CORDIAL: Can Multimodal Large Language Models Effectively Understand Coherence Relationships?

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) are renowned for their superior instruction-following and reasoning capabilities across diverse problem domains. However, existing benchmarks primarily focus on assessing factual and logical correctness in downstream tasks, with limited emphasis on evaluating MLLMs' ability to interpret pragmatic cues and intermodal relationships. To address this gap, we assess the competency of MLLMs in performing *Multimodal Discourse Analysis* (MDA) using Coherence Relations. Our benchmark, CORDIAL, encompasses a broad spectrum of Coherence Relations across 3 different discourse domains at varying levels of granularity. Through our experiments on 10+ MLLMs employing different prompting strategies, we show that even top models like Gemini 1.5 Pro and GPT-4o fail to match the performance of simple classifier-based baselines. This study emphasizes the need to move beyond similarity-based metrics and adopt a discourse-driven framework for evaluating MLLMs, providing a more nuanced assessment of their capabilities. The benchmark and evaluation code will be released upon publication.

1 Introduction

The recent advancements in Multimodal Large Language Models (MLLMs) enable them to effectively capture diverse representations of problem domains (Alayrac et al., 2022; Chen et al., 2024c; Pichai, 2024; Liu et al., 2024a). These MLLMs are capable of adapting to various downstream tasks with limited data through Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2021) and In-Context Learning (ICL) (Brown et al., 2020) approaches. Existing Vision-based MLLM benchmarks assess different aspects of model performance such as Perception, Cognition, and Reasoning (Li et al., 2024) through various downstream tasks.

Current benchmark design strategies often focus on evaluating the ability of MLLMs to utilize the intersection of input sources to solve a common problem (Kruk et al., 2019). Although this helps assess the model's ability to interpret its inputs factually and logically, it does *not fully capture the model's understanding of the relationships between these modalities*. Similarly, benchmarks that evaluate the alignment between images and text (Thrush et al., 2022), utilize curated or synthetically generated image-text pairs. These methods focus solely on literal relations that measure the level of overlap between the image and text. On the other hand, pragmatic cues provide information on non-literal relations where the true intent/message of an example may not be directly referenced in both modalities as shown in Figure 1. These cues are leveraged routinely in real-world multimodal discourses, which are characterized by the use of multiple modes of communication to convey different components of a message. Multimodal Discourse Analysis (MDA) studies how the interaction between these different modes can create semiotic meaning (Kress, 2009).

To operationalize the assessment of these intermodal relationships, we turn to theories of *Discourse Coherence* (Hobbs, 1978), which offer a way to quantify the organization and flow of ideas across information sources. From these theories, we focus on the concept of *Coherence Relations* (Alikhani and Stone, 2019), which provides a finite structure to link different parts of a discourse. Recent studies have extended these traditionally text-only theories to multimodal discourses, showing that Coherence Relations can be effectively applied to image-text pairs (Alikhani et al., 2020). With Coherence Relations being a fundamental aspect of human communication, we evaluate whether MLLMs can effectively predict and verify these relations.

In this work, we propose the CORDIAL (COherence Relations in Discourse for Images

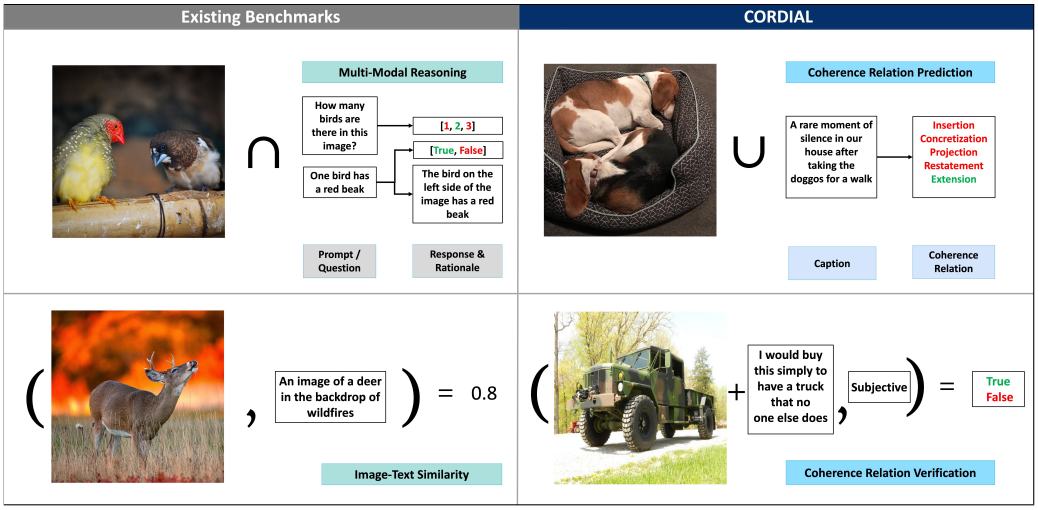


Figure 1: CORDIAL presents a combination of literal and pragmatic relations for analyzing the intermodal reasoning capabilities of MLLMs. We evaluate MLLMs on the task of Multimodal Discourse Analysis through the prediction and verification of Coherence Relations across three different discourse domains.

And Language), the first benchmark for evaluating MLLMs on the task of MDA. CORDIAL consists of a diverse set of Coherence Relations across three different discourse domains: Disaster Management, Social Media, and Online Articles. Each domain also offers different levels of complexity in the evaluated Coherence Relations, from binary relations to more challenging settings such as multi-class and multi-label relations assigned by human annotators. We evaluate the performance of 10+ MLLMs on CORDIAL, focusing on three research questions:

RQ1: Can MLLMs predict Coherence Relations effectively?

RQ2: Can MLLMs verify Coherence Relations accurately?

RQ3: Can we teach MLLMs to understand Coherence Relations better?

Our analysis reveals that both Coherence Relation prediction (RQ1) and verification (RQ2) are challenging tasks for MLLMs when these relations focus on pragmatic cues. Although larger MLLMs perform better than their smaller, open-source counterparts, traditional classifier baselines consistently outperform them across discourse domains. To summarize, our key takeaways are as follows:

- We propose CORDIAL, the first benchmark for evaluating MLLMs for Multi-modal Discourse Analysis (MDA) using Coherence Relations.
- Our experiments show that MLLMs struggle to predict and verify Coherence Relations, especially when these relations are more pragmatic.
- We demonstrate the need for coherence-aware fine-tuning approaches to improve intermodal reasoning capabilities of MLLMs.

2 Related Work

Multimodal Large Language Models MLLMs are fundamentally generative models that combine Large Language Models (LLM) (Brown et al., 2020) with multimodal encoders (Dosovitskiy et al., 2021). In recent years, several new MLLMs have been released, based on various proprietary (OpenAI et al., 2024; Anthropic; Pichai, 2024) and open-source LLM backbones (Liu et al., 2023; Wu et al., 2024; Bai et al., 2023). These models have shown impressive performance on a variety of downstream reasoning tasks, including Visual Question Answering (Wu and Xie, 2024), Document Analysis (Lv et al., 2023), Embodied AI agents (Shek et al., 2024), etc.

MLLM Reasoning Benchmarks Recent works that have proposed benchmarks evaluating vision language reasoning, focus on assessing different facets of their input modalities. Visual Reasoning benchmarks measure the capability of these models to understand spatial and object-level relations among image components (Kamath et al., 2023; Rajabi and Kosecka, 2024; Nie et al., 2024; Thrush et al., 2022; Kamoi et al., 2024). Contextual Reasoning benchmarks demonstrate how MLLMs interpret in-context examples and compositional language prompts (Zong et al., 2024; Wu and Xie,

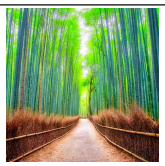
Dataset	Examples				
DisREL	 Part of my pile of branches after #HurricaneIrma - still no power in #Orlando **Coherence Relation:** Similar			 Floridians rescue stranded manatees as Irma sucks water from shores **Coherence Relation:** Complementary	
Tweet Subtitles	 Fresh never frozen jumbo wings tossed in a housemade buffalo sauce. Yum! **Coherence Relation:** Concretization	 Freshly picked off my allotment today, well chuffed. (strawberry) **Coherence Relation:** Insertion	 Cartel leader whose arrest sparked killings is sentenced to prison in Dallas court **Coherence Relation:** Projection	 Amazon Prime delivers anything these days! (delivering a cat) **Coherence Relation:** Extension	 Eiffel Tower shuts down as snow, freezing rain pummel France **Coherence Relation:** Restatement
CLUE	 A path winds through an ancient bamboo forest **Coherence Relations:** Visible	 A model walks the runway for the collection during, fashion week **Coherence Relations:** Visible, Meta, Action	 A city in winter is such a beautiful city **Coherence Relations:** Subjective, Story	 People know that curb appeal is not a thing to take lightly when remodeling a home **Coherence Relations:** Story	 Seals fighting for a spot to sleep on the rocks **Coherence Relations:** Action

Table 1: Examples from each dataset for all Coherence Relations. The words in red are important cues present in the caption, while the words in orange show pragmatic cues inferred from the image-text pair. The relations highlighted in blue are the selected relations for CLUE Single-Label.

2024; Shao et al., 2024; Zeng et al., 2024). Finally, Knowledge-based reasoning assesses how models recall knowledge from intrinsic and extrinsic sources to answer factual and logical questions (Johnson et al., 2016; Xenos et al., 2023; Lu et al., 2022). Although these benchmarks measure how multimodal prompts can be efficiently understood to solve a candidate task, intermodal reasoning with real-world discourses has been less studied.

Image-Text Relationships Quantifying image-text relationships accurately has been an active area of research in the era of Vision Language Models (VLMs). Traditional VLMs translate images and text into a common representation space and compute the degree of similarity based on the distance between these embeddings (Radford et al., 2021; Jia et al., 2021; Caron et al., 2021; Hessel et al., 2021). However, these methods failed to capture human preferences in image-text matching accurately across different task domain benchmarks (Anantha Ramakrishnan et al., 2024b; Ross et al., 2024; Anantha Ramakrishnan et al., 2024a). To include human feedback in the process of predicting similarity scores, content-based models trained on human-annotated similarity scores were introduced (Wu et al., 2023; Kirstain et al., 2023; Xu et al., 2023). Apart from similarity scores, taxonomies have been proposed to quantify different types of linkages between image-text pairs (Marsh and White, 2003; Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019; Bateman, 2014). In particular, multimodal coherence relations have been shown to sufficiently capture different aspects of image-text intents for various vision-language tasks (Alikhani et al., 2019; Inan et al., 2021; Alikhani et al., 2023, 2020; Xu et al., 2022).

3 The CORDIAL Benchmark

3.1 Motivation

With Coherence Relations providing a finite representation of image-text linkages, we aim to measure MLLM performance through relation classification and verification tasks. Traditional alignment benchmarks often evaluate models using similarity scores. But multiple states of alignment between image-text pairs can exist, at the object-level,

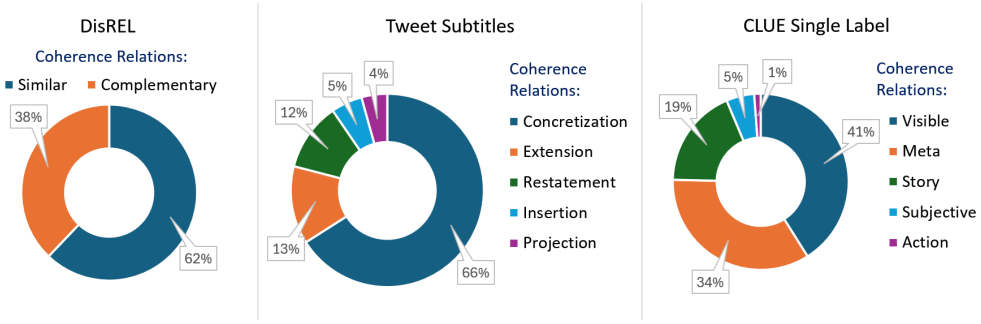


Figure 2: An overview of the Image-Text label (i.e., Coherence Relations) distributions across CORDIAL

scene-level, or even at the discourse-level (Xu et al., 2022). A pragmatic understanding of the context surrounding these pairs informs our ability to describe this alignment accurately. Thus, similarity scores alone may not be sufficient to capture the true performance of MLLMs. Additionally, with Coherence Relations being context-driven, the type of relations present in a discourse can vary across different domains. This necessitates the evaluation of MLLMs on multiple real-world discourse domains to assess their generalization capabilities. With MLLMs-as-a-judge (Chen et al., 2024a) becoming more popular in tasks where acquiring human judgment is expensive and time-consuming, the importance of this task is further highlighted. We carefully pick and curate *real-world image-text pairs* with *expert human annotations* with the preprocessing details described in Appendix Section A. The three different discourse domains we evaluate are: Disaster Management, Social Media, and Online Articles.

3.2 Coherence Relations

Each dataset we include in CORDIAL assesses a unique set of Coherence Relations. To understand how communication in a discourse can be quantified by Coherence Relations, we turn to the Theory of Coherence (Hobbs, 1978). We define communication as the transfer of information and ideas from a speaker to a listener. For successful communication, a discourse needs to satisfy 4 conditions: (1) The message contents should be present in the discourse (2) The message must be relevant to the overall context of the discourse (3) Any new/unpredictable attributes of the message must build on the listener's existing world knowledge (4) The speaker must provide cues to guide the listener to graph their intended meaning. The goal of defining Coherence Relations is to serve any of the above-mentioned communicative functions. This way, for tasks such as MDA, we can analyze the communicative patterns present in a multimodal discourse. We consider Coherence Relations to be a constrained set of connections that describe the structural and causal relationships between different parts of a discourse. Consider the examples from Table 1, certain relations such as Visible and Concretization deal with presenting the same message content across modalities. On the other hand, relations such as Insertion and Extension require the reader to understand the union of information along with the context surrounding each modality to get the full message.

3.3 Data Sources

To construct our benchmark, we leverage existing datasets that provide image-text pairs along with human-annotated Coherence Relations across different discourse domains. We select three datasets that offer a diverse set of Coherence Relations: DisRel (Disaster Management), Tweet Subtitles (Social Media), and CLUE (Online Articles).

DisRel This dataset (Sosea et al., 2021) explores the relationship of image-text pairs from disaster-related tweets, with labels collected through crowdsourcing on Amazon MTurk. The dataset contains 4600 multimodal tweets with a test set size of *500 examples* with a 50% split between the two classes:

- **Similar**: The image and text share the same focus and attempt to convey the same message. There exists a significant overlap in the information conveyed between modalities.
- **Complementary**: The image and text do not share the same focus, but one modality helps understand the other better. Both modalities provide independent information which when combined, provide a more complete picture of the

message/event. There may be divergence in the information conveyed between modalities.

Tweet Subtitles To measure cross-modal coherence relations between image and text, this dataset (Xu et al., 2022) contains 16000 image-text pairs sourced from Twitter on open-domain topics. The test set for this dataset consists of *1600 examples*, which is 10% of the entire dataset. The dataset provides single-label annotations from expert annotators on 3 entity-level and 2 scene-level relations:

- **Insertion (Entity-level)**: Both the text and the image focus on the same visual entity but it is not explicitly mentioned in the text.
- **Concretization (Entity-level)**: Both the text and image contain a mention of the main visual entity but may differ in types of details shared.
- **Projection (Entity-level)**: The main entity mentioned in the text is implicitly related to the visual objects present in the image. The image contains a reference to objects related to the main entity rather than the entity itself.
- **Restatement (Scene-level)**: The text directly describes the image contents. Both modalities convey the same message.
- **Extension (Scene-level)**: The image expands upon the story or idea in the text, presenting new elements or elaborations, effectively filling in narrative gaps left by the text.

CLUE This dataset presents a novel conceptualization of image-text relations by extending text-only coherence relations to the multimodal setting (Alikhani et al., 2020). The publicly available version of the dataset contains 4770 image-text pairs sourced from the Conceptual Captions Dataset (Sharma et al., 2018). The samples were provided multi-label annotations by expert annotators for 5 different relationship types:

- **Visible**: The text presents information that is intended to recognizably characterize what is depicted in the image.
- **Action**: The text describes an extended, dynamic process in which the moment captured in the image is a representative snapshot.
- **Meta**: The text allows the reader to draw inferences not just about the scene depicted in the image but about the production and presentation of the image itself.
- **Subjective**: The text provides information about the speaker's reaction to, or evaluation of, what is depicted in the image.
- **Story**: The text provides a freestanding description of the circumstances depicted in the image, analogous to including instructional, explanatory, and other background relations.

We evaluate this dataset in two different settings: Multi-Label (ML) and Single-Label (SL). In the ML setting, we treat the dataset as a multi-label classification task where MLLMs predict all applicable labels. For CLUE SL, we follow the original dataset's label mapping strategy to select the most applicable label from the present annotations for each sample (Alikhani et al., 2020). This provides two different settings for evaluating MLLM's understanding of coherence relations on the same image-text pairs with *1183 examples* in the test set.

3.4 Baseline Classifier

Our goal of including a baseline classifier is to capture the existing signal in our datasets and to provide a reference point for MLLM performance. Understanding that human annotations can be noisy, we utilize this simple, generalizable classifier to identify relations where MLLMs are particularly under-performing on our benchmark. We employ CLIP Text and Image encoders to extract multimodal embeddings in a zero-shot manner (Radford et al., 2021). We then train a Multi-Layer Perceptron (MLP) classifier using these embeddings on the train sets of each of these datasets to predict Coherence Relations. This ensures that our classifier is not biased towards any specific domain and can generalize across different discourse contexts. More details about the classifier are present in Appendix Section F.

4 Experiments

To answer our research questions, we conduct experiments on the CORDIAL benchmark with top open-source and proprietary MLLMs. For (RQ1), we evaluate the performance of 12 MLLMs from 9 different model families across our benchmark along with a classifier baseline. The 4 settings in our benchmark are structured with increasing difficulty, with DisRel and Tweet Subtitles being the simpler settings while CLUE Single-Label (SL) and CLUE Multi-Label (ML) are more complex. To answer (RQ2), we pick a selection of MLLMs

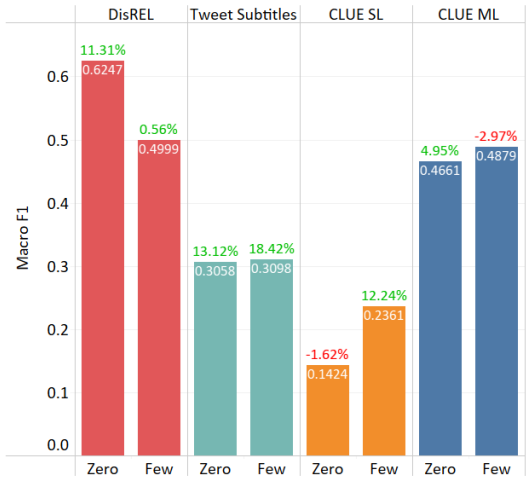


Figure 3: % Loss/Gain after fine-tuning Llama 3.2-V. Fine-tuning shows significant performance gains, either on zero-shot or few-shot prompts across all 4 settings

and investigate their ability to verify coherence relations as correct or incorrect when provided along with image-text pairs. This provides a measure of the model's grasp of concepts such as discourse coherence and intermodal reasoning. For understanding (RQ3), we evaluate the effectiveness of different prompting strategies in enabling these MLLMs to discern coherence relations. We also fine-tune an MLLM on our benchmark to see if it can enhance its intermodal reasoning capability.

4.1 Models Evaluated

We evaluate **4 proprietary MLLMs**: GPT-4o (OpenAI et al., 2024), Gemini 1.5 Flash (Pichai, 2024), Gemini 1.5 Pro (Pichai, 2024), and Claude 3.5 Sonnet v2 (Anthropic) and **8 open-source MLLMs:** LLaVA 1.6 (7B, 13B, 34B) (Liu et al., 2024b), LLaVA OneVision 7B (Li et al., 2025), Qwen2-VL-7B (Wang et al., 2024), Llama 3.2 11B Instruct (Meta AI), Phi3.5 Vision Instruct (Abdin et al., 2024), and InternVL 2.5 26B (Chen et al., 2024b). We selected these model families as they demonstrated acceptable prompt adherence as described in Appendix Sections B, C. We also include a pre-trained classifier fine-tuned for the task of coherence relation prediction. We selected GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet v2 as they were among the better-performing MLLMs on our benchmark for verification, with more details provided in Appendix Section D.

4.2 Evaluation Metrics

On the task of coherence relation prediction, we report the per-class F1 score and overall F1 score

Model	Prompt	Sim	Compl	Macro F1
Random Guess	Baseline	0.490	0.478	0.484
LLaVA 1.6 7B	Zero	0.253	0.541	0.397
	CoT	0.544	0.489	0.516 ↑30.0%
LLaVA 1.6 13B	Zero	0.666	0.000	0.333
	CoT	0.408	0.675	0.542 ↑62.8%
LLaVA 1.6 34B	Zero	0.000	0.666	0.333
	Few	0.139	0.679	0.409 ↑22.8%
	CoT	0.353	0.571	0.462 ↑38.7%
LLaVA OneVision 7B	Zero	0.626	0.391	0.509
	Few	0.549	0.541	0.545 ↑7.1%
	CoT	0.549	0.601	0.575 ↑13.0%
Qwen2-VL 7B	Zero	0.654	0.268	0.461
	Few	0.664	0.148	0.406 ↓11.9%
	CoT	0.446	0.602	0.524 ↑13.7%
Llama 3.2 Vision 11B	Zero	0.388	0.635	0.512
	Few	0.509	0.479	0.494 ↓3.5%
	CoT	0.292	0.615	0.453 ↓11.5%
Phi3.5 Vision 4.2B	Zero	0.655	0.177	0.416
	Few	0.409	0.662	0.536 ↑28.8%
	CoT	0.549	0.601	0.575 ↑38.2%
InternVL 2.5 26B	Zero	0.618	0.698	0.658
	Few	0.633	0.633	0.633 ↓3.8%
	CoT	0.393	0.670	0.531 ↓19.3%
GPT-4o	Zero	0.025	0.667	0.346
	Few	0.443	0.667	0.555 ↑60.4%
	CoT	0.361	0.676	0.519 ↑50.0%
Gemini 1.5 Flash	Zero	0.714	0.715	0.715
	Few	0.363	0.688	0.525 ↓26.6%
	CoT	0.593	0.699	0.646 ↓9.7%
Gemini 1.5 Pro	Zero	0.719	0.679	0.699
	Few	0.611	**0.727**	0.669 ↓4.3%
	CoT	0.630	0.717	0.673 ↓3.7%
Claude 3.5 Sonnet v2	Zero	0.722	0.615	0.669
	Few	0.710	0.559	0.634 ↓5.2%
	CoT	0.603	0.703	0.653 ↓2.4%
CLIP Classifier	Baseline	**0.750**	0.715	**0.733**

Table 2: Results for Coherence Relation Prediction on DisRel. The coherence relations predicted are Similar (Sim) and Complementary (Compl).

Dataset	CR	Claude	Gemini	GPT4o
DisREL	Similar	70.4%	57.2%	14.8%
	Complementary	91.2%	10.8%	96.8%
	Overall	**80.8%**	34.0%	55.8%
Tweet Subtitles	Insertion	20.59%	0.0%	11.76%
	Concretization	74.1%	57.35%	37.61%
	Projection	81.82%	0.0%	15.91%
	Restatement	65.73%	64.34%	21.68%
	Extension	66.29%	0.0%	38.29%
	Overall	**70.44%**	47.69%	34.56%
CLUE SL	Visible	83.37%	90.21%	75.4%
	Subjective	58.0%	20.0%	52.0%
	Action	72.73%	9.09%	54.55%
	Story	29.12%	3.85%	35.71%
	Meta	9.98%	0.0%	0.8%
	Overall	**42.77%**	35.0%	36.52%
CLUE ML	Overall	**48.82%**	32.71%	44.21%

Table 3: Accuracy of MLLMs in verifying each Coherence Relation (CR) of every dataset.

across all 4 settings. We select Macro F1 for overall performance as it treats all classes equally, which is important for our benchmark as it contains imbalanced classes. We report response accuracy for measuring performance on the verification task.

Model	Prompt	Ins	Concr	Proj	Restmt	Ext	Macro F1
Random Guess	Baseline	0.094	0.340	0.068	0.123	0.165	0.158
LLaVA 1.6 7B	Zero	0.000	0.693	0.062	0.066	0.082	0.181
	CoT	0.019	0.822	0.081	0.050	0.114	0.217 ↑19.9%
LLaVA 1.6 13B	Zero	0.085	0.044	0.000	0.000	0.095	0.045
	CoT	0.070	0.477	0.000	0.122	0.054	0.145 ↑222.2%
LLaVA 1.6 34B	Zero	0.000	0.176	0.094	0.104	0.253	0.125
	Few	0.026	0.630	0.198	0.060	0.211	0.225 ↑80.0%
	CoT	0.024	0.063	0.108	0.154	0.169	0.104 ↓16.8%
LLaVA OneVision 7B	Zero	0.023	0.000	0.066	0.125	0.032	0.049
	Few	0.067	0.000	0.087	0.071	0.177	0.081 ↑65.3%
	CoT	0.062	0.005	0.057	0.124	0.101	0.070 ↑42.9%
Qwen2-VL 7B	Zero	0.000	0.728	0.121	0.142	0.011	0.201
	Few	0.094	0.148	0.078	0.144	0.068	0.106 ↓47.3%
	CoT	0.156	0.167	0.068	0.170	0.000	0.112 ↓44.3%
Llama 3.2 Vision 11B	Zero	0.000	0.779	0.000	0.093	0.000	0.175
	Few	0.035	0.388	0.000	0.092	0.113	0.126 ↓28.0%
	CoT	0.097	0.421	0.055	0.167	0.086	0.165 ↓5.7%
Phi3.5 Vision 4.2B	Zero	0.043	0.790	0.109	0.171	0.030	0.229
	Few	0.183	0.179	0.000	0.159	0.093	0.123 ↓46.3%
	CoT	0.025	0.745	0.164	0.156	0.022	0.223 ↓2.6%
InternVL 2.5 26B	Zero	0.101	0.389	0.090	0.090	0.011	0.136
	Few	0.090	0.002	0.041	0.292	0.000	0.085 ↓37.5%
	CoT	0.118	0.450	0.102	0.199	0.083	0.190 ↑39.7%
GPT-4o	Zero	0.126	0.564	0.111	0.200	0.167	0.234
	Few	0.171	0.599	0.131	0.268	0.199	0.274 ↑17.1%
	CoT	0.076	0.346	0.146	0.217	0.187	0.194 ↓17.1%
Gemini 1.5 Flash	Zero	0.172	0.783	0.138	0.183	0.011	0.257
	Few	0.027	0.681	0.139	0.257	0.193	0.259 ↑0.8%
	CoT	0.068	0.734	0.133	0.259	0.071	0.253 ↓1.6%
Gemini 1.5 Pro	Zero	0.200	0.692	0.141	0.290	0.034	0.271
	Few	0.113	0.661	0.247	0.270	0.000	0.258 ↓4.8%
	CoT	0.102	0.657	0.101	0.278	0.022	0.232 ↓14.4%
Claude 3.5 Sonnet v2	Zero	0.132	0.764	0.183	0.328	0.175	0.316
	Few	0.144	0.567	0.122	0.285	0.246	0.273 ↓13.6%
	CoT	0.180	0.725	0.138	0.316	0.256	0.323 ↑2.2%
CLIP Classifier	Baseline	**0.542**	**0.866**	**0.286**	**0.388**	**0.514**	**0.519**

Table 4: Results for Coherence Relation Prediction on Tweet Subtitles. The Coherence Relations predicted are Insertion (Ins), Concretization (Concr), Projection (Proj), Restatement (Restmt) and Extension (Ext).

Model	Prompt	Visible	Subj	Action	Story	Meta	Macro F1
Random Guess	Baseline	0.233	0.069	0.030	0.162	0.266	0.152
LLaVA 1.6 7B	Zero	0.484	0.135	0.000	0.158	0.096	0.174
	CoT	0.534	0.198	0.068	0.043	0.004	0.169 ↓2.9%
LLaVA 1.6 13B	Zero	0.541	0.027	0.039	0.158	0.000	0.153
	CoT	0.529	0.043	0.054	0.034	0.016	0.135 ↓11.8%
LLaVA 1.6 34B	Zero	0.545	0.000	0.000	0.012	0.004	0.112
	Few	0.457	0.097	0.058	0.318	0.086	0.203 ↑81.3%
	CoT	0.537	0.143	0.062	0.210	0.004	0.191 ↑70.5%
LLaVA OneVision 7B	Zero	0.541	0.000	0.087	0.043	0.000	0.134
	Few	0.146	0.000	0.025	0.172	0.243	0.117 ↓12.7%
	CoT	0.535	0.000	0.048	0.092	0.000	0.135 ↑0.7%
Qwen2-VL 7B	Zero	0.533	0.068	0.000	0.034	0.000	0.127
	Few	0.539	0.000	0.000	0.000	0.004	0.109 ↓14.2%
	CoT	0.530	0.156	0.057	0.080	0.004	0.166 ↑30.7%
Llama 3.2 Vision 11B	Zero	0.537	0.136	0.098	0.023	0.000	0.159
	Few	0.542	0.000	0.026	0.000	0.000	0.114 ↓28.3%
	CoT	0.533	0.189	0.026	0.083	0.020	0.170 ↑6.9%
Phi3.5 Vision 4.2B	Zero	0.542	0.038	0.053	0.104	0.000	0.147
	Few	0.485	0.256	0.021	0.255	0.162	0.236 ↑60.5%
	CoT	0.534	0.000	0.087	0.083	0.000	0.141 ↓4.1%
InternVL 2.5 26B	Zero	0.558	0.273	0.071	0.312	0.027	0.248
	Few	0.498	0.211	0.048	0.253	0.127	0.228 ↓8.1%
	CoT	0.537	0.333	0.052	0.254	0.087	0.252 ↑1.6%
GPT-4o	Zero	0.544	0.345	0.064	0.178	0.065	0.239
	Few	0.549	0.352	0.023	0.390	0.134	0.289 ↑20.9%
	CoT	0.558	0.321	0.054	0.324	0.024	0.256 ↑7.1%
Gemini 1.5 Flash	Zero	0.543	0.215	0.091	0.168	0.020	0.207
	Few	0.543	0.380	0.054	0.402	0.071	0.290 ↑40.1%
	CoT	0.557	0.300	0.000	0.329	0.072	0.252 ↑21.7%
Gemini 1.5 Pro	Zero	**0.559**	0.329	0.039	0.440	0.112	0.296
	Few	0.531	0.391	0.070	0.451	0.253	0.339 ↑14.5%
	CoT	0.558	0.330	0.000	0.350	0.057	0.259 ↓12.5%
Claude 3.5 Sonnet v2	Zero	0.516	0.408	0.070	0.439	0.113	0.309
	Few	0.467	**0.430**	0.077	0.434	0.338	0.349 ↑12.9%
	CoT	0.537	0.378	0.058	0.382	0.119	0.295 ↓4.5%
CLIP Classifier	Baseline	0.548	0.270	**0.150**	**0.479**	**0.687**	**0.427**

Table 5: Results for Coherence Relation Prediction on CLUE Single-Label. The Coherence Relations predicted are Visible, Subjective (Subj), Action, Story and Meta

4.3 Prompting Strategies and Fine-tuning

In addition to zero-shot evaluation, we also investigate the contribution of few-shot and Chain-of-Thought (CoT) prompting strategies in enabling MLLMs to learn coherence relations better. For few-shot, we include one example per coherence relation in each prompt as examples in the 3 single-label classification settings. For multi-label classification on CLUE ML, we include 6 different examples covering different combinations of relations in our prompt. To perform CoT, we include a reasoning step in our prompt that asks the model to generate a rationale before predicting the coherence relation. More details about the prompt templates used for each of the tasks are present in Sections C.1 and D.1 of our appendix. We fine-tune the Llama 3.2 11B Instruct model on our benchmark to measure the impact of task-specifc fine-tuning in open-source MLLMs with hyperparameter selection described in Appendix Section E.

4.4 Main Results

MLLMs Struggle with Coherence Relations From our results in Tables 2, 4, 5, 6 we observe that no MLLM shows improvements over our baseline classifier on Macro F1 scores across all settings. When strictly looking at zero-shot prompts, Claude 3.5 Sonnet v2 performs the best on Tweet Subtitles, CLUE ML, and CLUE SL while Gemini 1.5 Flash performs the best on DisRel. However, the CLIP Classifier can outperform these MLLMs by 2.4% on DisRel, 64.1% on Tweet Subtitles, 38.6% on CLUE SL, and 5.6% on CLUE ML in terms of Macro F1 score. This shows that although these datasets have clearly discernible visual and text features that help in predicting coherence relations, MLLMs aren't able to comprehend them effectively. The trend extends to both proprietary and open-source MLLMs regardless of their size. Our results reiterate the need for benchmarks such as CORDIAL to evaluate the intermodal reasoning capabilities of MLLMs.

Pragmatic Relations are Challenging In single-label prediction settings, we observe that MLLMs come close to the baseline classifier's scores on DisRel, containing the image-text relations that are more literal (Similar, Complementary). On the other hand, there exists a significant gap in performance in other single-label datasets. Looking into per-relation F1 scores, pragmatic relation cate-

gories such as Insertion, Projection, and Extension are particularly challenging for MLLMs. A similar trend is observed in CLUE SL and CLUE ML where MLLMs struggle with relation categories such as Story and Meta.

Verification Accuracy Depends on Settings Analyzing the verification performance of MLLMs in Table 3, we observe that the performance of MLLMs on the verification task is highly dependent on the setting. Across all settings, Claude 3.5 Sonnet v2 performs the best, with an accuracy of 80.8% on DisRel, 70.4% on Tweet Subtitles, 42.8% on CLUE SL and 48.5% on CLUE ML. This shows that MLLMs are able to verify coherence relations better in settings where the relations are more literal and easier to understand. However, the performance of MLLMs on the verification task is significantly lower in settings where the relations are more non-literal and pragmatic.

Inconsistency of Prompting Strategies In our experiments with few-shot and CoT prompting strategies, we observe that the performance of MLLMs is inconsistent across different settings and model families. Across DisRel, Tweet Subtitles, CLUE SL and CLUE ML, a total of 7, 8, 10 and 10 MLLMs respectively show improvements in performance with either few-shot or CoT prompting strategies. However, only 2 MLLMs: LLaVA OneVision 7B and GPT-4o show improvements across all settings. Overall, we observe that in the more difficult settings (CLUE SL and CLUE ML), more number of models are able to leverage one of these alternate prompting strategies to improve their performance. But, even with additional examples or reasoning steps, MLLMs are not able to outperform the baseline classifier. This shows that Coherence Relation Prediction is a fundamentally difficult task that cannot be taught to MLLMs only through prompting strategies.

Fine-tuning Improves MLLM Reasoning Looking at Figure 3, we observe that fine-tuning the Llama 3.2 Vision model on our benchmark proves beneficial for coherence relation prediction. In both DisRel and Tweet Subtitles, we see gains in both zero-shot and few-shot prompt scores with Llama 3.2 Vision up to 18.42% compared to its original performance. On both CLUE ML and SL, we see improvements in either zero-shot or few-shot performance with minimal performance loss on the other. This shows that MLLMs are able to learn to recognize coherence relations better when fine-tuned on a task-specific dataset. Coherence-aware fine-tuning can be a promising direction for improving their reasoning and cognition abilities.

Model Biases Inhibit Prediction Performance Looking at the per-class F1 scores across MLLMs, we observe they are biased towards certain relation categories. This includes the prediction of only a small subset of relations across all samples in an evaluation setting. From Figure 2, we acknowledge that the distribution of relation categories in our benchmark is imbalanced. However, this response imbalance of MLLMs is observed even on majority classes such as Concretization in Tweet Subtitles and Meta relations in CLUE SL and ML. This shows that despite providing few-shot examples and prompt optimization strategies, MLLMs display biases towards certain relation categories. When we look at the results of our fine-tuned model, we can see that prediction results on relations ignored by the base model are improved. This shows that fine-tuning can help mitigate these reasoning biases in MLLMs.

5 Conclusions

We propose CORDIAL, a novel benchmark to evaluate how MLLMs perform MDA using Coherence Relations. Our experiments show existing state-of-the-art MLLMs struggle to match simple baseline classifiers in predicting Coherence Relations across different discourse domains. We also show the impact of evaluating different prompt strategies and the importance of using diverse datasets to probe intermodal reasoning capabilities of MLLMs. Finally, we show that fine-tuning MLLMs on coherence relations can help alleviate model biases and improve their performance on these tasks. This work highlights the need for MLLM benchmarks to evolve beyond factual & perceptual assessment tasks and focus on understanding both literal and pragmatic relationships between multimodal components of real-world discourses. We hope that CORDIAL will serve as a stepping stone for future research in MDA and encourage the community to explore new methods to improve MLLMs on these tasks.

Limitations

While our proposed benchmark provides a comprehensive assessment of intermodal reasoning in current MLLMs, several limitations must be acknowledged. Firstly, the benchmark is currently limited to analyzing coherence relations in single-turn discourses. This is due to a lack of publically available datasets that provide multi-turn image-text pairs with annotated coherence relations. We plan to extend our benchmark to include multi-turn discourse relations as future work. Secondly, although we analyze different discourse domains in our benchmark, we lack a unified set of coherence relations that can be applied across all domains. The difficulty in defining a universal set of coherence relations is due to the varying nature of discourse in different domains. This limits our ability to analyze the inter-domain performance of MLLMs on the same set of relations. Finally, our benchmark is currently limited to the English language and must be extended to multi-lingual discourses as well.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv [cs.CL]*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, A Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, O Vinyals, Andrew Zisserman, and K Simonyan. 2022. Flamingo: A visual language model for few-shot learning. *Neural Inf Process Syst*, abs/2204.14198.

Malihe Alikhani, Baber Khalid, and Matthew Stone. 2023. Image-text coherence and its implications for multimodal AI. *Front. Artif. Intell.*, 6:1048874.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North*, pages 570–575, Stroudsburg, PA, USA. Association for Computational Linguistics.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Stroudsburg, PA, USA. Association for Computational Linguistics.

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aashish Anantha Ramakrishnan, Sharon X Huang, and Dongwon Lee. 2024a. ANCHOR: LLM-driven news subject conditioning for text-to-image synthesis. *arXiv [cs.CV]*.

Aashish Anantha Ramakrishnan, Sharon X Huang, and Dongwon Lee. 2024b. ANNA: Abstractive text-to-image synthesis with filtered news captions. In *Proceedings of the Third Workshop on Advances in Language and Vision Research*. Association for Computational Linguistics.

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/claude/sonnet. Accessed: 2025-2-14.

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 [cs.CV]*.

John A Bateman. 2014. *Text and Image: A critical introduction to the visual/verbal divide*, 1st edition edition. Routledge.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv:2005.14165 [cs]*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660. IEEE.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv [cs.CV]*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024c. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.*, 67(12).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jerry R Hobbs. 1978. *Why is discourse coherent?*, volume 176. SRI International Menlo Park, CA.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv [cs.CL]*.

Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, and Malihe Alikhani. 2021. COSMic: A coherence-aware generation metric for image descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3419–3430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 139:4904–4916.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2016. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv [cs.CV]*, pages 1988–1997.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. 2024. VisOnlyQA: Large vision language models still struggle with visual perception of geometric information. *arXiv [cs.CL]*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gunther Kress. 2009. *Multimodality: A social semiotic approach to contemporary communication*. Routledge, London, England.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024. A survey on benchmarks of multimodal large language models. *arXiv [cs.CL]*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2023. KOSMOS-2.5: A multimodal literate model. *arXiv [cs.CL]*.

Emily E Marsh and M White. 2003. A taxonomy of relationships between images and text. *J. Documentation*, 59:647–672.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2025-2-2.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. 2024. MMRel: A relation understanding benchmark in the MLLM era. *arXiv [cs.CV]*.

OpenAI, Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A J Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz,

Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. GPT-4o system card. *arXiv [cs.CL]*.

Sundar Pichai. 2024. Our next-generation model: Gemini 1.5. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/. Accessed: 2025-2-2.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv:2103.00020 [cs]*.

Navid Rajabi and Jana Kosecka. 2024. GSR-bench: A benchmark for grounded spatial reasoning evaluation via multimodal LLMs. In *NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward*.

Candace Ross, Melissa Hall, Adriana Romero-Soriano, and Adina Williams. 2024. What makes a good metric? evaluating automatic metrics for text-to-image consistency. In *First Conference on Language Modeling*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Chak Lam Shek, Xiyang Wu, Wesley A Suttle, Carl Busart, Erin Zaroukian, Dinesh Manocha, Pratap Tokekar, and Amrit Singh Bedi. 2024. LANCAR: Leveraging language for context-aware robot locomotion in unstructured environments. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9612–9619. IEEE.

Tiberiu Sosea, Iustin Sirbu, Cornelia Caragea, Doina Caragea, and Traian Rebedea. 2021. Using the image-text relationship to improve multimodal disaster tweet classification. *Int Conf Inf Syst Crisis Response Manag*, pages 691–704.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Stroudsburg, PA, USA. Association for Computational Linguistics.

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. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv [cs.CV]*.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score: Better aligning text-to-image models with human preference. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2096–2105. IEEE.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv [cs.CV]*.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023. A simple baseline for knowledge-based visual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14871–14877, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chunpu Xu, Hanzhuo Tan, Jing Li, and Piji Li. 2022. Understanding social media cross-modality discourse in linguistic space. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2459–2471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and evaluating human preferences for text-to-image generation. *arXiv [cs.CV]*.

Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can MLLMs perform text-to-image in-context learning? In *First Conference on Language Modeling*.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2024. VL-ICL bench: The devil in the details of multimodal in-context learning. In *The Thirteenth International Conference on Learning Representations*.

Appendix

Model	Prompt	Visible	Subj	Action	Story	Meta	Macro F1
LLaVA 1.6 7B	Zero	0.864	0.117	0.113	0.048	0.029	0.234
	CoT	0.848	0.245	0.247	0.058	0.013	0.282 ↑20.5%
LLaVA 1.6 13B	Zero	0.869	0.147	0.389	0.115	0.401	0.384
	CoT	0.849	0.095	0.237	0.090	0.048	0.264 ↓31.2%
LLaVA 1.6 34B	Zero	0.868	0.165	0.470	0.369	0.298	0.434
	Few	0.859	0.000	0.471	0.453	0.166	0.390 ↓10.1%
	CoT	0.858	0.117	0.317	0.175	0.163	0.326 ↓24.9%
LLaVA OneVision 7B	Zero	0.820	0.034	0.380	0.024	0.000	0.252
	Few	0.757	0.109	0.510	0.150	0.000	0.305 ↑21.0%
	CoT	0.856	0.150	0.349	0.213	0.154	0.345 ↑36.9%
Qwen2-VL 7B	Zero	0.864	0.045	0.211	0.086	0.013	0.244
	Few	0.864	0.162	0.461	0.368	0.017	0.374 ↑53.3%
	CoT	0.865	0.082	0.094	0.080	0.021	0.228 ↓6.6%
Llama 3.2 Vision 11B	Zero	0.869	0.157	0.424	0.349	0.284	0.417
	Few	0.828	0.248	0.571	0.443	0.499	0.518 ↑24.2%
	CoT	0.850	0.183	0.391	0.420	0.371	0.443 ↑6.2%
Phi3.5 Vision 4.2B	Zero	0.866	0.000	0.092	0.036	0.013	0.201
	Few	0.527	0.226	0.311	0.490	0.036	0.318 ↑58.2%
	CoT	0.819	0.047	0.475	0.294	0.064	0.340 ↑69.2%
InternVL 2.5 26B	Zero	0.822	0.291	0.448	0.324	0.029	0.383
	Few	0.496	0.266	0.491	0.400	0.128	0.356 ↓7.0%
	CoT	0.757	0.397	0.444	0.331	0.059	0.397 ↑3.7%
GPT-4o	Zero	0.858	0.451	0.453	0.291	0.060	0.423
	Few	0.874	0.495	0.561	0.525	0.123	0.515 ↑21.7%
	CoT	0.865	0.506	0.357	0.354	0.084	0.433 ↑2.4%
Gemini 1.5 Flash	Zero	0.875	0.368	0.554	0.355	0.065	0.443
	Few	0.847	0.420	0.648	0.480	0.163	0.512 ↑15.6%
	CoT	0.871	0.419	0.308	0.358	0.109	0.413 ↓6.8%
Gemini 1.5 Pro	Zero	0.884	0.485	0.544	0.313	0.106	0.467
	Few	0.866	0.532	0.668	0.464	0.206	0.547 ↑17.1%
	CoT	0.880	0.403	0.180	0.278	0.090	0.366 ↓21.6%
Claude 3.5 Sonnet v2	Zero	0.891	**0.535**	**0.681**	0.479	0.220	0.561
	Few	0.829	0.503	0.643	0.553	0.360	0.578 ↑3.0%
	CoT	0.876	0.515	0.596	0.389	0.174	0.510 ↓9.1%
CLIP Classifier	Baseline	**0.905**	0.176	0.627	**0.615**	**0.642**	**0.593**

Table 6: Results for Coherence Relation Prediction on the CLUE Multi-Label dataset. The Coherence Relations predicted are Visible, Subjective (Subj), Action, Story and Meta with multiple relations being applicable to a single image-text pair.

A Data Preparation

This section sheds light on the methods used while preparing all the datasets mentioned in this paper for model evaluation. We verify all three datasets used to construct this benchmark have a permissive license that allows usage for research purposes without restrictions (DisRel - MIT License, Tweet Subtitles - MIT License, CLUE - Sourced from Conceptual Captions and free for research use).

A.1 DisREL

Due to limited number of samples in the **Unrelated** category, these image-text pairs were discarded from our train and test set. All placeholder instances of `<URL>` were removed from the text as a part of our data cleaning.

A.2 Tweet Subtitles

This dataset contains two types of captions for tweets: actual and text generated by an image captioning model. We use only the **actual** caption as part of our evaluation.

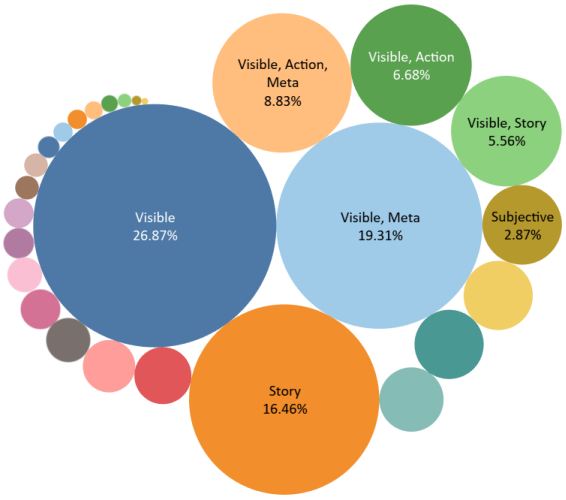


Figure 4: An overview of the Image-Text Label (i.e., Coherence Relations) distribution across CLUE ML

A.3 CLUE

The labels other than the ones mentioned in Section 3.3 were disregarded from our train and test set for both settings, due to the lack of examples. We construct the CLUE Single-Label dataset with the same heuristic used by Alikhani et al. (2020):

Step 1: If the set contains a *Meta* relation, assign it to the image-text pair. Else, proceed to the next step.

Step 2: If the set contains a *Visible* relation and doesn't contain either a *Meta* or *Subjective* relation, assign it to the image-text pair. Else, proceed to the next step.

Step 3: If none of the above rules are met, randomly sample one relation from the 5 available, and assign it to the pair.

B Model Availability

This section focuses on the details of model availability and parameters, that we use in Section 4.1. For all models, we set temperature to 0 or `do_sample=False`, maximum output tokens to 512 and the random seed set to 42, wherever possible to ensure reproducibility. The model responses in this paper were collected between January 12, 2025 and February 12, 2025.

B.1 Proprietary Models

OpenAI GPT: We access the GPT-4o model via the official OpenAI API. We evaluate `gpt-4o-2024-08-06`.

Anthropic Claude: We access Claude 3.5 Sonnet v2 via the Vertex AI API, using Google Cloud. We evaluate `claude-3-5-sonnet-v2@20241022`.

Google Gemini: We access Gemini 1.5 Flash and Gemini 1.5 Pro via the Vertex AI API, using Google Cloud. We evaluate `gemini-1.5-flash-002` and `gemini-1.5-pro-002`.

B.2 Open Source Models

We evaluate models published on Huggingface Hub. LLaVA 1.6 34B and Llama 3.2 11B Vision were evaluated using the LMDeploy [1] framework. We evaluate Qwen2-VL using code released by the authors. All other models, were evaluated using the VLLM [2] framework. Refer to Table 7 for the models we evaluate.

Model	Model ID
InternVL 2.5 26B	`OpenGVLab/InternVL2_5-26B`
Llama 3.2 Vision 11B	`meta-llama/Llama-3.2-11B-Vision-Instruct`
LLaVA 1.6 7B	`llava-hf/llava-v1.6-mistral-7b-hf`
LLaVA 1.6 13B	`llava-hf/llava-v1.6-vicuna-13b-hf`
LLaVA 1.6 34B	`liuhaotian/llava-v1.6-34b`
LLaVA OneVision 7B	`llava-hf/llava-onevision-qwen2-7b-ov-hf`
Phi 3.5 Vision	`microsoft/Phi-3.5-vision-instruct`
Qwen2-VL-7B	`Qwen/Qwen2-VL-7B-Instruct`
Claude 3.5 Sonnet v2	`claude-3-5-sonnet-v2@20241022`
GPT-4o	`gpt-4o-2024-08-06`
Gemini 1.5 Flash	`gemini-1.5-flash-002`
Gemini 1.5 Pro	`gemini-1.5-pro-002`

Table 7: MLLMs we evaluate in this paper. For open-source models, this table shows the model names in Huggingface.

C MLLM Evaluation Details

This section provides details about the *evaluation* task (RQ1) mentioned in Section 4.1.

C.1 Prompt Templates

As mentioned in Section 4.3, we make use of Zero-Shot, Few-Shot and Chain of Thought prompting for evaluation. Every prompting strategy utilizes three different messages:

- **System Message:** We explain the task and the definitions of each Coherence Relation present in the dataset being evaluated.
- **User Message:** This message is used to reiterate the task again, along with the required output format. The image and text that need to be evaluated, is also added here.
- **Assistant Message:** We use this optional message for certain models, to guide its responses towards the intended output format.

The different prompts and system messages used on each data source as mentioned in Section 3.3, is present in the appendix.

C.2 Few Shot Prompting

In this prompting strategy, we utilize user-assistant message pairs that are inserted right after the user message which specifies output format. For the Tweet Subtitles and CLUE Single-Label datasets, we utilize **5-shot examples** to include all possible coherence relations. In the case of CLUE Multi-Label and DisREL, we utilize **6-shot examples** and **2-shot examples** respectively.

We do not evaluate LLaVA 1.6 7B and 13B using this prompting technique, as our prompt (text + multimodal tokens) does not fit into the context length (4096) of these models.

C.3 Chain-of-Thought Prompting

We instruct the model to analyze the image-text pair, before assigning a Coherence Relation in this prompting strategy. We incorporate the instruction "Let's think step by step", to make the model respond with concise sentences that detail its reasoning process.

C.4 Preprocessing Images for Claude

We noticed that some images were above the 5 MB per file size limit imposed by Anthropic for their API. As per their recommendations, we evaluate Claude on images that are resized to 1.3 megapixels, while preserving the aspect ratio.

C.5 Postprocessing MLLM Responses

In the case of single-label datasets, we remove instances of the phrase "Coherence Relation:" along with other punctuation and whitespace. If there exists only one occurrence of a particular coherence relation, we use that as the prediction result for the image-text pair.

While working with CLUE Multi-Label responses, we remove instances of the phrase "Coherence Relations:". All valid JSON in the response is parsed using regular expressions. If the output format is comma-separated values, those responses are parsed appropriately.

[1] https://github.com/InternLM/lmdeploy
[2] https://github.com/vllm-project/vllm

After this, if we cannot find any valid label for an image-text pair from the MLLM's response, we discard the sample from our test set. To ensure test set consistency, we discarded around **200 samples** across all datasets and calculated the final evaluation metrics as mentioned in Section 4.2.

D MLLM Verification Details

This section provides details about the *verification* task (RQ2) mentioned in Section 4.1.

D.1 Prompt Templates

For this task, we utilize a Chain-of-Thought prompting strategy. Each model is given the same system message as before, but along with the image-text pair, we also give the ground truth Coherence Relation. The model is then asked to respond with a True/False answer, along with its rationale for its response.

D.2 Preprocessing Images for Claude

We use the same strategy as mentioned in Section C.4, only for the images that don't come under the file size limit.

D.3 Postprocessing MLLM Responses

We parse boolean values from each MLLM response, and assign **False** to an image-text pair, only if there is any occurrence of the same. For CLUE ML, we provide only overall verification accuracies since it is a multi-label verification problem.

E Fine-tuning Details

We fine-tune LLaMA 3.2 Vision 11B Instruct (`unsloth/Llama-3.2-11B-Vision-Instruct` in Huggingface) using the Unsloth[3] framework. We opted for this framework due to its memory efficiency and rapid fine-tuning. We perform Parameter Efficient Fine-Tuning (PEFT) of all layers (Vision & Language) and modules (Attention & MLP) present. We use the hyperparameters mentioned in Section E.1 on each dataset for fine-tuning. Other parameters have been initialized to their default values.

E.1 Hyperparameters

Common Parameters

- LoRA Parameters: `r=16`
- `num_train_epochs = 3`
- `warmup_steps = 100` since our train sets are relatively small.
- `per_device_train_batch_size = 32`
- `gradient_accumulation_steps = 1`
- `dtype = torch.bfloat16`
- `optim = adamw_torch`
- `weight_decay = 0.01`
- `lr_scheduler_type = cosine`

[3] https://unsloth.ai/blog/vision

Model	Prompt	Sim	Compl	Macro F1
FT-Llama 3.2 Vision 11B	Zero	0.629	0.620	**0.625**
	Few	**0.673**	0.327	0.500 ↓20.0%
Llama 3.2 Vision 11B	Zero	0.388	**0.635**	0.512
	Few	0.509	0.479	0.494 ↓3.5%

Table 8: Per-class Coherence Relation Prediction of Fine-tuned LLama 3.2 Vision 11B (FT-Llama) on the DisRel dataset. The coherence relations predicted are Similar and Complementary.

Model	Prompt	Ins	Concr	Proj	Restmt	Ext	Macro F1
FT-Llama 3.2 Vision 11B	Zero	**0.440**	**0.853**	0.045	0.042	0.148	0.306
	Few	0.231	0.752	**0.213**	**0.100**	**0.254**	**0.310** ↑1.3%
Llama 3.2 Vision 11B	Zero	0.000	0.779	0.000	0.093	0.000	0.175
	Few	0.035	0.388	0.000	0.092	0.113	0.126 ↓28.0%

Table 9: Per-class Coherence Relation Prediction of Fine-tuned LLama 3.2 Vision 11B (FT-Llama) on the Tweet Subtitles dataset. The Coherence Relations predicted are Insertion (Ins), Concretization (Concr), Projection (Proj), Restatement (Restmt) and Extension (Ext).

Model	Prompt	Visible	Subj	Action	Story	Meta	Macro F1
FT-Llama 3.2 Vision 11B	Zero	**0.547**	0.074	0.042	0.045	0.004	0.142
	Few	0.516	**0.230**	0.053	**0.228**	**0.155**	**0.236** ↑66.2%
Llama 3.2 Vision 11B	Zero	0.537	0.136	**0.098**	0.023	0.000	0.159
	Few	0.542	0.000	0.026	0.000	0.000	0.114 ↓28.3%

Table 10: Per-class Coherence Relation Prediction of Fine-tuned LLama 3.2 Vision 11B (FT-Llama) on the CLUE Single-Label dataset. The Coherence Relations predicted are Visible, Subjective (Subj), Action, Story and Meta

Model	Prompt	Visible	Subj	Action	Story	Meta	Macro F1
FT-Llama 3.2 Vision 11B	Zero	0.864	0.228	0.520	0.287	0.431	0.466
	Few	0.864	0.158	**0.586**	0.282	**0.549**	0.488 ↑4.7%
Llama 3.2 Vision 11B	Zero	**0.869**	0.157	0.424	0.349	0.284	0.417
	Few	0.828	**0.248**	0.571	**0.443**	0.499	**0.518** ↑24.2%

Table 11: Per-class Coherence Relation Prediction of Fine-tuned LLama 3.2 Vision 11B (FT-Llama) on the CLUE Multi-Label dataset. The Coherence Relations predicted are Visible, Subjective (Subj), Action, Story and Meta with multiple relations being applicable to a single image-text pair.

DisREL

- LoRA Parameters: `lora_alpha=16`
- Learning Rate = $1e^{-5}$

Tweet Subtitles

- LoRA Parameters: `lora_alpha=16`
- Learning Rate = $1e^{-5}$

CLUE Single-Label

- LoRA Parameters: `lora_alpha=16`
- Learning Rate = $1e^{-5}$

CLUE Multi-Label

- LoRA Parameters: `lora_alpha=8`
- Learning Rate = $1e^{-7}$

E.2 Train Set Preparation for CLUE

During experimentation, we noticed that models fine-tuned on CLUE Single-Label and Multi-Label, tend to skew their responses towards the majority classes (Visible, Story and Meta) in the dataset. In order to curb this behavior, we decided to randomly sample **200 examples** from the CLUE Single-Label train set for these coherence relations alone. The same image-text pairs were used for the multi-label setting as well.

F Baseline Classifier Details

As mentioned in Section 3.4, we employ CLIP Text and Image Encoders (`openai/clip-vit-large-patch14` in Huggingface) in a zero-shot manner to extract multi-modal embeddings. These embeddings are then concatenated together, to form a tensor of size 1536. This multi-modal tensor is then passed through a Multi-Layer Perceptron with two hidden layers of size 512 and 256, along with an output layer equal to the number of Coherence Relations in each dataset. The MLP uses RELU in between each layer for introducing non-linearity, and a Dropout of 0.2 between the first two layers.

A validation split of 10% was created from the train sets. The DisREL, Tweet Subtitles and CLUE Single-Label classifiers were trained using the Cross Entropy Loss, whereas the CLUE Multi-Label classifier used the Binary Cross Entropy Loss along with a Sigmoid Layer. Due to the large class imbalance in CLUE Single-Label, we use a weighted loss function in that classifier alone. Every model was trained with a batch size of 32, using the Adam Optimizer and a learning rate of $1e^{-5}$. Table 12 shows the number of epochs, for which each classifier was trained in every setting.

Dataset	Number of Epochs
DisREL	15
Tweet Subtitles	25
CLUE Single-Label	25
CLUE Multi-Label	50

Table 12: Number of epochs for which each classifier was trained.

G Computational Resources

To evaluate and fine-tune open-source models, we use 2 NVIDIA H100 80GB HBM3 and 2 NVIDIA A100 SXM4 GPUs for around two days worth of computation.

System Message for DisREL

You are an expert linguist and your task is to predict the Coherence Relations of a given image-text pair. A coherence relation captures the structural, logical, and purposeful relationships between an image and its text, capturing the author's intent.

These are the possible coherence relations you can assign to an image-text pair:
- Similar: The image and text provide the same information and share the same focus. There exists significant overlap in information conveyed between modalities.
- Complementary: The image and text do not provide the same information or share the same focus but one modality helps understand the other better.

System Message for Tweet Subtitles

You are an expert linguist and your task is to predict the Coherence Relations of a given image-text pair. A coherence relation captures the structural, logical, and purposeful relationships between an image and its text, capturing the author's intent.

These are the possible coherence relations you can assign to an image-text pair:
- Insertion: The salient object described in the image is not explicitly mentioned in the text.
- Concretization: Both the text and image contain a mention of the main visual entity.
- Projection: The main entity mentioned in the text is implicitly related to the visual objects present in the image.
- Restatement: The text directly describes the image contents.
- Extension: The image expands upon the story or idea in the text, presenting new elements or elaborations, effectively filling in narrative gaps left by the text.

System Message for CLUE Single-Label and Multi-Label

You are an expert linguist and your task is to predict the Coherence Relations of a given image-text pair. A coherence relation captures the structural, logical, and purposeful relationships between an image and its text, capturing the author's intent.

These are the possible coherence relations you can assign to an image-text pair:

- Visible: The text presents information that is intended to recognizably characterize what is depicted in the image.
- Action: The text describes an extended, dynamic process of which the moment captured in the image is a representative snapshot.
- Meta: The text allows the reader to draw inferences not just about the scene depicted in the image but about the production and presentation of the image itself.
- Subjective: The text provides information about the speaker's reaction to, or evaluation of, what is depicted in the image.
- Story: The text provides a free-standing description of the circumstances depicted in the image, analogous to including instructional, explanatory and other background relations.

Zero/Few Shot Prompt for DisREL, Tweet Subtitles and CLUE Single-Label

System
<insert-system-message>

User
Based on provided information, predict the most applicable Coherence Relation for the next image-text pair. Output only one relation (<insert-coherence-relations) and do not include any other information in your response.

Use the format "Coherence Relation: <insert-coherence-relation>" for your response.
(Added to finetuned LLaMA 3.2 Vision's prompt in CLUE Single-Label, to enhance output format adherence.)

<add-few-shot-examples>

<insert-image-text-pair>

Assistant
Coherence Relation:

CoT Prompt for DisREL, Tweet Subtitles and CLUE Single-Label

System
<insert-system-message>

User
Before assigning a coherence relation, let's think step by step and analyze the image-text pair in depth.

<insert-image-text-pair>

Assistant
Analysis: <add-analysis-from-model>

User
Based on provided information, predict the most applicable Coherence Relation for the next image-text pair. Output only one relation (<insert-coherence-relations>) and do not include any other information in your response.

Assistant
Coherence Relation:

Zero/Few Shot Prompt for CLUE Multi-Label

System
<insert-system-message>

User
Based on provided information, predict the correct Coherence Relations for the next image-text pair. Output them as a JSON value to the key labels" and do not include any other information in your response. (Default output format for all models)

Give your predicted labels as comma separated values. Do not include any other information in your response.
(Alternate output format for LLaMA 3.2, Phi 3.5, Qwen2-VL and LLaVA-OneVision)

Use the format "Coherence Relation: <insert-coherence-relation>" for your response.
(Added to LLaVA 1.6 13B prompt to enhance output format adherence.)

<add-few-shot-examples>

<insert-image-text-pair>

Assistant
Coherence Relations:

CoT Prompt for CLUE Multi-Label

System
<insert-system-message>

User
Before assigning a coherence relation, let's think step by step and analyze the image-text pair in depth.

<insert-image-text-pair>

Assistant
Analysis: <add-analysis-from-model>

User
Now, using your analysis, predict the correct Coherence Relations for the image-text pair. Output them as a JSON value to the key labels" and do not include any other information in your response. (Default output format for all models)

Give your predicted labels as comma separated values. Do not include any other information in your response.
(Alternate output format for LLaMA 3.2, Phi 3.5, Qwen2-VL and LLaVA OneVision)

Use the format "Coherence Relation: <insert-coherence-relation>" for your response.
(Added to LLaVA 1.6 13B prompt to enhance output format adherence.)

Assistant
Coherence Relations:

Verification Prompt Template

System
<insert-system-message>

User
Based on provided information, reply True (if appropriate) or False (if not appropriate) for the following image-text pair. Give your rationale behind it.

<insert-image-text-pair>
<insert-coherence-relation>

Sample Assistant Response
<True/False>
Rationale: <model-response>