Logic Unseen: Revealing the Logical Blindspots of Vision-Language Models

Yuchen Zhou[1,2*], Jiayu Tang[1], Shuo Yang[3], Xiaoyan Xiao[1], Yuqin Dai[4], Wenhao Yang[5], Chao Gou[1†], Xiaobo Xia[2†], Tat-Seng Chua[2]

[1]School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China
[2]School of Computing, National University of Singapore, Singapore, Singapore
[3]Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[4]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[5]School of Artificial Intelligence, Nanjing University, Nanjing, China

Abstract

Vision-Language Models (VLMs), exemplified by CLIP, have emerged as foundational for multimodal intelligence. However, their capacity for logical understanding remains significantly underexplored, resulting in critical "logical blindspots" that limit their reliability in practical applications. To systematically diagnose this, we introduce LogicBench, a comprehensive benchmark with over 50,000 vision-language pairs across 9 logical categories and 4 diverse scenarios: images, videos, anomaly detection, and medical diagnostics. Our evaluation reveals that existing VLMs, even the state-of-the-art ones, fall at over 40 accuracy points below human performance, particularly in challenging tasks like Causality and Conditionality, highlighting their reliance on surface semantics over critical logical structures. To bridge this gap, we propose LogicCLIP, a novel training framework designed to boost VLMs' logical sensitivity through advancements in both data generation and optimization objectives. LogicCLIP utilizes logic-aware data generation and a contrastive learning strategy that combines coarse-grained alignment, a fine-grained multiple-choice objective, and a novel logical structure-aware objective. Extensive experiments demonstrate LogicCLIP's substantial improvements in logical comprehension across all LogicBench domains, significantly outperforming baselines. Moreover, LogicCLIP retains, and often surpasses, competitive performance on general vision-language benchmarks, demonstrating that the enhanced logical understanding does not come at the expense of general alignment. We believe LogicBench and LogicCLIP will be important resources for advancing VLM logical capabilities.

Github — https://github.com/yuchen2199/Logic-Unseen

Introduction

"Logic is the anatomy of thought." *— John Locke*

Logic, as the intrinsic structure of human reasoning, permeates every facet of how we perceive the world, communicate effectively, and solve complex problems (Kowalski 2011). In natural language, logical constructs such as conjunction, disjunction, negation, and causality play a crucial role in shaping meaning beyond superficial semantics. For example, consider the medical instructions *"Do not mix drug A **and** drug B"* versus *"Do not mix drug A **or** drug B"*. A subtle change in the logical connective system radically alters the meaning, potentially affecting realistic vital decisions. As multimodal AI systems are increasingly deployed in critical applications (Liu et al. 2025f; Lu et al. 2024; Zhou et al. 2024; Luo et al. 2025b; Liu et al. 2025d; Zhou et al. 2025b), such logical understanding becomes essential. Vision-Language Models (VLMs) must not only perceive images and interpret texts but also understand the complex logical structures that connect them, thereby preventing potentially catastrophic real-world errors.

Although VLMs, exemplified by CLIP (Radford et al. 2021; Cherti et al. 2023), excel in image-text alignment and have served as foundational encoders for various applications, such as open-world perception (Wang, Chan, and Loy 2023; Sun et al. 2024; Zhou, Tan, and Gou 2024; Yan et al. 2025) and Multimodal Large Language Models (MLLMs) (Li et al. 2023; Zhang et al. 2025b; Luo et al. 2025c), their ability to comprehend and reason over logical structures remains significantly underexplored. Figure 1 presents a preliminary analysis revealing these concerning **logical blindspots** in CLIP. As demonstrated, we perturb correctly matched image-text pairs by modifying object categories, attributes, and logical structures within the captions to construct erroneous descriptions. Our findings indicate that CLIP is highly sensitive to changes in objects or attributes, resulting in a significant drop in CLIP scores. However, it struggles to distinguish subtle logical shifts that fundamentally alter the entire sentence's meaning, sometimes even assigning higher matching scores to these logically incorrect descriptions. From left to right in Figure 1, CLIP fails to capture subtle variations in logical structures such as causality, inclusion, comparison, and conjunction. These deficiencies suggest that CLIP relies heavily on superficial semantic alignment, often missing deeper logical structures within natural language, which leads to logical blindspots.

To bridge this gap, in this paper, we first introduce **Log-**

*This work was conducted while Yuchen Zhou was at the NExT++ Research Center, mentored by Prof. Chua and Dr. Xia.
†Corresponding authors.

Figure 1: Preliminary analysis of OpenAI CLIP-L/14 reveals logical blindspots. While CLIP is highly sensitive to changes in objects or attributes, resulting in a significant drop in CLIP scores, it demonstrably struggles to distinguish subtle logical shifts that fundamentally alter the sentence's meaning, sometimes even assigning higher matching scores to descriptions containing logical errors. These examples include perturbations in causal, inclusion, comparison, and conjunctive logical relations.

icBench, a novel and comprehensive benchmark specifically designed to systematically diagnose the logical understanding capabilities of VLMs. Specifically, LogicBench spans a wide range of domains, including everyday images, web videos, anomaly detection, and medical scenarios. It covers nine common logical categories, namely conjunction, disjunction, negation, contrast, comparison, condition, causality, temporality, and inclusion. This broad coverage permits an in-depth evaluation of VLMs' ability to comprehend logic across various contexts. Moreover, we propose two diagnostic tasks to probe these capabilities: (1) Logic-Aware Retrieval, which requires models to understand the logical structure of sentences and retrieve images from diverse data that match specific logical relations, simulating real-world tasks such as search engines, content moderation, and recommendation systems; and (2) Logic Multiple-Choice Questions (MCQ), which challenges models to perform fine-grained reasoning on subtitles with adversarial logical perturbations, selecting the most accurate description of an image from closely related options.

Furthermore, we propose **LogicCLIP**, a novel training framework that elevates the logical sensitivity of VLMs through logic-aware contrastive learning. Technically, we introduce a large-scale hard negative sample generation pipeline, where Large Language Models (LLMs) are employed to craft logically perturbed negative captions from original human-written descriptions meticulously. These captions serve as crucial supervisory signals during contrastive fine-tuning, enabling the model to better align visual content with logically coherent language. Moreover, our proposed logic-aware learning approach encourages the model not only to focus on surface-level semantics but also to explicitly attend to critical logical structures. Extensive experimental results demonstrate that while existing VLMs exhibit significant limitations in logical reasoning, our LogicCLIP not only achieves state-of-the-art performance on LogicBench but also maintains, and in some cases surpasses, the performance of native CLIP on general vision-language benchmarks. Before delving into details, we summarize our contributions as follows:

- **(Benchmark)** We introduce LogicBench, a comprehensive benchmark specifically designed to systematically evaluate the logical understanding capabilities of current VLMs. LogicBench features over 50,000 logical vision-language pairs across 9 categories of logical relationships, encompassing 4 diverse application scenarios and 2 distinct evaluation tasks.
- **(Diagnosis)** We conduct the first systematic evaluation of VLMs' logical reasoning abilities, revealing significant "logical blindspots" and inherent limitations in comprehending complex logical structures.
- **(Solution)** We present LogicCLIP, a novel and effective training framework engineered to boost VLMs' logical sensitivity. Extensive experiments validate that LogicCLIP achieves substantial improvements in logical comprehension across various domains while simultaneously preserving, and often surpassing, competitive performance on standard vision-language benchmarks.

Related Work

Contrastive Language-Image Learning. CLIP (Radford et al. 2021; Liu et al. 2025b,c) has demonstrated impressive zero-shot generalization capabilities by learning cross-modal alignment from massive web data, which has led to its widespread use in various downstream tasks such as recognition (Yu et al. 2025; Zhou, Liu, and Gou 2024), detection (Esmaeilpour et al. 2022), segmentation (Liu et al. 2025e), and diagnosis (Lu et al. 2024). CLIP has also become the most commonly used vision encoder for recent MLLMs (Li et al. 2023; Chen et al. 2024; Liu et al. 2024b;

Figure 2: The overview of our LogicBench. A comprehensive benchmark for evaluating logical reasoning across various domains, including 9 common logical categories, 4 key application scenarios, and 2 challenging tasks.

Tang et al. 2025b,a; Zeng et al. 2025b), enabling LLMs to understand images. However, significant challenges persist. Recent works have highlighted CLIP's limitations in specific areas such as negation (Ko and Park 2025; Park et al. 2025), compositionality (Hsieh et al. 2023), and spatial reasoning (Wang et al. 2025), which hinder the reliability and performance of multimodal AI systems (Tong et al. 2024; Zeng et al. 2025a; Luo et al. 2025a; Liu et al. 2025a; Zhou et al. 2023; Liu, Zhou, and Gou 2023; Huang et al. 2024; Guo, Zhou, and Gou 2024; Li et al. 2025b,a; Zhang et al. 2025a). Benchmarks such as CREPE (Ma et al. 2023) and CC-Neg (Singh et al. 2025) have focused on the compositional understanding, relying on linguistic templates, but they lack diversity in real-world query phrasing. More recently, NegBench (Alhamoud et al. 2025) utilized LLaMA 3.1 (Grattafiori et al. 2024) to generate fluent negations, but the reliance on a single LLM introduces potential bias and hallucinations.

Synthetic Data for Model Training. With the growing prominence of LLMs, the use of synthetic data for training models has become a key area of exploration (Zhou et al. 2025a; Li and Li 2025). The generation of large-scale high-quality synthetic data has shown promise in enhancing model performance, especially when used to augment real data. Several works (Tian et al. 2024; Liu et al. 2024a) have shown that models trained entirely on synthetic images and descriptions can achieve performance comparable to those trained on real-world data.

LogicBench

Definition & Statistics

LogicBench is structured around three key dimensions: a defined set of logical categories, diverse application scenarios, and specifically designed logical-aware tasks for evaluation.

Logical Categories. LogicBench includes 9 fundamental logical categories, collectively forming 51,360 logical image-text pairs. Among these, the most frequent categories are conjunction (20.15%), negation (15.63%), and disjunction (13.36%), while the less common categories include comparison (6.29%) and condition (6.01%), although they still provide a sufficient sample size. These categories enable a detailed analysis of VLMs in their ability to interpret complex logical relations beyond superficial matching. The specific categories and examples are illustrated in Figure 2(a). *See Supp. Mat. A for more details and statistics.*

Application Scenarios. To comprehensively evaluate VLMs' logical understanding in real-world contexts, LogicBench incorporates data from four distinct scenarios: natural scene images, videos, anomaly detection, and medical diagnostics, as shown in Figure 2(b). Specifically, ***(1) Image***: Sourced from the CC12M (Changpinyo et al. 2021) and MSCOCO (Lin et al. 2014) validation sets, these datasets feature highly diverse images from everyday environments. ***(2) Video***: From the MSRVTT dataset (Xu et al. 2016), containing popular video clips from a commercial video search engine, it is used to test the VLMs' ability to handle dynamic logical understanding. ***(3) Anomaly***: Using the DADA-2000 dataset (Fang et al. 2019), this scenario involves identifying traffic anomalies and accidents, requiring precise logical reasoning to identify unusual or hazardous events. ***(4) Medicine***: Leveraging the Open-i dataset (Demner-Fushman et al. 2015), which includes radiology reports alongside medical images, this scenario demands nuanced logical interpretation for clinical relevance, such as understanding negation and causality in medical findings.

Logical-aware Tasks. We design two diagnostic tasks with varying levels of granularity: Logic-Aware Retrieval and Logical Multiple Choice Questions (MCQ), as shown in Figure 2(c). Specifically, ***(1) Logic-Aware Retrieval*** requires the models to understand the logical structure of sentences and retrieve images from a diverse database that match specific logical relations. This simulates real-world applications such as search engines, content moderation, and recommendation systems. ***(2) Logical MCQ*** challenges models to perform discriminative reasoning over subtitles containing adversarial logical perturbations. Models must select the most accurate description of a given image from a set of closely

related options, requiring a precise understanding of subtle logical distinctions. *See Supp. Mat. A for visualizations.*

Data Construction Pipeline

Creating a large-scale, high-quality benchmark to diagnose VLMs' capabilities to understand logical structures presents two inherent challenges. *The first challenge is how to acquire a large-scale and logic-aware positive sample set.* Directly using MLLMs to generate logical captions may introduce hallucinations and bias. These generated sentences are often crafted to fulfill the structural needs of logic, which can result in semantic misalignments with the corresponding visual content. While using human experts to annotate from scratch may serve as an alternative, the high cost associated with this approach severely limits the benchmark's scale. For example, the successful SpatialBench (Wang et al. 2025), despite considerable efforts, includes fewer than 300 MCQs. To address this, we leverage existing human-annotated vision-language datasets and filter out captions containing logical structures to serve as our positive samples by utilizing spaCy and regular expression scripts for syntactic analysis. This approach enables us to process large datasets automatically, identifying samples with embedded logical relationships. Moreover, the selected samples are naturally annotated without the introduction of artificial complex logical structures, ensuring both the authenticity and practical relevance of the data.

The second challenge lies in generating a large-scale set of logic-aware negative samples based on the positive samples. Previous works (Ma et al. 2023; Singh et al. 2025) often rely on template-based perturbations of captions, which could result in grammatically awkward sentences with low diversity and poor-quality negative samples. Other methods use a single LLM to generate negative samples (Alhamoud et al. 2025), which addresses some of the issues but still introduces significant bias due to the lack of diversity. Our solution involves using multiple widely validated LLMs (Qwen 2.5-max, DeepSeek-V3, Gemini-2.5-pro, GPT-4.1, and LLaMA 3.3 70B) to generate diverse and high-quality negative samples. This approach mitigates the bias introduced by relying on a single LLM. After generating both positive and negative samples, we engage human experts to review and select the most accurate and reliable samples, ensuring logical consistency and high quality. The dataset is then finalized, leading to LogicBench, which is a large-scale, high-quality benchmark that not only meets the scale requirements but also incorporates sufficient logical complexity, naturalness, and diversity to effectively diagnose current VLMs. Figure 3 illustrates the LogicBench construction pipeline, *which is further detailed in Supp.Mat.A.*

LogicCLIP

As noted, CLIP relies on superficial semantic alignment, matching image features with text based on simple co-occurrence patterns. However, it struggles to capture deeper logical structures, missing subtle semantic shifts in complex logical relationships. This leads to the phenomenon of "logical blindspots". The root cause of these blindspots lies in both the *training data* and *optimization objectives*. First, the training data lacks sufficient explicit examples of complex logical relationships. Second, the optimization objective does not explicitly account for logical structures. The combination of factors prevents CLIP from learning these intricate logical relationships. To address this, we propose LogicCLIP, which tackles "logical blindspots" challenges through both data augmentation and optimization.

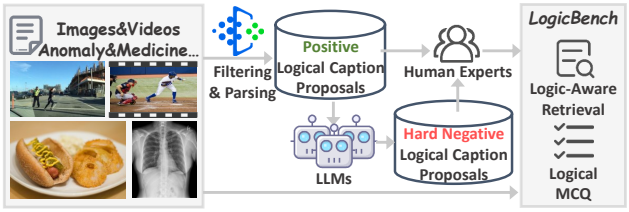

Figure 3: LogicBench Construction Pipeline. We select vision-language pairs from human-annotated datasets, then parse and filter them for logical relations to form positive caption proposals. Multiple LLMs generate hard negative captions by perturbing the positive ones. Human experts review all proposals to create LogicBench.

Logic-Aware Training Data Generation

We expand the MSCOCO training set (Lin et al. 2014), which includes high-quality human-annotated captions from daily image scenes, to generate our logic-aware training data. We aim to enhance the model's ability to adapt to diverse logical structures and comprehend logical relations in various scenarios. This is accomplished by utilizing naturally occurring logical sentence patterns from general daily contexts. The process involves two key steps:

Positive Logical Sample Parsing. We begin by leveraging human-annotated captions from MSCOCO and applying spaCy and regular expression scripts for syntactic analysis. Let $\mathcal{I}$ be the set of images, $\mathcal{C}$ the set of captions, and $\mathcal{L} = \{L_1, \ldots, L_k\}$ the set of predefined logical categories. For each image $I \in \mathcal{I}$ with associated captions $C_I \subset \mathcal{C}$, we filter and identify captions $c_{\text{pos}} \in C_I$ that contain specific logical structures. Each such caption c_{pos} is then associated with a non-empty subset of logical categories $L \subseteq \mathcal{L}$. These (I, c_{pos}, L) pairs form the positive samples, which the model is trained to recognize. *See Supp. Mat. B for more details.*

Negative Logical Sample Generation. To create challenging negative samples, we generate a diverse set using multiple well-validated LLMs (Qwen 2.5-max, DeepSeek-V3, Gemini-2.5-pro, GPT-4.1, and LLaMA 3.3 70B). For each positive sample (I, c_{pos}, L), LLMs perturb c to generate a set of three negative captions, denoted as $C_{\text{neg}} = \{c_{\text{neg},1}, c_{\text{neg},2}, c_{\text{neg},3}\}$. Each caption $c_{\text{neg},m} \in C_{\text{neg}}$ is semantically plausible but logically incorrect or misleading for image I given the logical categories in L. This strategy ensures significant diversity in the negative sample set and pushes the model's capacity to differentiate subtle logical nuances.

The complete Logic-aware Training Data $\mathcal{D}$ consists of samples structured as $D = (I, L, c_{\text{pos}}, C_{\text{neg}})$, where each sample includes an image, its logical categories, a positive

caption, and a set of three negative captions. Notably, $\mathcal{D}$ is derived from general daily images and is not specifically curated for the four LogicBench scenarios. Nevertheless, as shown in Sec. V, LogicCLIP demonstrates remarkable generalization, performing robustly across logically critical domains such as video, anomaly detection, and medical diagnostics, even when fine-tuned solely on these daily images.

Logic-Aware Contrastive Learning

To equip VLMs with robust logical understanding capabilities, our optimization strategy integrates three distinct and complementary objectives: a coarse-grained standard CLIP contrastive objective, a fine-grained hard multiple-choice objective, and a novel logical structure-aware objective, which collectively enhance VLMs' understanding of logical relations within both visual and linguistic modalities.

Standard CLIP Contrastive Objective. The primary objective of this component is to improve the model's overall logic-aware visual-language alignment at a coarse-grained level. By ensuring that the model learns to align image content with logical text descriptions, this foundational objective sets the stage for more fine-grained logical understanding. For a given batch $\mathcal{B} = \{(I_i, C_i)\}_{i=1}^{N}$, where N is the number of image-caption pairs, we first extract their respective representations. An image encoder E_{I} produces image features $\mathbf{v}_i = E_{\text{I}}(I_i)$, and a text encoder E_{T} generates text features $\mathbf{t}_i = E_{\text{T}}(C_i)$. The cosine similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ is computed, where $\mathbf{S}_{ij} = \cos(\mathbf{v}_i, \mathbf{t}_j)$. We apply a symmetric cross-entropy loss to this matrix, encouraging high similarity for correct image-caption pairs $(\mathbf{v}_i, \mathbf{t}_i)$ and low similarity for incorrect pairs:

$$L_{\text{CLIP}} = \mathcal{L}_{\text{CE}}(\mathbf{S}, \text{labels}),$$

where labels indicate an identity matrix with ones on the diagonal indicating correct matches.

Fine-Grained Multiple-Choice Objective. This objective aims to improve the model's ability to distinguish subtle logical differences, particularly when confronted with challenging negative samples. It directly leverages the image-positive sample-negative sample sets generated during data creation. Given a batch $\mathcal{B}_{\text{MC}} = \{(I_i, c_{\text{pos},i}, C_{\text{neg},i})\}_{i=1}^{N}$, where I_i is an image, $c_{\text{pos},i}$ is its corresponding positive logical caption, and $C_{\text{neg},i} = \{c_{\text{neg},i,1}, c_{\text{neg},i,2}, c_{\text{neg},i,3}\}$ are the three hard negative captions generated by logical perturbations, we construct a set of four description options $C_{\text{options},i} = \{c_{\text{pos},i}\} \cup C_{\text{neg},i}$. The model is tasked with identifying the correct caption $c_{\text{pos},i}$ among these four options for each image I_i. We compute the cosine similarity between the image feature $\mathbf{v}_i = E_{\text{I}}(I_i)$. This leads to a set of logits $\mathbf{l}_i = \{\cos(\mathbf{v}_i, \mathbf{t}_{\text{option},m})\}_{m=1}^{4}$, representing the similarity between the image and each option. The multiple-choice loss L_{MC} is then computed by applying a cross-entropy loss over the logits, with the index of $c_{\text{pos},i}$ as the ground-truth target:

$$L_{\text{MC}} = -\log \frac{\exp(\mathbf{l}_{i,\text{pos}})}{\sum_{m=1}^{4} \exp(\mathbf{l}_{i,m})},$$

where $\mathbf{l}_{i,\text{pos}}$ denotes the logit corresponding to the positive caption $c_{\text{pos},i}$. This encourages the model to assign higher scores to logically correct captions and lower scores to incorrect ones, fostering fine-grained logical discrimination.

Logical Structure-Aware Objective. The aim of this objective is to explicitly guide VLMs to identify specific logical structures within a given text description, further enhancing the model's attention to logical relations rather than merely focusing on surface-level semantic matching. For each caption C_i in a batch, we obtain its text encoder feature $\mathbf{t}_i = E_{\text{T}}(C_i)$. We then feed $\mathbf{t}_i$ into a logical classifier F_{Logic}, which outputs a vector of predicted scores for each logical category in $\mathcal{L}$. The classifier assigns a score for each category, indicating the likelihood that caption C_i belongs to that category. Since a caption C_i can contain multiple logical categories, its ground-truth label is a multi-hot encoded vector $y_i \in \{0, 1\}^k$, where $y_{i,j} = 1$ if C_i contains logical category L_j and 0 otherwise. We train this logical classifier using binary cross-entropy loss L_{Logic}. This loss explicitly encourages the model to accurately predict all logical categories present in the text description, thereby enhancing its direct comprehension of logical structures.

Total Optimization Objective. The total optimization objective is a weighted sum of these three components:

$$L_{\text{Total}} = \alpha L_{\text{CLIP}} + \beta L_{\text{MC}} + \gamma L_{\text{Logic}},$$

where α, β, and γ are hyperparameters that control the contribution of each loss term. This composite objective ensures that LogicCLIP maintains general vision-language alignment, accurately distinguishes subtle logical differences, and explicitly understands various logical relations through direct classification.

Experiments

Experimental Settings

Evaluated Models. We conduct a comprehensive evaluation of state-of-the-art VLMs on LogicBench and general benchmarks, including OpenAI's native CLIP (base-32 and large-14 versions) (Radford et al. 2021), LaCLIP (Fan et al. 2023), LaionCLIP (Schuhmann et al. 2022), DatacompCLIP (Gadre et al. 2023), NegCLIP (Yuksekgonul et al. 2023), TripletCLIP (Patel et al. 2024), MetaCLIP (Xu et al. 2024), ILCLIP (Zheng et al. 2024), ConCLIP (Singh et al. 2025), and NegFull (Alhamoud et al. 2025). Additionally, we include a set of human evaluation experiments by randomly selecting 10% of the samples from LogicBench. Independent human evaluators answered each question and provided assessments, serving as a reference for human performance.

Benchmarks & Metrics. In addition to LogicBench, we also evaluate the performance of VLMs on general benchmarks to comprehensively assess model capabilities. Following (Wang et al. 2025), we use COCO (Lin et al. 2014) and Flickr30k (Plummer et al. 2015) for text-image retrieval. For MCQ tasks, we report accuracy. For retrieval tasks, we report Recall@1 and Recall@5.

Training Settings. To evaluate the efficacy of our LogicCLIP framework, we fine-tune four representative pretrained VLMs: OpenAI CLIP-B, OpenAI CLIP-L, NegCLIP, and NegFull. We then compare the fine-tuned models with their respective base versions on both LogicBench and general

Model	LogicBench								General Benchmark			
	Image			*Video*			*Anomoly*	*Medicine*	*COCO*		*Flickr30K*	
	MCQ	**R@1**	**R@5**	**MCQ**	**R@1**	**R@5**	**MCQ**	**MCQ**	**R@1**	**R@5**	**R@1**	**R@5**
Human	96.32	-	-	93.97	-	-	94.23	87.00	-	-	-	-
LaCLIP	42.28	28.80	52.47	42.55	44.40	69.73	35.14	13.50	31.67	56.12	57.58	82.14
LaionCLIP	50.81	36.41	61.73	46.32	56.40	78.53	30.92	15.40	39.37	65.42	66.76	88.38
DatacompCLIP	36.60	8.94	23.34	35.47	22.53	44.80	32.93	7.40	10.98	27.02	18.10	40.36
TripletCLIP-CC3M	44.37	3.50	10.39	39.81	7.33	20.67	16.62	26.20	3.59	11.29	8.72	22.08
TripletCLIP-CC12M	56.32	10.18	26.15	48.49	21.33	44.93	15.66	17.50	11.25	28.20	25.36	51.76
MetaCLIP	54.05	33.06	58.26	44.34	52.80	76.80	25.85	17.00	36.62	62.48	63.84	86.14
ILCLIP	54.55	9.88	25.21	30.94	14.80	33.33	24.00	17.30	12.35	30.27	20.52	43.36
ConCLIP	38.58	23.00	45.09	42.08	44.53	68.67	30.22	8.30	27.67	52.26	56.74	82.68
NegFull	60.04	27.00	50.98	55.85	45.60	71.60	31.22	17.90	30.60	55.49	57.68	82.54
NegCLIP	55.27	39.38	67.06	53.96	59.07	82.00	32.03	21.30	41.52	68.43	67.44	89.50
OpenAI CLIP-B	38.56	27.88	50.85	35.57	49.60	72.40	28.36	16.50	30.44	55.97	58.78	83.54
OpenAI CLIP-L	36.90	34.24	58.38	35.66	53.73	76.40	24.75	12.80	36.52	61.06	65.00	87.26
LogicCLIP-NegFull (Ours)	79.34	39.85	67.64	70.85	57.60	82.93	57.83	37.60	42.77	69.18	67.20	89.44
LogicCLIP-Neg (Ours)	**85.69**	42.30	69.50	**82.55**	**59.33**	82.67	56.38	**46.60**	44.38	71.28	69.84	90.58
LogicCLIP-B (Ours)	81.91	39.82	67.00	77.92	57.60	**83.33**	45.33	33.80	42.54	69.74	68.78	89.66
LogicCLIP-L (Ours)	83.93	**44.28**	**71.46**	79.53	58.53	83.07	**62.30**	35.90	**45.27**	**71.54**	**71.98**	**91.76**

Table 1: Performance comparison of different models on LogicBench and general benchmark datasets.

benchmarks. All models are fine-tuned for 16 epochs with a 1,000-step linear warmup. We employ the AdamW optimizer with a weight decay of 0.2. Batch sizes are set to 256 for OpenAI CLIP-B and NegCLIP. Due to memory constraints, OpenAI CLIP-L is fine-tuned with a reduced batch size of 64. Experiments are implemented using PyTorch and conducted on a single NVIDIA A100-80G GPU. The hyperparameters α, β, and γ are set to 4, 2, and 1, respectively.

Results & Insights

Logic: A major challenge faced by current VLMs. To begin, we evaluate the performance of various VLMs on LogicBench, as shown in Table 1. These results reveal significant shortcomings of current VLMs in logical understanding. Even for relatively simple image scenes, models like LaionCLIP and MetaCLIP, which perform well on general benchmarks, achieve only 50.81% and 54.05% on the MCQ task, respectively. The strongest baseline, NegFull, scores 60.04% on MCQ but only 27.00% on R@1. Furthermore, models show even weaker performance in more complex scenarios such as video, anomaly detection, and medical diagnostics, with MCQ scores dropping significantly. Other competitive pre-trained models, such as LaCLIP (42.28% on image MCQ), LaionCLIP (50.81% on image MCQ), and the original OpenAI CLIP-B (38.56% on image MCQ), achieve more moderate scores. In contrast, human evaluators perform exceptionally well across all four LogicBench scenarios, with scores surpassing 90% in image, video, and anomaly tasks. These results collectively underscore that the current VLMs are inadequate for handling logical relations.

LogicCLIP achieves remarkable improvements on logical tasks. In contrast, our LogicCLIP consistently demonstrates superior performance across all LogicBench scenarios, significantly outperforming its original versions. LogicCLIP-B and LogicCLIP-L show a significant boost in performance on LogicBench, particularly in MCQ tasks and R@1 across all application scenarios. For example, LogicCLIP-B achieves an MCQ score of 81.91%, far surpassing the OpenAI CLIP-B score of 38.56%. Similarly, LogicCLIP-L elevates Image MCQ from 36.90% to 83.93%, and LogicCLIP-Neg achieves the highest Image MCQ score at 85.69% compared to NegCLIP's 55.27%. These substantial improvements extend across other logical tasks within LogicBench: LogicCLIP-B improves Video MCQ from 35.57% to 77.92%, Anomaly MCQ from 28.36% to 45.33%, and Medicine MCQ from 16.50% to 33.80%.

Logic-aware training also improves performance on general benchmarks. On general benchmarks, our model not only maintains the performance of its base model but often achieves significant improvements. For instance, LogicCLIP-B improves the R@1 accuracy on COCO from 30.44% (OpenAI CLIP-B) to 42.54%, and on Flickr30K, the R@1 accuracy increases from 58.78% to 68.78%. These results indicate that LogicCLIP enhances the model's ability to capture deeper logical relationships, prompting it to build richer, more nuanced image and text representations. As a result, the overall quality of image-text alignment is improved. This enhanced understanding likely contributes to better generalization on general retrieval tasks. Our findings suggest that LogicCLIP has the potential to guide the development of more robust and powerful VLMs.

Broad generalization across unseen logical domains. It is important to emphasize the remarkable generalization capabilities of LogicCLIP. While our training data is derived solely from daily images, the logic-aware abilities acquired by LogicCLIP effectively transfer to unseen, specialized domains within LogicBench. For instance, LogicCLIP-B boosts Video MCQ from 35.57% to an impressive 77.92%. LogicCLIP-L boosts Anomaly MCQ from 24.75% to 62.30%, and LogicCLIP-Neg improves Medicine MCQ from 21.30% to 46.60%. This demonstrates that LogicCLIP learns a fundamental understanding of logical structures that

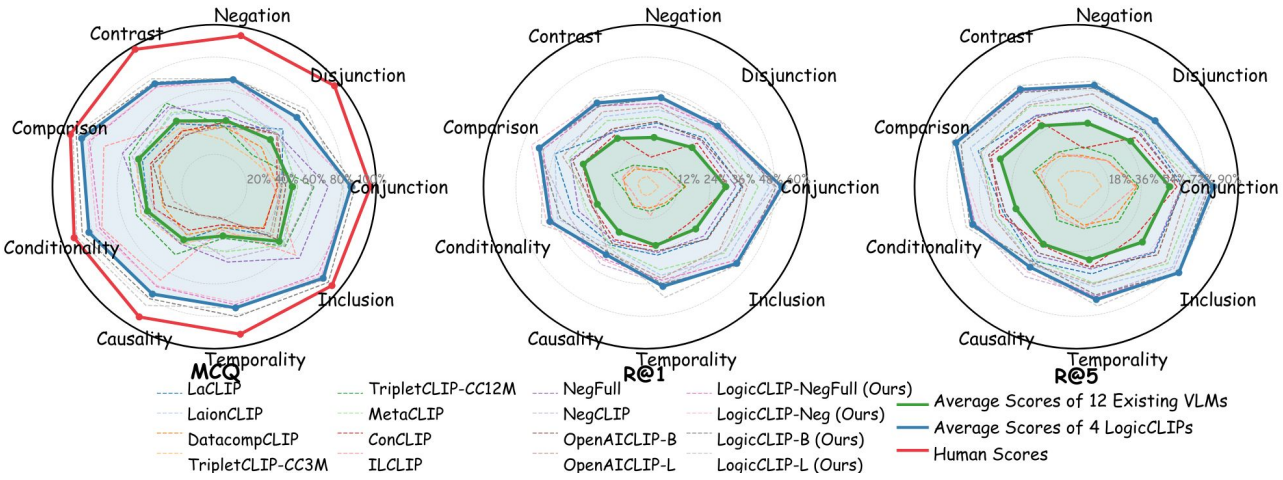

Figure 4: Performance comparison across different logical structures. *See Supp. Mat. C for more detailed results.*

is transferable across diverse visual contexts.

Varying difficulty across logical categories. Beyond overall performance, analyzing results across specific logical structures offers deeper insights into model capabilities and limitations. Figure 4 details the MCQ, R@1, and R@5 scores for 9 distinct logical categories. Our analysis reveals a clear spectrum of difficulty for current baseline VLMs in discerning different logical relationships. ***(1)*** *Temporality, Causality, Conditionality, and Negation prove to be the most challenging*, with average baseline MCQ scores of 31.18%, 35.15%, 41.23%, and 41.47%, respectively. For example, OpenAI CLIP-B/L consistently scores below 20% MCQ on Temporality and Causality, highlighting its weakness in handling nuanced temporal and causal relationships. ***(2)*** *Disjunction, Contrast, and Comparison present moderate challenges*, with average MCQ scores of 46.28%, 46.97%, and 47.71%. Baselines score in the 40-60% range, suggesting partial understanding but substantial room for improvement. ***(3)*** *Conjunction and Inclusion are comparatively less challenging for some strong baselines*, with average MCQ scores of 49.26% and 51.18%, respectively. LogicCLIP-NegFull achieves 70.30% on Conjunction and 68.60% on Inclusion, though standard CLIP variants show only moderate performance in these areas. Our LogicCLIP consistently demonstrates superior performance across all nine logical categories, dramatically overcoming the limitations of baseline models. In challenging categories like Temporality, Causality, and Conditionality, LogicCLIP's MCQ scores frequently leap from baseline 20-40% to over 80-90%. It also excels in other logical structures, including Comparison, Conjunction, and Inclusion, regularly achieving over 90% MCQ accuracy. *More analyses are in Supp. Mat. C.*

Conclusion

In this paper, we tackle the critical challenge of logical understanding in VLMs, a crucial yet underexplored capability for their reliable deployment. We introduce LogicBench, a comprehensive benchmark designed to assess the logical reasoning capabilities of VLMs. Our evaluation reveals significant logical blindspots in current VLMs, particularly in tasks involving Temporality, Causality, Conditionality, and Negation. We further analyze that these blindspots stem from a lack of explicit consideration of logic during both data collection and optimization objectives. To address these limitations, we propose LogicCLIP, a novel training framework that enhances VLMs' logical sensitivity by integrating a large-scale hard negative sample generation pipeline and a logic-aware contrastive learning strategy, both focused on improving logical understanding. Extensive experiments demonstrate that LogicCLIP not only outperforms state-of-the-art models on LogicBench but also retains competitive performance on general vision-language benchmarks. Additionally, LogicCLIP generalizes effectively across diverse domains such as video, anomaly detection, and medical diagnostics. Our work highlights the critical role of explicit logical training in VLMs and sets the stage for building more robust and reliable models for real-world applications.

Acknowledgements

This research/project is supported by National Natural Science Foundation of China under Grant 62373387, Beijing Natural Science Foundation (4252048), Shenzhen Fundamental Research Program (Grant No. JCYJ20240813151301003), and International Program for Candidates, Sun Yat-Sen University. This research/project is also supported in part by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP2024-002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Xiaobo Xia is partially supported by MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China (Grant No. 2421002). Shuo Yang is supported by NSFC Young Scientists Fund (No. 62506096) and Shenzhen Fundamental Research Program (Grant No. JCYJ20250604145514018).

References

Alhamoud, K.; Alshammari, S.; Tian, Y.; Li, G.; Torr, P. H.; Kim, Y.; and Ghassemi, M. 2025. Vision-language models do not understand negation. In *CVPR*, 29612–29622.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 3558–3568.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 24185–24198.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2818–2829.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Esmaeilpour, S.; Liu, B.; Robertson, E.; and Shu, L. 2022. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *AAAI*, volume 36, 6568–6576.

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2023. Improving clip training with language rewrites. In *NeurIPS*, 35544–35575.

Fang, J.; Yan, D.; Qiao, J.; Xue, J.; Wang, H.; and Li, S. 2019. Dada-2000: Can driving accident be predicted by driver attention$f$ analyzed by a benchmark. In *ITSC*, 4303–4309.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. 2023. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, 27092–27112.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, Z.; Zhou, Y.; and Gou, C. 2024. Drivinggen: Efficient safety-critical driving video generation with latent diffusion models. In *ICME*, 1–6. IEEE.

Hsieh, C.-Y.; Zhang, J.; Ma, Z.; Kembhavi, A.; and Krishna, R. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS*, 31096–31116.

Huang, Z.; Zhou, Y.; Zhu, J.; and Gou, C. 2024. Driver Scanpath Prediction Based On Inverse Reinforcement Learning. In *ICASSP*, 8306–8310.

Ko, H.; and Park, C.-M. 2025. Bringing CLIP to the Clinic: Dynamic Soft Labels and Negation-Aware Learning for Medical Analysis. In *CVPR*, 25897–25906.

Kowalski, R. 2011. *Computational logic and human thinking: how to be artificially intelligent*. Cambridge University Press.

Li, H.; and Li, B. 2025. Enhancing Vision-Language Compositional Understanding with Multimodal Synthetic Data. In *CVPR*, 24849–24861.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742.

Li, Y.; Yang, J.; Shen, Z.; Han, L.; Xu, H.; and Tang, R. 2025a. Catp: Contextually adaptive token pruning for efficient and enhanced multimodal in-context learning. *arXiv preprint arXiv:2508.07871*.

Li, Y.; Yun, T.; Yang, J.; Feng, P.; Huang, J.; and Tang, R. 2025b. TACO: Enhancing Multimodal In-context Learning via Task Mapping-Guided Sequence Configuration. *arXiv preprint arXiv:2505.17098*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, C.; Wan, Z.; Wang, H.; Chen, Y.; Qaiser, T.; Jin, C.; Yousefi, F.; Burlutskiy, N.; and Arcucci, R. 2024a. Can Medical Vision-Language Pre-training Succeed with Purely Synthetic Data? *arXiv preprint arXiv:2410.13523*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024b. Improved baselines with visual instruction tuning. In *CVPR*, 26296–26306.

Liu, X.; Xia, X.; Huang, Z.; Ng, S.-K.; and Chua, T.-S. 2025a. Towards modality generalization: A benchmark and prospective analysis. *ACM MM*.

Liu, X.; Xia, X.; Ng, S.-K.; and Chua, T.-S. 2025b. Continual multimodal contrastive learning. In *NeurIPS*.

Liu, X.; Xia, X.; Ng, S.-K.; and Chua, T.-S. 2025c. Principled multimodal representation learning. *arXiv preprint arXiv:2507.17343*.

Liu, X.; Xia, X.; Zhao, W.; Zhang, M.; Yu, X.; Su, X.; and Yang, S. e. 2025d. L-MTP: Leap Multi-Token Prediction Beyond Adjacent Context for Large Language Models. In *NeurIPS*.

Liu, X.; Zhou, Y.; and Gou, C. 2023. Learning from interaction-enhanced scene graph for pedestrian collision risk assessment. *IEEE T-IV*, 8(9): 4237–4248.

Liu, Y.; Wang, G.; Zhang, J.; Liu, Q.; and Huang, D. 2025e. Unveiling the Knowledge of CLIP for Training-Free Open-Vocabulary Semantic Segmentation. In *AAAI*, volume 39, 5649–5657.

Liu, Y.; Zhang, Y.; Cai, J.; Jiang, X.; Hu, Y.; Yao, J.; Wang, Y.; and Xie, W. 2025f. Lamra: Large multimodal model as your advanced retrieval assistant. In *CVPR*, 4015–4025.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3): 863–874.

Luo, R.; Shan, R.; Chen, L.; Liu, Z.; Wang, L.; Yang, M.; and Xia, X. 2025a. VCM: Vision Concept Modeling Based on Implicit Contrastive Learning with Vision-Language Instruction Fine-Tuning. *arXiv preprint arXiv:2504.19627*.

Luo, R.; Wang, L.; He, W.; and Xia, X. 2025b. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*.

Luo, R.; Xia, X.; Wang, L.; Chen, L.; Shan, R.; Luo, J.; Yang, M.; and Chua, T.-S. 2025c. NExT-OMNI: Towards Any-to-Any Omnimodal Foundation Models with Discrete Flow Matching. *arXiv preprint arXiv:2510.13721*.

Ma, Z.; Hong, J.; Gul, M. O.; Gandhi, M.; Gao, I.; and Krishna, R. 2023. Crepe: Can vision-language foundation models reason compositionally? In *CVPR*, 10910–10921.

Park, J.; Lee, J.; Song, J.; Yu, S.; Jung, D.; and Yoon, S. 2025. Know" No"Better: A Data-Driven Approach for Enhancing Negation Awareness in CLIP. *arXiv preprint arXiv:2501.10913*.

Patel, M.; Kusumba, N. S. A.; Cheng, S.; Kim, C.; Gokhale, T.; Baral, C.; et al. 2024. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. In *NeurIPS*, 32731–32760.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2641–2649.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 25278–25294.

Singh, J.; Shrivastava, I.; Vatsa, M.; Singh, R.; and Bharati, A. 2025. Learning the Power of "No": Foundation Models with Negations. In *WACV*, 8002–8012. IEEE.

Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; and Wang, J. 2024. Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. In *CVPR*, 13019–13029.

Tang, F.; Gu, Z.; Lu, Z.; Liu, X.; Shen, S.; Meng, C.; Wang, W.; Zhang, W.; Shen, Y.; Lu, W.; Xiao, J.; and Zhuang, Y. 2025a. GUI-G^2: Gaussian Reward Modeling for GUI Grounding. arXiv:2507.15846.

Tang, F.; Xu, H.; Zhang, H.; Chen, S.; Wu, X.; Shen, Y.; Zhang, W.; Hou, G.; Tan, Z.; Yan, Y.; Song, K.; Shao, J.; Lu, W.; Xiao, J.; and Zhuang, Y. 2025b. A Survey on (M)LLM-Based GUI Agents. arXiv:2504.13865.

Tian, Y.; Fan, L.; Chen, K.; Katabi, D.; Krishnan, D.; and Isola, P. 2024. Learning vision from models rivals learning vision from data. In *CVPR*, 15887–15898.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 9568–9578.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, 2555–2563.

Wang, Z.; Zhou, S.; He, S.; Huang, H.; Yang, L.; Zhang, Z.; Cheng, X.; Ji, S.; Jin, T.; Zhao, H.; et al. 2025. SpatialCLIP: Learning 3D-aware Image Representations from Spatially Discriminative Language. In *CVPR*, 29656–29666.

Xu, H.; Xie, S.; Tan, X.; Huang, P.-Y.; Howes, R.; Sharma, V.; Li, S.-W.; Ghosh, G.; Zettlemoyer, L.; and Feichtenhofer, C. 2024. Demystifying CLIP Data. In *ICLR*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Yan, J.; Ren, R.; Liu, J.; Xu, S.; Wang, L.; Wang, Y.; Wang, Y.; Zhang, L.; Chen, X.; Sun, C.; et al. 2025. TeleEgo: Benchmarking Egocentric AI Assistants in the Wild. *arXiv preprint arXiv:2510.23981*.

Yu, Y.; Cao, C.; Zhang, Y.; Lv, Q.; Min, L.; and Zhang, Y. 2025. Building a multi-modal spatiotemporal expert for zero-shot action recognition with clip. In *AAAI*, 9689–9697.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It? In *ICLR*.

Zeng, P.; Yin, J.; Sun, H.; Dai, Y.; Jiang, M.; Zhang, M.; and Lu, S. 2025a. MRED-14: A Benchmark for Low-Energy Residential Floor Plan Generation with 14 Flexible Inputs. In *ACM MM*, 11298–11307.

Zeng, P.; Yin, J.; Zhang, M.; Dai, Y.; Li, J.; Jin, Z.; and Lu, S. 2025b. Card: Cross-modal agent framework for generative and editable residential design. In *EMNLP*, 9315–9330.

Zhang, D.; Chen, X.; Luo, J.; Jia, M.; Sun, C.; Ren, R.; Liu, J.; Sun, H.; and Li, X. 2025a. Infinite Video Understanding. *arXiv preprint arXiv:2507.09068*.

Zhang, H.; Zhang, W.; Qu, H.; and Liu, J. 2025b. Enhancing human-centered dynamic scene understanding via multiple llms collaborated reasoning. *Visual Intelligence*, 3(1): 3.

Zheng, C.; Zhang, J.; Kembhavi, A.; and Krishna, R. 2024. Iterated Learning Improves Compositionality in Large Vision-Language Models. In *CVPR*, 13785–13795.

Zhou, Y.; Liu, L.; and Gou, C. 2024. Learning from observer gaze: Zero-shot attention prediction oriented by human-object interaction recognition. In *CVPR*, 28390–28400.

Zhou, Y.; Tan, G.; and Gou, C. 2024. Hierarchical home action understanding with implicit and explicit prior knowledge. In *ICASSP*, 4015–4019. IEEE.

Zhou, Y.; Tan, G.; Li, M.; and Gou, C. 2023. Learning from easy to hard pairs: Multi-step reasoning network for human-object interaction detection. In *ACM MM*, 4368–4377.

Zhou, Y.; Tang, J.; Xiao, X.; Lin, Y.; Liu, L.; Guo, Z.; Fei, H.; Xia, X.; and Gou, C. 2025a. Where, What, Why: Towards Explainable Driver Attention Prediction. In *ICCV*.

Zhou, Y.; Xia, X.; Lin, Z.; Han, B.; and Liu, T. 2024. Few-shot adversarial prompt learning on vision-language models. In *NeurIPS*, 3122–3156.

Zhou, Z.; Xia, X.; Ma, F.; Fan, H.; Yang, Y.; and Chua, T.-S. 2025b. DreamDPO: Aligning Text-to-3D Generation with Human Preferences via Direct Preference Optimization. In *ICML*.