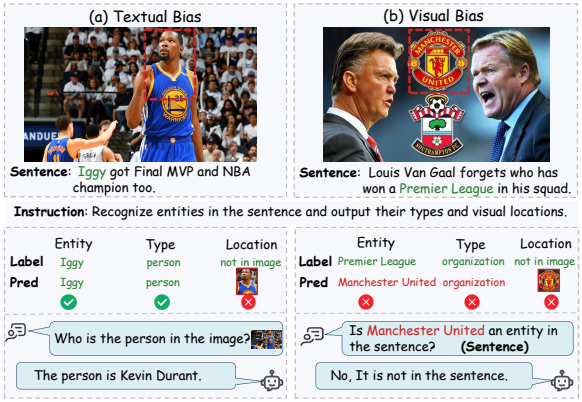

Figure 1: **Error Patterns from Modality Bias in GMNER.** Typical failure cases due to the model's tendency to hallucinate correlations based on unimodal heuristics rather than rigorous cross-modal verification.

Recently, Multimodal Large Language Models (MLLMs, Bai et al., 2025; Yue et al., 2025) have achieved remarkable performance on various vision-language tasks. This progress has motivated researchers to explore their use for GMNER (Tang et al., 2025c,a). However, these approaches typically employ MLLMs as auxiliary tools like image descriptors within cascaded pipelines, which inevitably introduce cumulative error propagation (Tang et al., 2025b) and incur additional computational costs (Ok et al., 2024).

In this work, we take the first step toward exploring the potential of MLLMs for end-to-end GMNER by reformulating it as a generative reasoning task. Our investigation reveals that direct application of MLLMs to GMNER faces a critical pathology: *modality bias*, characterized by the model's tendency to hallucinate correlations based on unimodal heuristics rather than rigorous cross-modal verification. As shown in Figure 1 (a), textual bias causes the model to disregard visual evidence: despite correctly recognizing "Kevin

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 43518–43539

Durant" with image-only input, it incorrectly grounds the text-only entity "Iggy" to the bounding box of "Kevin Durant". Symmetrically, visual bias leads to the neglect of textual semantics. In Figure 1 (b), the model overrides textual context, erroneously recalling "Manchester United" as a named entity driven by visual cues, ignoring its absence in the textual context. We further conduct quantitative analyses that empirically confirm the severity and prevalence of modality bias for different MLLMs in Table 4. These reveal that MLLMs are prone to taking cognitive shortcuts rather than engaging in rigorous deduction, required for strict cross-modal grounding.

To address this, we propose **M**odality-aware **C**onsistency **R**easoning (**MCR**), which enforces structured cross-modal reasoning to mitigate modality bias through Multi-style Reasoning Schema Injection (MRSI) and Constraint-guided Verifiable Optimization (CVO). Specifically, MRSI transforms abstract constraints into executable reasoning chains by synthesizing and injecting diverse reasoning templates to explicitly model the structural dependencies. Furthermore, to empower the model to autonomously explore reasoning trajectories within these structural bounds, CVO is proposed to dynamically align intermediate reasoning process with Group Relative Policy Optimization (GRPO, Guo et al., 2025). This optimization mechanism punishes unimodal shortcuts and encourages the model to generate constraint-faithful rationales, effectively rectifying the intrinsic modality bias. Extensive experiments on Multimodal Named Entity Recognition (Huang et al., 2024; Yu et al., 2023) and Visual Grounding (He et al., 2023) benchmarks verify that our method, applied to Qwen2.5-VL and Mimo-VL, achieves superior performance compared to existing baselines. In-depth analyses confirm that our design explicitly facilitates cross-modal reasoning, effectively mitigating modality bias.

In summary, our contributions are as follows:

- We identify *modality bias* in MLLM-based end-to-end GMNER, revealing that models are prone to unimodal cognitive shortcuts.
- We propose a MCR framework, which enforces explicit, constraint-faithful reasoning through schema injection and verifiable optimization against modality bias.
- We achieve superior performance on multiple benchmarks, demonstrating that structured reasoning is essential for precise cross-modal grounding.

2 Related Work

2.1 Multimodal Named Entity Recognition

Multimodal Named Entity Recognition (MNER, Moon et al., 2018) extracts and classifies named entities from image-text pairs. As a fine-grained extension, Grounded MNER (GMNER, Yu et al., 2023) requires the model to simultaneously recognize the named entities and localize visually present entities via bounding boxes. Existing studies primarily focus on refining cross-modal alignment to suppress visual noise (Liu et al., 2024b; Bao et al., 2024) and enhancing generalization for unseen entities (Wang et al., 2024b). With the emergence of Multimodal Large Language Models (MLLMs, Bai et al., 2025; Yue et al., 2025), recent studies (Tang et al., 2025a,c; Ok et al., 2024) have started to integrate them into GMNER. These approaches primarily exploit the vast semantic priors inherent in MLLMs to refine and align multimodal feature representations, thereby facilitating more accurate entity-image association.

In this work, we move beyond feature alignment to fully exploit the cross-modal reasoning potential of MLLMs, enabling a holistic multimodal interplay for rigorous consistency verification.

2.2 Reasoning in MLLMs

MLLMs have achieved remarkable success across a wide range of domains (Zhu et al., 2025; Zhang et al., 2025a; Zheng et al., 2025; He et al., 2026; Li et al., 2025a; Jiang et al., 2026), demonstrating exceptional capabilities in integrating and reasoning over heterogeneous data. Recent advancements in MLLMs have catalyzed a paradigm shift in complex reasoning tasks (Li et al., 2025c; Chen et al., 2025a; Kumar et al., 2025; Wei et al., 2025; Wang et al., 2025c; Qu et al., 2025; Tong et al., 2026). By introducing explicit reasoning processes into language-level Chains of Thought (CoT, Wei et al., 2022), these models decompose intricate problems into granular, sequential sub-steps (Wang et al., 2022b; Gao et al., 2023; Wang et al., 2025a; Chen et al., 2025b). This reasoning-centric paradigm has proven instrumental in mitigating hallucinations arising from modality misalignment (Zhang et al., 2023; Li et al., 2025b; Zhang et al., 2025b; Wu

et al., 2025), significantly enhancing both the accuracy and stability of multi-step inference.

2.3 Modality bias in MLLMs

Recent works identify *modality bias* (Zhang et al., 2025d; Leng et al., 2024; Zhang et al., 2025c) in MLLMs, observing that models often exhibit intrinsic inclinations toward specific modalities. To address this, prevailing strategies typically employ Reinforcement Learning from Human Feedback (RLHF, Ouyang et al., 2022) by curating extensive preference datasets (Ouyang et al., 2022; Wang et al., 2024a) to enable MLLMs to distinguish between hallucinated and grounded content, effectively mitigating bias and hallucinations.

In this work, we attribute modality bias in GMNER to cognitive shortcuts, where models bypass rigorous verification in favor of unimodal heuristics. To rectify this, we propose Modality-aware Consistency Reasoning (MCR) to explicitly model the interplay between modalities to verify entity existence and spatial alignment, thereby enforcing rigorous cross-modal consistency.

3 Task Formulation

Given a sentence s and its associated image v, Grounded Multimodal Named Entity Recognition (GMNER) can be decomposed into two subtasks:

Multimodal Named Entity Recognition (MNER). MNER recognizes entities in s and assigns each entity a predefined type. And it produces pairs (e_i, t_i), where e_i is an entity span in s and t_i denotes its corresponding type.

Entity Extraction & Grounding (EEG). EEG is parallels generalized Visual Grounding (VG). For each textual entity e_i, decide whether it is visually present in v. If present, output its bounding box b_i; otherwise, output None. Accordingly, the GMNER output can be formulated as:

$$\mathcal{Y} = \{(e_i,\ t_i,\ l_i)\}_{i=1}^{k_1}, \tag{1}$$

where k_1 indicate the numbers of output triples in a sample, and l_i is formed as:

$$l_i = \begin{cases} b_i = (x_1, y_1, x_2, y_2), & e_i \text{ is grounded}, \\ \text{None}, & e_i \text{ is ungrounded}, \end{cases} \tag{2}$$

where $(x_1,\ y_1)$ and $(x_2,\ y_2)$ are the coordinates of the top-left and bottom-right corners.

4 Methodology

To fully leverage multimodal evidence and ensure cross-modal consistency, we propose **M**odality-aware **C**onsistency **R**easoning (**MCR**) including **M**ulti-style **R**easoning **S**chema **I**njection (**MRSI**) and **C**onstraint-guided **V**erifiable **O**ptimization (**CVO**). The framework of MCR is illustrated in Figure 2. MRSI organizes the diverse reasoning schema with modality-specific constraints and enforces explicit reasoning, while CVO leverages the reasoning schema together with GRPO to further strengthen the model's reasoning capability.

4.1 Multi-style Reasoning Schema Injection

To address the modality bias caused by insufficient cross-modal consistency reasoning, we propose MRSI, which injects constraint-centered and diverse reasoning schema into the inference process (Zhou et al., 2025; Zhoubian et al., 2025) to strengthen cross-modal verification. Specifically, MRSI is guided by four core constraints, covering entity recognition $\mathcal{C}_s$, type classification $\mathcal{C}_t$, visual entailment $\mathcal{C}_e$ and visual grounding $\mathcal{C}_u$:

$$\mathcal{C} = \{\mathcal{C}_s, \mathcal{C}_t, \mathcal{C}_e, \mathcal{C}_u\}. \tag{3}$$

Each constraint aligns with the task and its relevant modality. See Appendix B for an example. The resulting reasoning schema in Figure 2 reflects both task- and modality- specific considerations. As shown in Appendix C, through templates, LLMs, or MLLMs, we transform $(s, v, \mathcal{C}, \mathcal{Y}_\tau)$ into programmatic reasoning steps z with multiple styles on the labeled set $\mathcal{D}_\mathcal{G}$:

$$\mathcal{D}_\mathcal{R} = \bigcup_{(s,v,\mathcal{Y}) \in \mathcal{D}_\mathcal{G}} \Gamma_\phi(z \mid s,\ v,\ \mathcal{C},\ \mathcal{Y}), \tag{4}$$

where Γ_ϕ denotes template extractors, LLMs or MLLMs, and $\mathcal{D}_\mathcal{R}$ is the obtained CoT training dataset. The diversity of reasoning schema prevents the sampled trajectories from collapsing into overly similar outputs, avoiding the negligible advantages and gradient vanishing (Xiong et al., 2025; Yao et al., 2025). We use $\mathcal{D}_1$ (a subset of $\mathcal{D}_\mathcal{R}$) to inject reasoning schema into MLLMs (Tang et al., 2025d; Koksal and Alatan, 2025) via:

$$\begin{aligned} \mathcal{L}_{\text{MRSI}} = -\ \mathbb{E}_{(x,v,z,y)\sim\mathcal{D}_1} \big[& \log \pi_{\text{MLLM}}(z \mid x, v) \\ & + \log \pi_{\text{MLLM}}(y \mid x, v, z) \big], \end{aligned} \tag{5}$$

where $\pi_{\text{MLLM}}(z \mid x, v)$ and $\pi_{\text{MLLM}}(y \mid x, v, z)$ respectively denote the probability that MLLMs

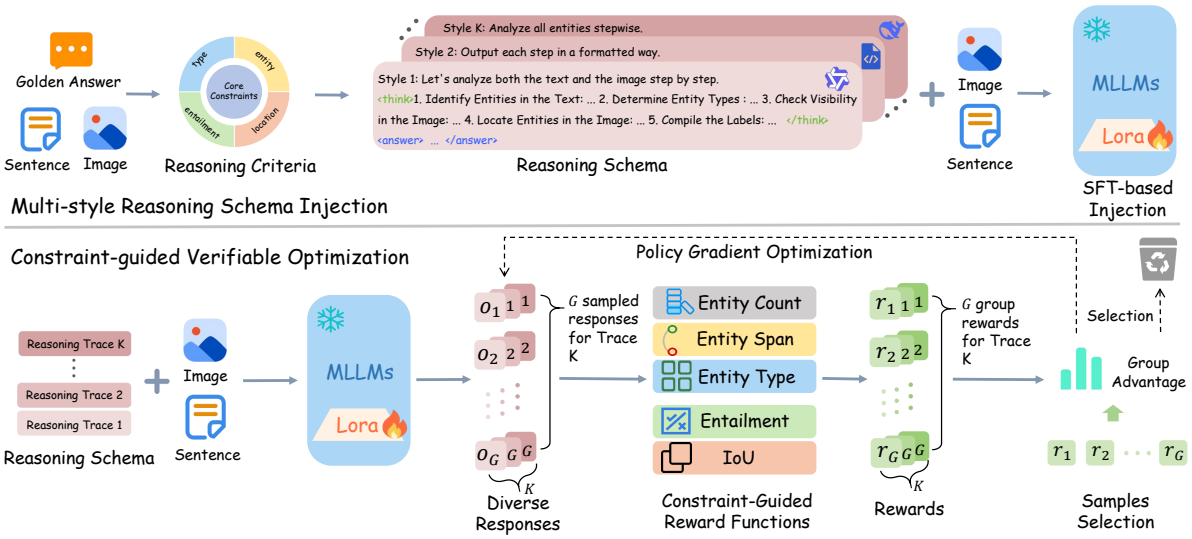

Figure 2: **The Framework of MCR.** The framework consists of two stages: (1) **Multi-style Reasoning Schema Injection** constructs diverse reasoning schema $\mathcal{D}_{\mathcal{R}}$ by treating the core constraints as reasoning criteria and generating multiple reasoning styles from templates, LLMs, and MLLMs based on the image–text inputs and labels. A subset of $\mathcal{D}_{\mathcal{R}}$ is injected into MLLMs through supervised fine-tuning. (2) **Constraint-guided Verifiable Optimization** uses the remaining of $\mathcal{D}_{\mathcal{R}}$ and optimizes the model with verifiable reward functions derived from the core constraints, together with the GRPO algorithm, to enhance cross-modal consistency reasoning.

generate a reasoning path given the image–text pair and predict the answer based on the generated path. Through explicitly introducing $\mathcal{C}$ and z, MRSI compels the model to retain a reasoning path for cross-modal consistency checking.

4.2 Constraint-guided Verifiable Optimization

Reinforcement Learning with Verifiable Rewards (Guo et al., 2025) replaces reward models with reward functions and has shown strong performance on reasoning tasks. Following this, we introduce CVO to enhance cross-modal reasoning.

4.2.1 Constraint-guided Verifiable Reward

Inspired by prior similar reward functions (Liu et al., 2025; Roit et al., 2023; Wang et al., 2025b), we anchor on $\mathcal{C}$ and design rule-based verifiable reward functions for GMNER, including entity count, entity span, entity type, entailment, and localization rewards.

Entity Count, Span and Type Rewards. The entity count reward R_c encourages broader exploration while still maintaining precision, preventing the model from becoming overly conservative. It assigns a score by comparing the number of predicted entity triples with the ground-truth count. See Appendix D.1 for details.

In computing the entity span reward R_s, we compute the token-level F1 score for every predicted–gold entity pair in a sample and perform optimal matching using the Hungarian algorithm. See Appendix D.2 for details. The entity span reward for a sample is defined as the average token-level F1 over all matched pairs:

$$R_s = \frac{1}{k} \sum_{(i,j)\in\mathcal{N}} F_{ij}, \tag{6}$$

where F_{ij} denotes the token-level F1 score between the i-th predicted entity and the j-th gold entity, $\mathcal{N}$ denotes the set of successfully matched entity pairs, and $k = |\mathcal{N}|$ is the number of matched pairs. And the type reward R_{type} is formed as:

$$R_t = \frac{1}{k} \sum_{i}^{k} \mathbb{1}\{\hat{t}_i = t_i\}, \tag{7}$$

where $\mathbb{1}$ denotes the indicator function, which equals 1 if the predicted type $\hat{t}_i$ matches the gold type t_i and 0 otherwise. The sample-level reward R_t is the average over the k matched pairs.

Visual Grounding and Entailment Rewards. The grounding reward function R_u considers the Intersection-over-Union (IoU) metric, which measures the overlap between a predicted bbox and a gold bbox as the intersection area divided by their union. R_u is formed as:

$$R_u = \frac{1}{k} \sum_{i}^{k} \max(0, \frac{\text{IoU}_i - \sigma}{1 - \sigma}), \tag{8}$$

where IoU_i denotes the IoU between the i-th predicted and gold bbox, and we apply a threshold σ and bounding boxes with an IoU below the threshold σ are set 0 of IoU (Liu et al., 2025), while bounding boxes with an IoU above the threshold σ are linearly mapped into $[0, 1]$. The sample-level reward R_u is the average over the k matched pairs. And the entailment reward R_e is formed as:

$$R_e = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{\hat{v}_i = v_i\}, \\ v_i = \mathbb{1}\{l_i \neq \text{None}\}, \;\; \hat{v}_i = \mathbb{1}\{\hat{l}_i \neq \text{None}\}, \tag{9}$$

where $\mathbb{1}$ denotes the indicator function, which returns 1 if the condition inside the braces is true and 0 otherwise, v_i and $\hat{v}_i$ respectively indicate whether the gold and predicted entities are visible. For each matched entity, the reward is set to 1 if the predicted and gold locations are both None or both non-None; otherwise, it is 0. The sample-level reward R_e is the average over the k matched pairs.

Finally, our overall reward is a weighted combination of the above rewards:

$$R = \lambda_1 R_c + \lambda_2 R_s + \lambda_3 R_t + \lambda_4 R_u + \lambda_5 R_e, \tag{10}$$

where λ_j denotes the weight of each reward term.

4.2.2 Optimization with Verifiable Rewards

Given verifiable rewards R and remaining data $\mathcal{D}_2 = \mathcal{D}_\mathcal{R} \backslash \mathcal{D}_1$, CVO optimizes the policy to align cross-modal verification with the core constraints. For each query q sampled, the current policy $\pi_{\theta_{\text{old}}}$ generates G diverse responses $\{o_1, o_2, \ldots, o_G\}$. The verifiable reward functions compute scores $\{r_1, r_2, \ldots, r_G\}$ for each response by R. We then obtain the group advantage A_i:

$$A_i = \frac{r_i - \mu_G}{\sigma_G}, \tag{11}$$

where μ_G denotes the empirical mean of the group rewards computed over $\{r_1, r_2, \ldots, r_G\}$, σ_G denotes their empirical standard deviation. To further improve training efficiency and reduce collapse risk, we apply sampling-based filtering to $\mathcal{D}_2$ based on reward distribution statistics. See the Appendix D.3 for details. CVO updates the policy with a GRPO style objective. The learning objective uses a clipped importance ratio to prevent overly aggressive updates and a length normalization to keep responses comparable. The objective function is defined as follows:

$$\mathcal{J}_{\text{CVO}}(\theta) = \mathbb{E}[q \sim P(Q),\; o \sim \pi_{\theta_{\text{old}}}(O \mid q)] \\ \frac{1}{|o|}\sum_{t=1}^{|o|} \min\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})} A_t, \right. \\ \left. \text{clip}\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) A_t \right), \tag{12}$$

where clipping with threshold ε prevents overly aggressive updates and length normalization ensures comparability across responses. The design yields stable group preference optimization without a critic and supports constraint anchored reasoning in multimodal settings.

5 Experiments

We briefly introduce the datasets, baselines, and evaluation metrics used in our experiments, with further details provided in the Appendix E.

5.1 Experimental Setup

Datasets. Training datasets are from three sources: Twitter-GMNER (Yu et al., 2023) for GMNER, a multi-image MNER dataset MNER-MI (Huang et al., 2024), and a visual grounding dataset, GREC (He et al., 2023). We evaluate Twitter-GMNER along with its two subtasks (MNER and EEG) in the main experiments, and conduct additional evaluations on MNER-MI and GREC.

Baselines. Following prior work (Tang et al., 2025b), we categorize existing approaches into pipeline and unified methods based on whether textual entity extraction and visual region prediction are executed within a single pass. Furthermore, we investigate the applicability of open-source and close-source MLLMs within an end-to-end paradigm. To establish robust benchmarks, we implement *Chain-of-Thought* (*CoT*, Wei et al., 2022), *Few-shot prompting* (Brown et al., 2020), and *Supervised Fine-tuning* (*SFT*, Ouyang et al., 2022) as strong baselines for comparison.

Evaluation Metrics. Following Yu et al. (2023), we evaluate GMNER, MNER and VG using Precision, Recall and F1 score. For the subtasks in GMNER, MNER identifies and classifies entities, while EEG grounds named entities in the image.

5.2 Main Results

As shown in Table 1, among training-free methods, incorporating explicit reasoning via Chain-of-Thought (CoT) or Few-shot prompting yields

Type	Methods	GMNER			MNER			EEG		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Pipeline	ITA-VinVL-EVG (Wang et al., 2022a)	52.4	50.8	51.6	80.4	78.4	79.4	56.6	54.8	55.7
	BARTMNER-VinVL-EVG (Yu et al., 2023)	52.5	52.4	52.5	80.7	80.1	80.4	55.7	55.6	55.7
	SCANNER (Ok et al., 2024)	68.3	68.7	68.5	-	-	-	-	-	-
	ReFineG (Tang et al., 2025a)	54.1	60.2	57.0		-	-	-	-	-
	UnCo (Tang et al., 2025c)	-	-	64.6	-	-	81.7	-	-	69.6
Unified	MNER-QG (Jia et al., 2023)	53.0	54.8	53.9	78.2	78.6	78.4	58.5	56.6	57.5
	H-Index (Yu et al., 2023)	56.2	56.7	56.4	79.4	80.1	79.7	60.9	61.5	61.2
	TIGER (Wang et al., 2023a)	55.8	57.5	56.6	79.9	80.1	80.3	60.7	61.8	61.3
	MQSPN (Tang et al., 2025b)	59.0	58.5	58.8	81.2	79.7	80.4	61.9	62.9	62.4
End-to-End	**GLM4.5VL** (Hong et al., 2025)	33.0	44.4	37.8	43.0	57.9	49.4	36.2	48.7	41.6
	+CoT	40.9	50.3	45.1	53.7	66.1	59.3	44.6	54.8	49.2
	+CoT+3-Shot	43.2	55.5	48.5	53.1	68.3	59.7	47.2	60.7	53.1
	Qwen2.5VL-72B (Bai et al., 2025)	24.0	44.5	31.2	32.2	59.8	41.9	26.3	48.9	34.2
	+CoT	30.4	45.2	36.3	40.5	60.3	48.4	33.7	50.2	40.3
	+CoT+3-Shot	33.0	52.3	40.5	47.0	74.4	57.6	37.2	58.8	45.6
	Qwen2.5VL-7B (Bai et al., 2025)	5.40	14.1	7.80	9.80	25.3	14.1	6.10	15.9	8.80
	+CoT	11.4	13.6	12.4	20.2	24.0	21.9	12.9	15.4	14.1
	+CoT+3-Shot	16.5	33.7	22.2	27.5	56.2	37.0	18.3	37.3	24.5
	+SFT	63.3	62.0	62.7	83.0	81.3	82.2	65.8	64.4	65.1
	+MRSI **(ours)**	69.1	68.1	68.6	82.4	81.2	81.8	72.1	71.0	71.5
	+MRSI+CVO **(ours)**	**70.5**	**70.8**	**70.6**	**82.6**	**82.9**	**82.8**	**73.2**	**73.5**	**73.4**
	MimoVL-7B (Yue et al., 2025)	9.60	10.9	10.2	20.5	23.3	21.8	10.6	12.0	11.2
	+CoT	11.9	17.3	14.1	22.1	32.2	26.2	13.1	19.0	15.5
	+CoT+3-Shot	15.0	21.4	17.7	29.6	42.1	34.8	17.8	25.3	20.9
	+SFT	63.5	60.6	62.0	81.7	78.0	79.8	67.0	63.9	65.4
	+MRSI **(ours)**	66.1	65.5	65.8	81.5	80.8	81.1	69.8	69.2	69.5
	+MRSI+CVO **(ours)**	**69.4**	**69.7**	**69.6**	**82.2**	**82.5**	**82.3**	**72.8**	**73.1**	**72.9**

Table 1: **Comparison on GMNER, MNER and EEG. Pre**, **Rec** and **F1** respectively denote **Precision**, **Recall** and **F1 score**. Best in each block is bold.

notable performance gains compared to the direct application of MLLMs. Upon integrating our proposed MCR framework (with MRSI and CVO), all MLLMs consistently outperform existing baselines. Specifically, on GMNER, MCR improves over the previous best unified method MQSPN (Tang et al., 2025b) by 11.87% F1 scores and over the best pipeline method SCANNER (Ok et al., 2024) by 2.11% F1 scores. Moreover, using Qwen2.5VL-7B, MCR outperforms both Qwen2.5VL-72B and GLM4.5VL. And MCR respectively surpasses direct Supervised Fine-Tuning (SFT) by 8.05% and 7.57% F1 scores on Qwen2.5VL-7B and MimoVL-7B. On MNER, MCR outperforms all methods at least 2.33% F1 scores. On EEG, MCE exceeds the best unified method MQSPN by 10.97% F1 scores.

5.3 Performance on uni-modal bias dataset

Pipeline methods often decompose GMNER into MNER and VG, where the bidirectional modality biases in GMNER also manifest separately. Beyond GMNER, MCR further leverages datasets from these tasks for training and evaluation, using MNER-MI (Huang et al., 2024) for MNER and GREC (He et al., 2023) for VG. MNER-MI features weak text–image correlation, making visual bias more likely, while GREC may induce textual bias when a description corresponds to zero region. We evaluate SFT and MCR on MimoVL-7B and Qwen2.5VL-7B across these datasets.

Performance on MNER. Table 2 show that MCR outperforms SFT on both tasks. Within MCR, the second-stage CVO generally surpasses the first-stage MRSI. This indicates that proposed method effectively leverages visual information to support entity extraction and classification while reducing the impact of irrelevant noise in the image.

Performance on VG. In GREC, N-acc and Precision (He et al., 2023) respectively evaluate grounding accuracy for cases with zero target

Methods	MNER-MI			GREC-testA		GREC-testB	
	Pre	Rec	F1	N-acc	Pre	N-acc	Pre
Qwen2.5VL-7B	-	-	-	-	-	-	-
+SFT	84.0	83.6	83.8	70.6	81.1	70.2	62.5
+MRSI	82.1	81.8	82.0	74.2	89.1	70.3	71.5
+MRSI+CVO	84.1	**86.2**	**85.1**	74.7	**90.4**	69.7	72.8
MimoVL-7B	-	-	-	-	-	-	-
+SFT	81.8	80.8	81.2	65.7	66.1	69.6	45.6
+MRSI	83.2	83.4	83.3	72.6	89.4	68.6	71.6
+MRSI+CVO	**84.7**	85.0	84.8	**75.7**	90.4	**71.9**	**73.0**

Table 2: **Results on MNER-MI and GREC.** For GREC dataset, we remove cases where a single textual description corresponds to multiple image regions.

Methods	GMNER			MNER			EEG		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Ours	**70.5**	**70.8**	**70.6**	**82.6**	**82.9**	**82.8**	**73.2**	**73.5**	**73.4**
w/o MRSI	50.2	53.2	51.7	64.3	68.1	66.2	54.2	57.5	55.8
w/o CVO	68.9	68.6	68.8	81.9	81.5	81.7	72.2	71.9	72.0
w/o $\mathcal{D}_{\mathcal{R}}$	69.5	68.8	69.2	81.6	80.7	81.1	72.8	72.1	72.5
w/o Inst	67.3	66.1	66.7	81.7	80.2	80.9	70.8	69.5	70.1

Table 3: **Ablation results on GMNER, MNER, and EEG. w/o $\mathcal{D}_{\mathcal{R}}$** means removing diverse reasoning styles and training MCR with a single style. **w/o Inst** means removing the components that specify stepwise cross-modal verification and cautionary guidelines from the original instructions.

region and with a target region. N-acc reflects the models' ability to judge entailment between text and image and thus partially characterizes textual bias. As shown in Table 2, N-acc improves across models, indicating that our method effectively reduces text bias in MLLMs.

5.4 Component Analysis

MRSI is crucial for enabling cross-modal reasoning. As shown by w/o MRSI in Table 3, directly applying CVO on $\mathcal{D}_{\mathcal{R}}$ without MRSI leads to an 18.95% drop in F1. This indicates that the stepwise cross-modal verification path established by MRSI is a necessary prerequisite, as it provides the model with the essential reasoning structure.

CVO aligns reasoning with task objectives. As shown by w/o CVO in Table 3, applying MRSI on the same data used for CVO yields only a marginal 0.2% improvement. In contrast, strengthening the model's reasoning with CVO adds to a 2.0% gain in Table 1. This contrast highlights that MRSI provides the essential reasoning foundation, whereas CVO is crucial for pushing the model beyond the MRSI plateau and further enhancing cross-modal verification.

Multi-style reasoning schema enhance training performance. w/o $\mathcal{D}_{\mathcal{R}}$ eliminates reasoning diversity, and training MCR with a single reasoning style. It slows down the CVO stage and results in smaller performance gains. This degradation is caused by reduced coverage and exploration of reasoning trajectories and shortcut reliance.

Reasoning-related instruction components effectively guide the model. w/o Inst removes the components that specify stepwise cross-modal verification and cautionary guidelines for MCR training from the original instructions. Without this constraint-aware guidance, the model struggles to establish clear intermediate goals and consistent verification criteria, which weakens execution fidelity at key steps and results in a 3.9% drop in F1 score under the same training conditions.

5.5 Further Analysis

MCR effectively mitigates visual bias in GMNER. Directly inspecting every test image to quantify how MCR handles visual bias is impractical, so we introduce two indirect metrics. Based on whether a recalled entity appears in the input sentence, we define **N-Count** as the number of recalled entities that are absent from the sentence, and **N-Rate** as the proportion of such entities among all recalled entities. As shown in Table 4, training-free approaches such as CoT and few-shot prompting can partially alleviate visual bias in GMNER. In contrast, MCR reduces visual bias for Qwen2.5VL-7B and MimoVL-7B to a near-negligible level.

Model	Method	N-Rate (%)	N-Count
Qwen2.5VL-72B	Direct Prompt	29.2	1372
	CoT	15.6	610
	CoT+3-Shot	5.8	241
GLM4.5VL	Direct Prompt	26.9	898
	CoT	24.3	820
	CoT+3-Shot	13.8	464
Qwen2.5VL-7B	CoT+3-Shot	13.2	363
	MCR(**ours**)	0.1	3
MimoVL-7B	CoT+3-Shot	24.3	637
	MCR(**ours**)	0.2	5

Table 4: **Quantitative results of visual bias. N-Count** means the number of recalled entities that are absent from the sentence and **N-Rate** means the proportion of such entities among all recalled entities. Direct Prompt denotes a concise task instruction.

MCR effectively mitigates textual bias in GMNER. Inspired by the N-acc metric in GREC,

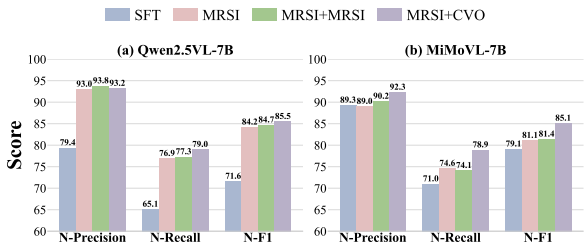

Figure 3: **Quantitative results of textual bias.** MCR effectively improves the models' ability to determine whether an entity is present, which in turn indicates that MCR mitigates textual bias.

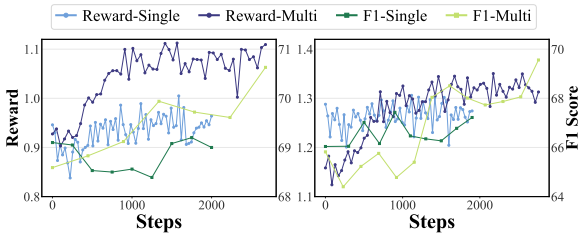

Figure 4: **Effect of multi-style vs. single-style reasoning schema on F1 and reward scores in CVO on Qwen2.5VL-7B (Left) and MimoVL-7B (Right). Single** means single-style reasoning schema and **Multi** means multi-style reasoning schema.

we introduce three metrics to quantify textual bias in GMNER. **N-Pre** measures the fraction of predicted text-only triples with location "None" that correctly match gold text-only triples, while **N-Rec** measures the proportion of gold text-only triples that the model correctly predicts as having no location. **N-F1** is the harmonic mean of N-Pre and N-Rec. As shown in Figure 3, MCR (MRSI + CVO) on Qwen2.5VL-7B improves over SFT by nearly 14% across all three metrics, indicating effective mitigation of textual bias. Notably, MRSI + MRSI denotes continuing MRSI training on the data used for CVO. The results show that simply scaling up data for MRSI brings only marginal improvement, highlighting that CVO is essential for refining cross-modal consistency and unlocking downstream performance gains.

Multi-style reasoning schema help improve the training effectiveness of CVO. Figure 4 compares single-style and multi-style reasoning schema during CVO training in terms of both reward and F1 score. At the early stage, single-style reasoning achieves higher rewards and F1 scores due to more focused supervision from MRSI. However, we observe that multi-style reasoning yields higher rewards and F1 scores finally, indicating its stronger ability to stimulate exploration and support more effective policy optimization. Moreover, it leads to more stable optimization, while single-style reasoning suffers from larger fluctuations.

Controlled policy exploration in CVO. As shown in Figure 5, when multi-style reasoning schema is used in CVO, the cross-entropy increases gradually and then stabilizes, indicating that the model performs controlled and effective exploration over diverse reasoning strategies. Meanwhile, the mean completion length first decreases and then converges, suggesting that the model gradually settles into more concise and consistent reasoning patterns. In contrast, when CVO is trained with single-style reasoning schema, the cross-entropy exhibits only a very slow increase, while the mean completion length fluctuates without clear convergence.

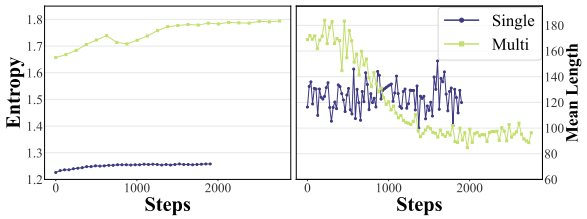

Figure 5: **Effect of multi-style vs. single-style reasoning schema on cross entropy (Left) and mean completion length (Right) in CVO on MimoVL-7B. Single** means single-style reasoning schema and **Multi** means multi-style reasoning schema.

Attention analysis. We analyze the underlying mechanism via attention distributions over visual and input sentence tokens during entity grounding generation. For cases with ground-truth grounding labeled as None, we use a post-SFT Qwen2.5VL-7B to select 50 correct predictions (Normal) and 50 incorrect predictions (Textual Bias), and compare them with MCR. As shown in Table 5, since generating entity grounding outputs inherently requires visual evidence, successful predictions exhibit an appropriate attention toward vision. In contrast, Textual Bias cases reveal the underlying failure mode, where attention is misallocated and disproportionately biased toward textual modality. MCR corrects this pattern and achieves even stronger visual focus, mitigating textual bias through structured cross-modal reasoning.

Cases	vision ratio (%)	text ratio (%)
Normal	71.8	28.2
Textual Bias	59.6	40.4
MCR(**ours**)	74.4	25.6

Table 5: **Underlying mechanism of unimodal shortcuts. Vision ratio** and **text ratio** denote the proportion of attention assigned to **visual tokens** and **sentence tokens** during entity grounding generation.

Training and inference time cost. We compare MCR and SFT in both training and inference cost, and further evaluate their training efficiency by extending SFT to match the training time of MCR. As shown in Table 6 , while MCR requires more training time than basic SFT, merely increasing the SFT training budget fails to yield comparable gains. In contrast, MCR breaks this optimization plateau and achieves substantially higher training efficiency. Moreover, MCR keeps inference cost relatively low, whereas previous pipeline methods often introduce substantial inference overhead due to cascaded components (Lin et al., 2025). Additional details on training cost are provided in the Appendix E.5.

Methods	Inference (min)	Training (h)	GMNER	Imp/h
SFT_{base}	3	6	63.6	-
SFT_{cont}	3	19 (+13)	64.8 (+1.2)	0.10
MCR(ours)	6	19 (+13)	70.6 (+7.0)	0.54

Table 6: **Training and inference cost comparison.** $\mathbf{SFT_{base}}$ denotes basic SFT, $\mathbf{SFT_{cont}}$ denotes SFT extended to matching MCR's training time, and **Imp/h** denotes the average GMNER F1 improvement per additional training hour relative to SFT_{base}.

Case Study. As shown in Figure 6, Naive End-to-end methods may incorrect grounding due to insufficient cross-modal verification, such as grounding the "NBA" logo to the textual entity "NFL". MCR mitigates this by explicitly reasoning over image–text consistency. More cases are provided in Appendix H.

6 Conclusion

In this work, we reformulate GMNER as an end-to-end generative reasoning task. We diagnose a critical pathology—*modality bias*—revealing that MLLMs often rely on unimodal shortcuts rather than rigorous cross-modal verification. To address this, we propose Modality-aware Consistency Reasoning (**MCR**), which enforces structured

Figure 6: **Case of MCR Mitigating Modality Bias.**

cross-modal reasoning to mitigate modality bias. Comprehensive evaluations on GMNER, MNER, VG and FMNERG benchmarks demonstrate that MCR effectively mitigates modality bias, enabling rigorous cross-modal verification and achieves superior performance compared to existing baselines. Besides, ablation experiments validates the necessity of our design choices and the stability of the optimization mechanism.

Limitations

Despite the promising performance of MCR in mitigating modality bias across GMNER, MNER, and VG tasks, our framework remains constrained by the inherent parametric knowledge limits of the underlying MLLMs. Specifically, MCR relies on the model's internal knowledge base for entity recognition; consequently, it may struggle to generalize to unseen entities that are absent from the pre-training corpus.

Acknowledgments

This work was supported in part by the Xinjiang Science and Technology Development Plan for the Two Innovation Demonstration Zones along the Silk Road Economic Belt under Grant 2024LQ03003, in part by the National Natural Science Foundation of China (62406091, 62276077, U23B2055, U24A20328, 62527822, 62350710797), in part by the Guangdong Basic and Applied Basic Research Foundation (2026A1515011718, 2024A1515011205), and in part by Shenzhen Science and Technology Program (KQTD20240729102154066).

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th*

ACM conference on recommender systems, pages 1204–1207.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xigang Bao, Mengyuan Tian, Luyao Wang, Zhiyuan Zha, and Biao Qin. 2024. Contrastive pre-training with multi-level alignment for grounded multimodal named entity recognition. In *Proceedings of the 2024 international conference on multimedia retrieval*, pages 795–803.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. 2025b. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. 2023. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*.

Zefeng He, Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Yu Cheng, and Yang Yang. 2026. Gems: Agent-native multimodal generation with memory and skills. *arXiv preprint arXiv:2603.28088*.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shizhou Huang, Bo Xu, Changqun Li, Jiabo Ye, and Xin Lin. 2024. Mner-mi: A multi-image dataset for multimodal named entity recognition in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11452–11462.

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 8032–8040.

Lingjie Jiang, Shaohan Huang, Xun Wu, Yixia Li, Guanhua Chen, Dongdong Zhang, and Furu Wei. 2026. Viscodex: Unified multimodal code generation via merging vision and coding models. In *The Fourteenth International Conference on Learning Representations*.

Aybora Koksal and A Aydin Alatan. 2025. Milchat: Introducing chain of thought reasoning and grpo to a multimodal small language model for remote sensing. *arXiv preprint arXiv:2505.07984*.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*.

Bo Li, Shaolin Zhu, and Lijie Wen. 2025a. Mit-10m: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025b. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

Jinyuan Li, Han Li, Di Sun, Jiahao Wang, Wenkun Zhang, Zan Wang, and Gang Pan. 2024. Llms as bridges: Reformulating grounded multimodal named entity recognition. *arXiv preprint arXiv:2402.09989*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025c. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Xinkui Lin, Yuhui Zhang, Yongxiu Xu, Kun Huang, Hongzhang Mu, Yubin Wang, Gaopeng Gou, Li Qian, Li Peng, Wei Liu, and 1 others. 2025. Makar: a multi-agent framework based knowledge-augmented reasoning for grounded multimodal named entity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6121–6141.

Jintao Liu, Chenglong Liu, and Kaiwen Wei. 2024a. Multi-view prompt for fine-grained multimodal named entity recognition and grounding. In *ECAI 2024*, pages 2693–2700. IOS Press.

Peipei Liu, Hong Li, Yimo Ren, Jie Liu, Shuaizong Si, Hongsong Zhu, and Limin Sun. 2024b. Hierarchical aligned multimodal learning for ner on tweet posts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18680–18688.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862*.

Hyunjong Ok, Taeho Kil, Sukmin Seo, and Jaeho Lee. 2024. Scanner: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities. *arXiv preprint arXiv:2404.01914*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, and 1 others. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, and 1 others. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 6252–6272.

Jielong Tang, Shuang Wang, Zhenxing Wang, Jianxing Yu, and Jian Yin. 2025a. Refineg: Synergizing small supervised models and llms for low-resource grounded multimodal ner. *arXiv preprint arXiv:2509.10975*.

Jielong Tang, Zhenxing Wang, Ziyang Gong, Jianxing Yu, Xiangwei Zhu, and Jian Yin. 2025b. Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25246–25254.

Jielong Tang, Yang Yang, Jianxing Yu, Zhen-Xing Wang, Haoyuan Liang, Liang Yao, and Jian Yin. 2025c. Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7644–7662.

Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025d. Thinking in character: Advancing role-playing agents with role-aware reasoning. *arXiv preprint arXiv:2506.01748*.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, Chaoran Tao, Zhiyuan Guo, Jizhou Yu, Tianhao Cheng, Zhiheng Xi, Changhao Jiang, Zhangyue Yin, Yining Zheng, Weifeng Ge, and 5 others. 2026. Game-RL: Synthesizing multimodal verifiable game data to boost VLMs' general reasoning. In *The Fourteenth International Conference on Learning Representations*.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Hongru Wang, Deng Cai, Wanjun Zhong, Shijue Huang, Jeff Z Pan, Zeming Liu, and Kam-Fai Wong. 2025a. Self-reasoning language models: Unfold hidden reasoning chains with few reasoning catalyst. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5578–5596.

Jieming Wang, Ziyan Li, Jianfei Yu, Li Yang, and Rui Xia. 2023a. Fine-grained multimodal named entity recognition and grounding with a generative framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3934–3943.

Jieming Wang, Ziyan Li, Jianfei Yu, Li Yang, and Rui Xia. 2023b. Fine-grained multimodal named entity recognition and grounding with a generative framework. In *Proceedings of the 31st ACM*

International Conference on Multimedia, pages 3934–3943.

Weiqin Wang, Yile Wang, Kehao Chen, and Hui Huang. 2025b. Beyond majority voting: Towards fine-grained and more reliable reward signal for test-time reinforcement learning. *arXiv preprint arXiv:2512.15146*.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and 1 others. 2025c. Mucar: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15037–15059.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022a. Ita: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 3176–3189.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Ziqi Wang, Chen Zhu, Zhi Zheng, Xinhang Li, Tong Xu, Yongyi He, Qi Liu, Ying Yu, and Enhong Chen. 2024b. Granular entity mapper: Advancing fine-grained multimodal named entity recognition and grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3211–3226.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. 2025. First sft, second rl, third upt: Continual improving multi-modal llm reasoning via unsupervised post-training. *arXiv preprint arXiv:2505.22453*.

Chuan Wu, Meng Su, Youxuan Fang, and Shaolin Zhu. 2025. Unveiling multimodal processing: Exploring activation patterns in multimodal llms for interpretability and efficiency. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9005–9016.

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. 2025. Reinforce-ada: An adaptive sampling framework for reinforce-style llm training. *arXiv preprint arXiv:2510.04996*.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and 1 others. 2025. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, and oothers. 2025. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*.

Pingrui Zhang, Xianqiang Gao, Yuhan Wu, Kehui Liu, Dong Wang, Zhigang Wang, Bin Zhao, Yan Ding, and Xuelong Li. 2025a. Moma-kitchen: A 100k+ benchmark for affordance-grounded last-mile navigation in mobile manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6315–6326.

Yu Zhang, Kehai Chen, Xuefeng Bai, Zhao Kang, Quanjiang Guo, and Min Zhang. 2024. Question-guided knowledge graph re-scoring and injection for knowledge graph question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. 2025b. Reasongen-r1: Cot for autoregressive image generation models through sft and rl. *arXiv preprint arXiv:2505.24875*.

Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025c. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*.

Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. 2025d. Debiasing multimodal large language models via noise-aware preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9423–9433.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, and 1 others. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735.

Xiangqing Zheng, Chengyue Wu, Kehai Chen, and Min Zhang. 2025. Locot2v-bench: A benchmark for long-form and complex text-to-video generation. *arXiv preprint arXiv:2510.26412*.

Ruiyang Zhou, Shuozhe Li, Amy Zhang, and Liu Leqi. 2025. Expo: Unlocking hard reasoning with self-explanation-guided reinforcement learning. *arXiv preprint arXiv:2507.02834*.

Sining Zhoubian, Dan Zhang, and Jie Tang. 2025. Rest-rl: Achieving accurate code reasoning of llms with optimized self-training and decoding. *arXiv preprint arXiv:2508.19576*.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30678–30701, Vienna, Austria. Association for Computational Linguistics.

A Ethical Considerations

A.1 Potential Risks

Potential risks associated with our work include the misuse of entity grounding capabilities for surveillance purposes and the possibility of model hallucinations leading to misinformation. We mitigate these risks by using only publicly available datasets and strictly filtering harmful content, but we urge practitioners to exercise caution and respect user privacy when deploying these models in real-world applications.

A.2 Use of LLM

In the preparation of this manuscript, we utilized Large Language Models (LLMs) for grammatical error correction and polishing to improve readability.

A.3 Code and Data

All images and generated text contexts, which we use to train MCR, strictly follow guidelines designed to exclude any harmful, unethical, or offensive content. Furthermore, the data used in MCR does not involve any comparisons of harmful, ethical, or offensive content between image pairs. Our code and curated dataset annotations will be released under an open-source license (e.g., MIT or CC-BY 4.0) upon acceptance, and we strictly adhere to the licensing terms and usage policies of the original datasets (Twitter-GMNER, MNER-MI, GREC) and backbone models used in this work.

B Modality-specific Constraints

Here is one of the instructions used for distillation, which requires the model to consider the relevant modalities during execution and to produce results consistent with the labels.

Modality- and Task- specific Constraints

Grounded Multimodal Named Entity Recognition (GMNER) task requires, given a text and a paired image, **recognize the meaningful and specific entities from the text** that are ... . **Entity type classification is primarily based on textual information**, with **visual cues considered when the text is insufficient to make a confident determination**. When you predict each entity's location, **if the entity appears in the image**, the corresponding location is a bounding box (bbox); **if the entity does not appear in the image**, the corresponding location is None ...
Given a piece of text and its paired image:<image><sentence>.
The ground-truth labels for this image-text pair in the GMNER task are: <label>.
Now **generate the reasoning process that leads from the image-text pair to the ground-truth labels**. The content of the thought process should be your reasoning on how to obtain the true label **based on the image and text input**.

C Multiple Styles Reasoning Schema

We construct diverse reasoning styles and paths using templates, LLMs, and MLLMs, and we design corresponding prompts for each style. We next illustrate the procedure with the GMNER task.

C.1 Instruction

To ensure the model understands the task, we first design a task-introduction instruction and use it across all experiments.

Instruction for GMNER

Here is a Grounded Multimodal Named Entity Recognition task. Given a text and a paired image, You need to identify all entities in the given text, assign each entity a category, and locate the corresponding entities in the image during the entity prediction.

Instructions to format the thought process. To ensure that the generated reasoning is produced in a formatted style constrained by fixed tags, we design two instruction prompts for distinct reasoning routes, and we next present one of the instructions.

Formal Type Prompt for GMNER

To accomplish this task, follow the steps below and place your reasoning and the results of each step inside the <process></process> tags.
1.First, prioritize the textual information; use the image only as a supplement. From the text, identify how many entities are present (zero or more) and list them. Important: do NOT extract entities that appear only in the image but not in the text, and do NOT omit entities that appear in the text but not in the image. Put the number of entities inside <entity_num></entity_num> tags.
2.Second, determine the type of each extracted entity. Use one of these labels: person, organization, location, miscellaneous. Put each entity and its type inside a <mner></mner> tag in the format: (entity text, entity type).
3.Third, decide whether each entity is visible in the image. For entities not visible in the image, place (entity text, invisible) inside an <entailment></entailment> tag. For entities that are visible, place (entity text, visible) inside an <entailment></entailment> tag.
4.Fourth, provide location information: for visible entities, give their bounding box as (x1, y1, x2, y2); for invisible entities, use None. Put each item inside a <location></location> tag in the format:(entity text, (x1, y1, x2, y2)) for visible entities;(entity text, None) for invisible entities.
After completing the steps above, synthesize your findings and produce the final answer of the task inside <answer></answer> tags.

Instructions to output thought process from LLMs. To elicit reasoning in a question–answer or few-conclusion style, we design two instruction prompts for distinct reasoning routes, and we next present one of the instructions. The LLM-augmented reasoning paths are generated using similar prompts.

Conclusion Type Prompt for GMNER

To accomplish this task, you need to follow the reasoning rules below to carry out step-by-step reasoning and accomplish the task objectives.
Goal and Reasoning rules:
1. Extract all special entities that appear in the TEXT (do not invent entities that only appear in the image but not in the text).
2. Assign a type to each special entity from (person, organization, location, miscellaneous). 3. Decide if each special entity is visible in the IMAGE.
4. If visible, provide the bbox coordinates as (x1, y1, x2, y2) of the special entity in the IMAGE; if not visible, use None as location information.
Put your reasoning in <process></process> tags and put the final answer in format of (entity text, entity type, None or bounding box) in <answer></answer>.

Instructions to output distilled thought process. We use the follow instruction to make MLLMs output distilled style reasoning paths.

Distill Type Prompt for GMNER

To accomplish this task, you need to carefully analyze this task step by step from the perspectives of the meaning of special entities, text semantics, image content, whether the special entities appear in the image, and the corresponding location information, and place the thought process within <process></process> tags, and put the triplet results in format of (entity text, entity type, None or bounding box) within <answer></answer> tags.

C.2 Reasoning Styles and Paths

The above diversified instructions yield varied reasoning styles and paths, and we next illustrate two example reasoning processes on a single sample. Different reasoning styles or paths produce different output formats, and even within the same sample, the ordering of entity triples as well as the ordering of the entity, type, and location within each triple can vary.

Thought Process v1

<process> Let's analyze both the text and the image step by step:
1. Identify Entities in the Text: The text mentions "Spurs" and "Premier League".
2. Determine Entity Types: "Spurs" refers to Tottenham Hotspur Football Club, which is an organization. "Premier League" is also an organization, specifically a professional football league.
3. Check Visibility in the Image: In the image, there is a visible logo and text related to "Spurs". The logo of a horse and the word "SPURS" can be clearly seen on the cake box. There is no visible mention or depiction of "Premier League" in the image.
4. Locate Entities in the Image: For "Spurs", the logo and text are present. By examining the image, the bounding box coordinates for "Spurs" can be determined as (405, 216, 558, 324), which encompasses the area where the "Spurs" logo. Since "Premier League" is not visible in the image, its location is marked as None.
5. Compile the Labels:
(Spurs, organization, (405, 216, 558, 324))
(Premier League, organization, None)
</process>

Thought Process v2

<process>
In the text, I've identified "Spurs" as an entity, which appears to be visible in the image at the coordinates (405, 216, 558, 324). Based on my understanding, "Spurs" refers to Tottenham Hotspur Football Club, an organization in the English football league system. I also recognize "Premier League" as an entity in the text, though it doesn't seem to be visually present in the image. The Premier League is the top level of the English football league system.
In total, there are two entities in the text.
</process>

D More Details about Rewards

D.1 Entity Count Rewards

After MRSI, the model tends to conservatively recall entities to avoid visual bias. To increase recall while preventing visual bias from reappearing, we introduce a entity count reward defined by the difference between the predicted and gold entity counts. The penalty is scaled by the true count: harsher penalties are applied when the true count is small, while penalties are more lenient when the true count is large. The reward is formed as:

$$w_o(q) = \begin{cases} 0.4, & 1 \le q \le 2, \\ 0.2, & 3 \le q \le 4, \\ 0.1, & q \ge 5, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

$$w_u(p) = \begin{cases} 0.5, & 0 \le p \le 2, \\ 0.3, & 3 \le p \le 4, \\ 0.2, & p \ge 5, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

$$R_c = \begin{cases} 1, & p = q, \\ \max\big(0,\ 1 - (p-q)w_o(q)\big), & p > q > 0, \\ \max\big(0,\ 1 - (q-p)w_u(p)\big), & 0 < p < q, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

where p and q respectively denote the numbers of predicted and gold entities in a sample, w_o and w_u are the penalty weights for excessive recall and insufficient recall. The reward R_{count} is computed as a function of the relationship between p and q.

D.2 Token-level F1 score

For each predicted entity span $\hat{e}_i = \{\hat{e}_{i,1}, \ldots, \hat{e}_{i,n}\}$ and gold entity span $e_j = \{e_{j,1}, \ldots, e_{j,m}\}$, where $\hat{e}_{i,k}$ and $e_{j,k}$ denote the k-th token in the predicted and gold spans respectively, we first compute the length of their longest contiguous token overlap w_{ij} between them. Then, we define the token-level precision and recall as:

$$P_{ij} = \frac{w_{ij}}{n},\ R_{ij} = \frac{w_{ij}}{m} \tag{16}$$

where n and m respectively denote the numbers of tokens in the predicted and gold spans. The token-level F1 score for this pair is finally computed as:

$$F_{ij} = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}}. \tag{17}$$

Given the token-level F1 matrix $\{F_{ij}\}$ over all predicted–gold span pairs, we apply the Hungarian algorithm to obtain an optimal one-to-one matching between predicted and gold entities. Let $\mathcal{M}$ denote the set of matched index pairs (i, j) and $k = |\mathcal{N}|$ be the number of matched pairs. The entity span reward for a sample is then defined as the average token-level F1 over all matched pairs:

$$R_s = \frac{1}{k} \sum_{(i,j)\in\mathcal{N}} F_{ij}. \tag{18}$$

D.3 Data Preparation

To prevent over-reliance on fixed templates and mitigate training collapse (Xiong et al., 2025) in CVO, we construct a multi-style set of reasoning schema $\mathcal{D}_{\mathcal{R}}$ during MRSI. We then use only a subset $\mathcal{D}_1$ for MRSI, and allocate the remainder $\mathcal{D}_2 = \mathcal{D}_{\mathcal{R}} \backslash \mathcal{D}_1$ to CVO for calibrating and optimizing cross-modal verification on core constraints. To further improve training efficiency and reduce collapse risk, we apply sampling-based filtering to $\mathcal{D}_2$. For each sample's G responses $\{o_1, o_2, \ldots, o_G\}$ with rewards $\{r_1, r_2, \ldots, r_G\}$, we compute the standard deviation, maximum reward, and median reward, and impose preset thresholds on these statistics. Specifically, only samples with a standard deviation of at least 0.1, a maximum group reward of at least 0.8, and a median group reward between 0.08 and 0.6 are retained for training. (Zhoubian et al., 2025).

E More Experiment Details

E.1 Datasets

Our training data are drawn from three sources: Twitter-GMNER (Yu et al., 2023) for GMNER and NER, a multi-image multimodal NER dataset

Dataset	Train	Val	Test
GMNER	7000	1500	1500
MNER-MI	6856	860	860
GREC	14000	5309	19066

Table 7: Dataset statistics used in our experiments.

MNER-MI (Huang et al., 2024), and generalized visual grounding dataset GREC (He et al., 2023). Because GREC includes cases where one textual description corresponds to multiple image regions, which is incompatible with the GMNER setting, we exclude samples in which a single textual description corresponds to multiple image regions. We evaluate Twitter-GMNER in the main experiments, and conduct additional evaluations on MNER-MI and GREC in Section 5.3. As shown in Table 7, we also report the number of raw datasets used during training and evaluation. For GMNER and MNER-MI, we use the full datasets. For GREC, we first filter out multi-target cases, then select 14,000 samples from the remaining data for training, while retaining all remaining validation and test samples. In total, we use 55,712 samples annotated with multi-style reasoning schema for training.

E.2 Baselines

Following prior work (Tang et al., 2025b), we categorize GMNER approaches into unified and pipeline methods. Unified methods use pretrained language models to extract entity–type–location triples in a single pass, while pipeline methods decompose the process into multiple stages handled by different models. Unified approaches reduce error propagation and improve over early pipelines (Wang et al., 2022a; Yu et al., 2023), but recent pipeline methods that incorporate LLMs as knowledge bases achieve substantially better performance (Li et al., 2024; Ok et al., 2024). In contrast to prior unified methods that still rely on auxiliary components and employ LLMs as auxiliary tools, we propose an end-to-end unified approach that uses MLLMs to complete all steps in a single inference.

Pipeline Methods. (1) **ITA-VinVL-EVG** (Wang et al., 2022a) formulates multimodal named entity recognition as an image–text alignment problem. (2) **BARTMNER-VinVL-EVG** (Yu et al., 2023) first uses generative model BART to identify entity type pairs, and then uses the Entity Extraction & Grounding (EEG) model to predict the bounding box for each pair. (3) **Scanner** (Ok et al., 2024) first identifies textual and visual entities using NER and visual grounding models, enriches entity semantics with LLMs and external knowledge bases, and finally matches textual entities to visual locations via a trained module. (4) **UnCo** (Tang et al., 2025c) adopts an uncertainty-aware collaboration between small models and large multimodal language models to refine grounded multimodal named entity recognition predictions. (5) **ReFineG** (Tang et al., 2025a) combines small supervised models and large language models to enhance low-resource grounded multimodal named entity recognition through refinement and knowledge transfer.

Unified Methods. (1) **MNER-QG** (Jia et al., 2023) formulates multimodal named entity recognition as an unified machine reading comprehension task, where entity queries are grounded to both textual context and visual evidence. (2) **H-index** (Yu et al., 2023) formulates GMNER to a sequence generation task with a multimodal BART model. (3) **TIGER** (Wang et al., 2023a) formulates fine-grained named entity recognition and grounding as a sequence generation task by converting entity-type-object triples into target text and employing T5 model to jointly predict entity spans, fine-grained types, and corresponding image objects. (4) MQSPN (Tang et al., 2025b) formulates grounded multimodal named entity recognition as a set prediction problem that employs multi-grained learnable queries to explicitly align textual entities with visual regions.

End-to-end Methods. GLM4.5VL, Qwen2.5VL, and MimoVL are multimodal large language models with strong capabilities in multimodal understanding, reasoning, and visual grounding. We evaluate these models under different prompting and training settings, including direct instruction prompting, Chain-of-Thought (CoT) prompting, and CoT with 3-shot demonstrations. In addition, we include a supervised fine-tuning (SFT) baseline, where the models are fine-tuned on GMNER training data.

E.3 Evaluation Metrics

GMNER. For GMNER and its two subtasks, MNER and EEG, we follow prior work (Yu et al., 2023) and evaluate performance using Precision (**Pre**), Recall (**Rec**), and the **F1** score.

Each sample contains zero or more entity triples $\{(e_i,\ t_i,\ l_i)\}_{i=1}^{k_1}$, and we compute the correctness of each entity triple as follow:

$$correct = \begin{cases} 1, & C_e \wedge C_t \wedge C_l, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

$$C_e/C_t = \begin{cases} 1, & e_i/t_i = \hat{e}_i/\hat{t}_i, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

$$C_l = \begin{cases} 1, & l_i = \hat{l}_i = \text{None}, \\ 1, & IoU(l_i, \hat{l}_i) \geq 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

$$IoU(l_i, \hat{l}_i) = \frac{area(l_i \cap \hat{l}_i)}{area(l_i \cup \hat{l}_i)}, \tag{22}$$

where C_e, C_t and C_l represent the correctness of entity, type and location; e_i, t_i and l_i represent the gold entity, type and location; $\hat{e}_i$, $\hat{t}_i$ and $\hat{l}_i$ represent the predicted entity, type and location; IoU denotes the IoU score between l_i and $\hat{t}_i$; and $area$ refers to the amount of two-dimensional space enclosed by a region. A predicted entity triple is regarded as correct only when the entity, type and location are all correct. Then Precision (Pre), Recall (Rec), and F1 score are used to evaluate the performance:

$$\text{Pre} = \frac{\#correct}{\#predict}, \quad \text{Rec} = \frac{\#correct}{\#gold}, \tag{23}$$

$$\text{F1} = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}}, \tag{24}$$

where $\#correct$, $\#predict$ and $\#gold$ respectively represent the number of triples of correct predictions and gold labels.

VG. Following prior work (He et al., 2023), we use **no-target accuracy** (**N-acc**) to measure localization accuracy when no target entity is present and Precision to measure accuracy when a target entity is present, and use Precision (Pre) to measure one-target entity localization. For a no-target sample, it considered a true positive (TP) when predicted bounding box is None, otherwise false negative (FN). Then N-acc is computed as follow:

$$\text{N-acc} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{25}$$

For a one-target sample, a prediction is counted as a correct localization only if $IoU \geq 0.5$.

Quantitative metrics for textual bias. Inspired by N-acc, we introduce N-Pre, N-Rec, and N-F1 to quantify textual bias in GMNER, particularly cases where the model assigns bounding boxes to entities absent from the image. For an entity triple whose location is None ($l_i = \text{None}$), we determine its correctness as follows:

$$n\text{-}correct = \begin{cases} 1, & C_e \wedge C_n, \\ 0, & \text{otherwise}, \end{cases} \tag{26}$$

$$C_n = \begin{cases} 1, & l_i = \hat{l}_i = \text{None}, \\ 0, & \text{otherwise}, \end{cases} \tag{27}$$

where C_n represent the correctness of no-target entity location. We compute the above metrics over all no-target entity triples:

$$\text{N-Pre} = \frac{\#n\text{-}correct}{\#n\text{-}predict}, \quad \text{N-Rec} = \frac{\#n\text{-}correct}{\#n\text{-}gold}, \tag{28}$$

$$\text{N-F1} = \frac{2 \times \text{N-Pre} \times \text{N-Rec}}{\text{N-Pre} + \text{N-Rec}}, \tag{29}$$

where $\#n\text{-}correct$, $\#n\text{-}predict$ and $\#n\text{-}gold$ respectively represent the number of triples of correct predictions and gold labels.

Quantitative metrics for visual bias. Directly inspecting every test image to quantify how MCR handles visual bias is impractical, so we introduce two indirect metrics that measure image-only entity recall. Based on whether a recalled entity appears in the input sentence, we define **N-Count** as the number of recalled entities that are absent from the sentence, and **N-Rate** as the proportion of such entities among all recalled entities. Specifically, for input sentence s and the model predicted entity triples $\hat{\mathcal{Y}} = \{(\hat{e}_i,\ \hat{t}_i,\ \hat{l}_i)\}_{i=1}^{k_2}$ where k_2 is the number of all entity triples, we compute N-Count and N-Rate as follow:

$$\text{N-Count} = \frac{1}{k_2} \sum_{i}^{k_2} \mathbb{1}\{\hat{e}_i \notin s\}, \tag{30}$$

$$\text{N-Rate} = \frac{\text{N-Count}}{k_2}, \tag{31}$$

where $\mathbb{1}$ denotes an indicator function that returns 1 if the predicted entity is mentioned in the sentence and 0 otherwise. A lower N-Count and N-Rate indicate weaker visual bias, as the model is less likely to hallucinate image-only entities as text mentions.

E.4 Implementation Details

We conduct all experiments on 8 NVIDIA Tesla L20 GPUs. Training and inference use the ms-swift (Zhao et al., 2025) framework, and decoding and sampling use the vLLM (Kwon et al., 2023) engine. All training procedures are conducted using LoRA (Hu et al., 2022).

MRSI. we generate diverse reasoning schema using a combination of template-based extraction, DeepSeek (Guo et al., 2025), Qwen2.5VL-72B, and Qwen3VL-30B-A3B (Bai et al., 2025). We train Qwen2.5VL for 2 epochs and MimoVL for 5 epochs with a learning rate of 0.0001 and cosine learning schedule. We use a batch size of 16 and train the model for 8 hours on 4 L20 GPUs.

CVO. During the CVO phase, we train for 2 epochs with a learning rate of 0.000005 and a batch size of 64. We use a warmup ratio of 0.05, sample 8 generations per input, set GRPO clipping thresholds to 0.15 and 0.25, and apply temperature 1.5 with top-k sampling ($k = 200$), top-p sampling ($p = 0.95$), and $\beta = 0.005$. We trained the model for 11 hours on 4 L20 GPUs.

E.5 Training and Inference Cost

Additional Overhead of MRSI and CVO over SFT. To better quantify the computational overhead introduced by MRSI and CVO, we report both training and inference costs under identical settings using vLLM on 4 NVIDIA Tesla L20 GPUs for Qwen2.5VL-7B. As shown in Table 8, both MRSI and CVO incur additional training and inference time compared to the SFT baseline. However, this increased cost translates into substantial performance gains: MRSI improves F1 by 5.0%, and further applying CVO reaching a total gain of 7.0% over SFT. Notably, the inference time remains the same for MRSI and CVO, indicating that the extra training cost of CVO does not affect deployment efficiency.

Methods	Inference (min)	Training (h)	GMNER
SFT_{base}	3	6	63.6
MRSI	6	8	68.6 (+5.0)
CVO	6	11	70.6 (+7.0)

Table 8: **Training and inference cost introduced by MRSI and CVO. SFT_{base}** denotes basic SFT, and **GMNER** denotes the F1 score on GMNER.

Additional Overhead from MRSI to CVO. Under the same post-MRSI setting, we compare continued MRSI and CVO on 4 NVIDIA Tesla L20 GPUs. Specifically, $\textbf{MRSI}_{\textbf{data}}$ continues supervised training with the same amount of data as CVO, while $\textbf{MRSI}_{\textbf{time}}$ further extends MRSI until reaching the same training time as CVO. As shown in Table 9, CVO consistently yields much larger gains than both variants of continued MRSI. These results show that simply scaling MRSI, either by using the same amount of data or by matching the training time of CVO, brings only limited improvement. In contrast, CVO provides a substantially stronger optimization signal and more effectively unlocks the model's reasoning potential.

Model	Method	Imp (%)
Qwen2.5VL-7B	$\text{MRSI}_{\text{data}}$	0.2
	$\text{MRSI}_{\text{time}}$	0.4
	CVO	2.0
MimoVL-7B	$\text{MRSI}_{\text{data}}$	0.2
	$\text{MRSI}_{\text{time}}$	0.5
	CVO	3.8

Table 9: **Performance improvement of different training strategies. $\text{MRSI}_{\text{data}}$** uses the same amount of training data as CVO, while $\textbf{MRSI}_{\textbf{time}}$ extends MRSI to the same training time as CVO. **Imp (%)** denotes the F1 gain on GMNER over MRSI.

F Sensitivity Analysis

We further conduct sensitivity analyses on the key hyperparameters in our constraints, including the IoU threshold σ, the reward weight for the number of entities λ_1, and the reward weight for entailment relations λ_5. As shown in Table 10, varying these thresholds or reward weights leads to only minor fluctuations in F1, all within a 1% range. Peak performance is generally achieved at moderate values, while more extreme settings still remain competitive. These results indicate that our method is robust to the specific choice of these hyperparameters.

G Robust generalization and reduced error propagation

To evaluate the generalization ability of MCR and its effectiveness in reducing error propagation compared with pipeline methods, we compare it with several strong pipeline baselines on

Param.	Value	Pre	Rec	F1
σ	0.0	69.6	69.9	69.8
	0.1	69.7	69.8	69.8
	0.2	69.8	70.3	70.0
	0.3	70.5	70.8	70.6
	0.4	69.5	69.8	69.7
	0.5	69.6	69.8	69.7
λ_1	0.0	68.1	70.7	69.3
	0.1	68.1	70.4	69.2
	0.2	69.0	70.8	69.9
	0.25	70.1	70.3	70.2
	0.3	70.5	70.8	70.6
	0.35	70.0	70.2	70.1
	0.4	69.7	69.8	69.8
	0.5	69.2	69.6	69.4
λ_5	0.0	68.9	69.4	69.1
	0.1	69.5	69.8	69.6
	0.2	70.5	70.8	70.6
	0.3	69.8	70.1	69.9
	0.4	69.5	69.8	69.7
	0.5	68.6	70.4	69.5

Table 10: Sensitivity analysis over different hyperparameter values.

Twitter-GMNER and Twitter-FMNERG (Wang et al., 2023b). Twitter-FMNERG extends Twitter-GMNER with fine-grained entity categories. For fair comparison, we remove the knowledge enhancement component from pipeline methods that use it, since MCR and several other baselines do not. As shown in Table 11, MCR also performs strongly on FMNERG, demonstrating good generalization. Moreover, MCR consistently outperforms the pipeline methods, with especially large gains on the EEG subtask. This highlights that the end-to-end MLLM-based design of MCR more effectively reduces error accumulation across stages, particularly in the grounding step that maps extracted entities to image regions.

Methods	Twitter-GMNER			Twitter-FMNERG		
	GMNER	MNER	EEG	GMNER	MNER	EEG
UnCo (Tang et al., 2025c)	64.6	81.7	69.6	53.6	67.7	68.3
MAKAR (Lin et al., 2025)	67.6	-	-	57.5	-	-
MVP (Liu et al., 2024a)	-	-	-	51.5	67.7	65.2
GEM (Wang et al., 2024b)	61.5	81.1	64.5	52.5	67.2	65.5
RiVEG (Li et al., 2024)	64.9	82.9	68.3	-	-	-
MCR(ours)	70.6	82.8	73.4	59.9	70.4	72.3

Table 11: **Comparison on Twitter-GMNER and Twitter-FMNERG across GMNER, MNER and EEG in terms of F1 scores.**

H Case Study

MCR effectively mitigates modality bias. As illustrated in Figure 7, (a) and (b) present two cases where MCR successfully mitigates textual bias. Naive End-to-end methods lead the model to assign incorrect image regions to textual entities. For example, the model incorrectly grounds the "NBA" logo to the textual entity "NFL", and assigns an elderly male to "Donald Trump".

By explicitly generating reasoning paths and reinforcing cross-modal consistency verification, MCR effectively alleviates these issues. Specifically, the model recognizes the distinct semantics of the "NBA" logo in the image and the "NFL" entity in the text and correctly concludes that they do not match. Similarly, it explicitly verifies whether the person in the image corresponds to "Donald Trump", thereby avoiding erroneous grounding.

(c) and (d) present two cases where MCR successfully mitigates visual bias. When MLLMs perform entity recognition and classification, they can be distracted by irrelevant visual elements, leading to the spurious recall of image-only entities or incorrect classification of textual entities. For instance, the model erroneously recalls the image-only entity "NBA", and misclassifies the human entity "Rory Calhoun" due to the presence of a cat in the image.

By explicitly prompting the model to surface multimodal evidence and reinforcing the principled use of such evidence, MCR effectively alleviates these issues. Specifically, the model clarifies the modality source of each entity and relies on internal knowledge and sentence semantics when determining entity types, rather than being misled by superficial visual cues.

Limitations of knowledge and entity span. As illustrated in Figure 8, (a) and (b) present two failure cases of MCR. Although MLLMs incorporate substantial knowledge during training, GMNER requires broad, cross-domain knowledge that inevitably includes entities beyond the model's coverage or cases where the model has acquired incorrect knowledge. As a result, MLLMs may still fail in such scenarios. For example, even though MCR possesses knowledge about "Lady Gaga", its limited visual knowledge of her appearance causes it to be misled by visually similar cues. Similarly, due to the lack of prior knowledge about "Ay Ziggy Zoomba", the model makes an

incorrect judgment from the outset. These cases illustrate that MCR remains constrained by the underlying model's knowledge coverage and visual familiarity. Figure 8(b) further shows that MLLMs still struggle with entity span detection.

Performance on unimodal-dominant cases. In GMNER, there exist unimodal-dominant cases where entity type prediction is independent of visual evidence, and the image may even act as a source of noise. In such cases, the cross-modal consistency mechanism in MCR performs cross-modal verification instead of enforcing strict text-image alignment. As described in Appendix C, MCR explicitly encourages the model to rely on linguistic input when visual grounding evidence is unnecessary or irrelevant. To evaluate MCR under this setting, we isolate unimodal-dominant samples from GMNER and analyze their entity type prediction errors. As shown in Table 12, MCR consistently yields lower error rates than SFT and larger models, indicating its effectiveness in reducing the impact of irrelevant visual cues in unimodal-dominant contexts.

Model	Method	Rate (%)	Count
Qwen2.5VL-72B	Direct Prompt	4.4	115
GLM4.5VL	Direct Prompt	5.3	138
Qwen2.5VL-7B	Direct Prompt	4.2	111
	SFT	3.0	78
	MCR(**ours**)	**2.7**	**71**
MimoVL-7B	Direct Prompt	4.9	128
	CoT+3-Shot	3.7	97
	MCR(**ours**)	**3.2**	**85**

Table 12: **Error analysis on unimodal-dominant cases. Rate** means the percentage of errors, and **Count** means the number of errors.

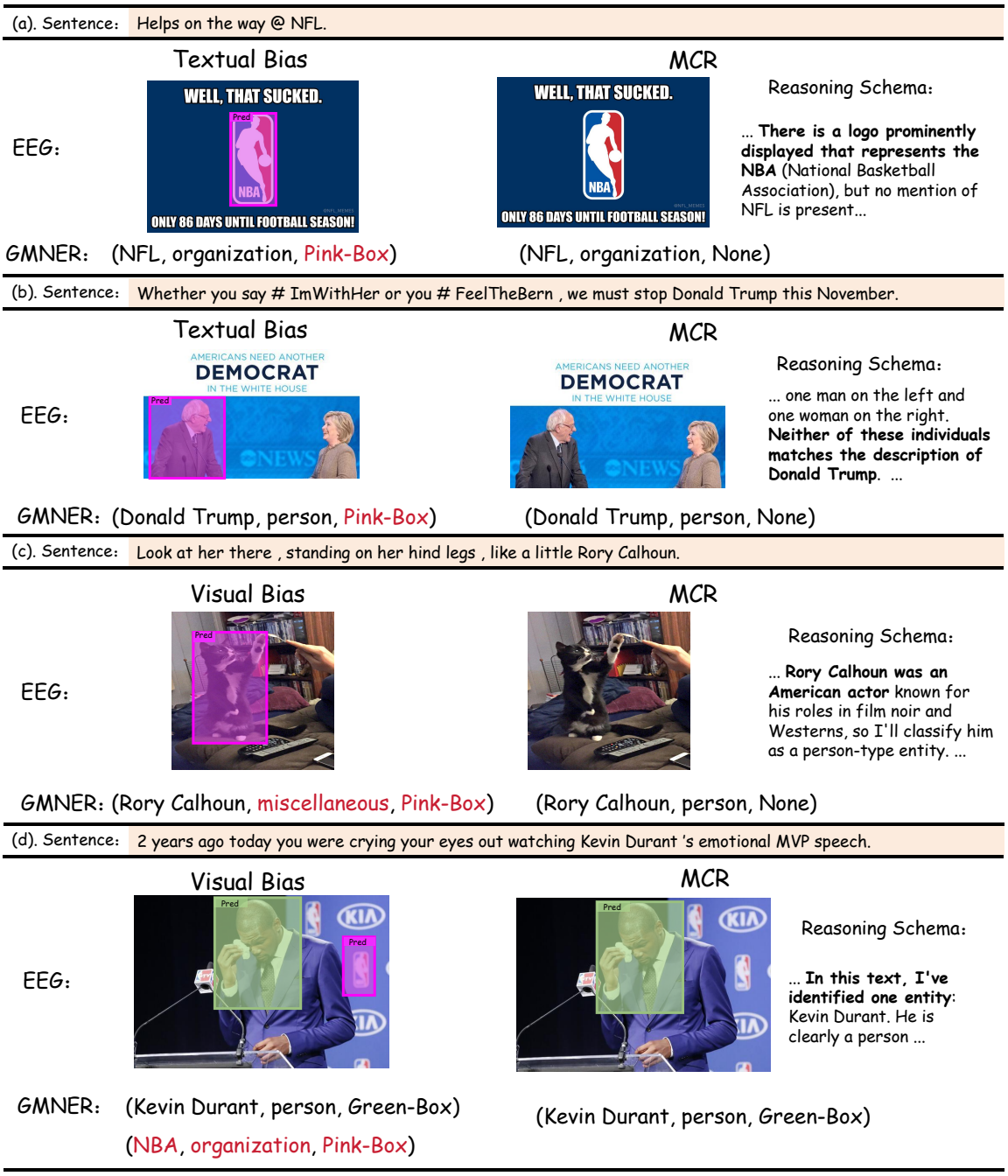

Figure 7: Case Studies of MCR Mitigating Modality Bias.

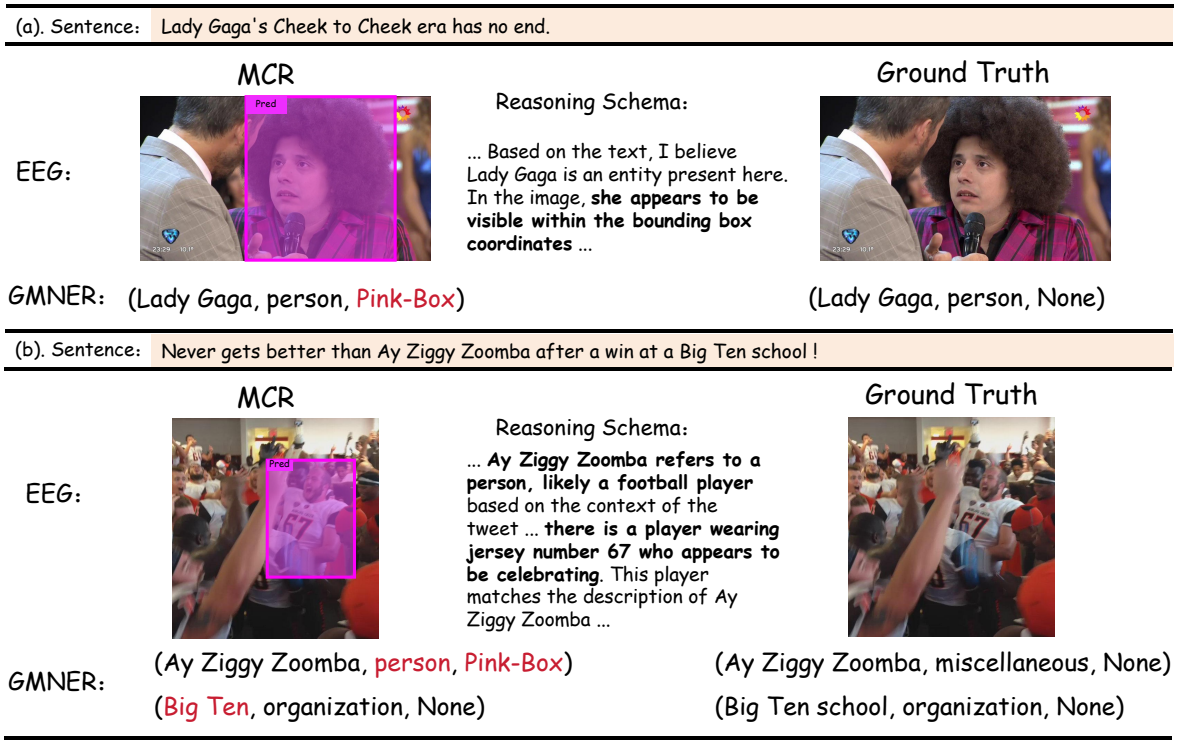

Figure 8: Failure Cases of MCR.