Explicit Logic Channel for Validation and Enhancement of MLLMs on Zero-Shot Tasks

Mei Chee Leong, Ying Gu, Hui Li Tan, Liyuan Li, and Nancy F. Chen

Institute for Infocomm Research (I^2R),
Present address: Institute of Advanced Intelligence and Computing (IAIC),
Agency for Science, Technology and Research (A*STAR), Singapore
{leongmc,guy,Li_Liyuan,Nancy_Chen}@a-star.edu.sg,hltan818@gmail.com

Abstract. Frontier Multimodal Large Language Models (MLLMs) exhibit remarkable capabilities in Visual-Language Comprehension (VLC) tasks. However, they are often deployed as zero-shot solution to new tasks in a black-box manner. Validating and understanding the behavior of these models become important for application to new task. We propose an Explicit Logic Channel, in parallel with the black-box model channel, to perform explicit logical reasoning for model validation, selection and enhancement. The frontier MLLM, encapsulating latent vision-language knowledge, can be considered as an Implicit Logic Channel. The proposed Explicit Logic Channel, mimicking human logical reasoning, incorporates a LLM, a VFM, and logical reasoning with probabilistic inference for factual, counterfactual, and relational reasoning over the explicit visual evidence. A Consistency Rate (CR) is proposed for cross-channel validation and model selection, even without ground-truth annotations. Additionally, cross-channel integration further improves performance in zero-shot tasks over MLLMs, grounded with explicit visual evidence to enhance trustworthiness. Comprehensive experiments conducted for two representative VLC tasks, *i.e.*, MC-VQA and HC-REC, on three challenging benchmarks, with 11 recent open-source MLLMs from 4 frontier families. Our systematic evaluations demonstrate the effectiveness of proposed ELC and CR for model validation, selection and improvement on MLLMs with enhanced explainability and trustworthiness.

Keywords: MLLM · Logical validation · Reliability

1 Introduction

Recently, MLLMs have advanced significantly, reflected by the frequent release of frontier models from leading AI organizations and the expansion of vision-language (VL) benchmarks in both scale and diversity [19, 36, 42, 74, 77], as well as numerous applications [45,60]. Despite these advances, recent research increasingly highlight the limitations of MLLMs in reliability, factuality, explainability, and logical reasoning [5, 14, 25, 33, 53, 64, 73]. Due to data privacy concerns, large model sizes, and the closed-source nature of models, frontier MLLMs are often used as black-boxes for zero-shot inference on new tasks. Therefore, it is crucial

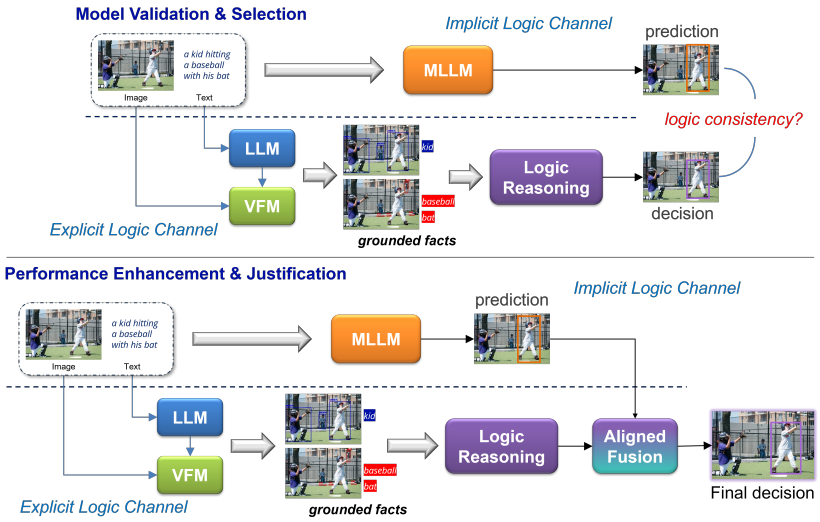

Fig. 1: Illustration of Explicit Logic Channel (ELC) for MLLM model validation and selection (upper) and performance enhancement (lower) for novel VLC application in zero-shot setting without the need of ground-truth annotation. In ELC, we combine LLM, VFM and Logical reasoning to derive a prediction on explicit and concrete visual evidence and logical validation for the VLC task.

to be able to identify reliable models and improve prediction accuracy without fine-tuning for new tasks [32, 63, 75, 76, 76, 78].

Vision-Language tasks, such as VQA and VG (Visual Grounding), in general, require a model to make a decision to a text query on visual information, needing a high-level understanding and concrete visual evidence to support the response. Recently, significant progresses have been achieved for enhancing the visual-language consistency and reliability, such as Grounded VQA [8–10, 26, 79] and VideoQA [15, 38, 66]. The advancements can be classified into three categories: Dataset, Training, and Metric. First, new datasets with additional annotations of related visual information are proposed, including attention region [8], object [9], and scene-graph [46]. Second, extensions of architecture and grounding objective function are proposed to force the model also providing grounded visual cue with the text answer on the new datasets with additional annotations [30, 31]. Third, corresponding to the formats of the extended datasets, new metrics are proposed which evaluate the performance on not only the text answer but also the visual information on the additional annotations [26, 54, 55]. These efforts focus on enhancing self-explanation of models on their responses. They might not be applicable for model validation in zero-shot applications without gt annotations.

As MLLMs are pre-trained on vast VL corpora, they are assumed to have implicitly learned human-like reasoning, forming an Implicit Logic Channel (ILC) that operates as a block-box. However, human makes decisions on explicit facts, relationships and logical rules [61]. To validate the correctness of ILC, we in-

troduce a novel Explicit Logic Channel (ELC), that runs in parallel with ILC. In ELC, a LLM is prompted to extract concept-level, task-related facts and relations from the input text. A VFM then grounds these facts explicitly in the image. Novel Logic Reasoning approaches are proposed to perform probabilistic inference on the grounded facts and relations and make the decision with explicit visual evidence. We further propose a Consistency Rate (CR) between the ILC and ELC, enabling model validation, justification, and selection, even without ground-truth (gt) annotations. The explicit visual evidence enhances user trust in predictions, while inconsistent samples are flagged for efficient manual inspection. In addition, based on auto-selected consistent sample set, we propose an aligned fusion to combine both channels for task performance enhancement and model verification without gt annotation and model fine-tuning.

To evaluate the effectiveness of ELC and CR for zero-shot VLC tasks, we conduct experiments on two representative tasks, *i.e.*, MC-VQA and HC-REC, using three recent challenging benchmarks, *i.e.*, NegBench, HC-RefCOCOg and HC-RefLoCo. We examine 11 frontier open-source MLLMs across four leading families, *i.e.*, Gemma, LLaVA, InternVL and QwenVL. Our experimental results show that, (a) the metric CR is strongly correlated with accuracy and provides reliable evaluation without gt annotation; (b) ELC with CR enables effective model validation and selection for VLC tasks without gt annotations; (c) the aligned fusion of ILC and ELC further improves VLC performance while offering explicit logical justification for enhanced trustworthiness.

Our main contributions include: (1) A general and adaptable Explicit Logic Channel with foundation models and Logic Reasoning, enabling validation, selection and enhancement of MLLMs on novel VLC tasks without gt annotation; (2) A logic consistency metric CR for model performance evaluation without requiring ground-truth; (3) A comprehensive study across 11 frontier MLLMs from four leading families, demonstrating the effectiveness of explicit logic consistency analysis for zero-shot VL applications.

2 Related work

LVLMs. Large Vision-Language Models (LVLMs) have grown rapidly in recent years [41, 70]. Representative VLMs, particularly CLIP cluster models [52, 62], and those developed on BLIP [37] and ALIGN [28], are pre-trained on massive VL corpora. Another major category is MLLMs, which adopt a pre-trained LLM as the backbone [71]. Visual features are extracted by a vision encoder, projected into the LLM's token space, and inserted into the token sequence for auto-regressive prediction. Recent MLLMs include LLaVA [43], GPT-4V [69], Gemini [22], InternVL [11], and Qwen-VL [3]. In this work, we perform evaluations on 11 frontier MLLMs from four leading families. The open-source models with less 10B parameters make them practical for real-world applications.

VLC benchmarks. Benchmarks for evaluating LVLM performance on VLC tasks have expanded rapidly [20, 41]. VLC capabilities are commonly assessed using Visual Question Answering (VQA) and Referring Expression Comprehen-

sion (REC) or Visual Grounding. Early VQA datasets typically involve multiple-choice questions or short textual answers, while later benchmarks broaden question types, domain knowledge, and tasks such as mathematical reasoning and chart understanding. Recent efforts also examine dataset bias [35] and negation understanding [2, 29]. Many standard VQA datasets have become less challenging for frontier models [41] due to extensive pre-training. Thus, we adopt a recent benchmark with negations to evaluate the effectiveness of zero-shot setting. The REC task aims to localize an image region based on a referring expression [67]. Standard datasets include RefCOCO [72], RefCOCO+ [72], and RefCOCOg [48], constructed from MS COCO. However, RefCOCO/+ might not be adequate for evaluating LVLM due to their concise referring phrases and limited vocabulary. On the other hand, zero-shot REC remains challenging [23]. Therefore, we conduct experiments on HC-RefCOCOg, which features richer descriptions (8.9 average words vs 3.3 and 3.4 in RefCOCO/+), and HC-RefLoCo [65], a recent challenge with long context expression of 93 words in average.

Logic reasoning on LLMs and VLM. Enhancing the logical reasoning capabilities of LLMs have attracted research attention [13, 51]. Prior work focuses on improving logical accuracy and consistency in generated responses. The approaches include introducing an external logic solver [16, 50] or building additional dataset with logically consistent annotations to fine-tune a model [17]. In BIRD [18], probabilistic inference is used to improve the trustworthiness of LLM. Existing NeuroSymbolic frameworks are sequential structures from Neuro-Networks to Logic Engines, and the programming methods typically require additional learning to perform logic reasoning [24, 39, 40, 68]. Different from these approaches, we propose a Dual-Channel framework and the ELC performs logic reasoning on grounded facts without additional training.

3 Methodology

A VLC task can be defined as follows: given an image I and a text expression T, the goal is to produce a logical decision D. When applying MLLM to a VLC problem, it can be formulated as $\hat{D} = \mathcal{F}_{MLLM}(I, T)$, where $\hat{D}$ denotes the predicted decision and $\mathcal{F}_{MLLM}()$ represents the function of MLLM to predict the correct decision. Recent studies show that an MLLM functions as a statistical prediction function, which predicts the next token based on the preceding sequence of visual and textual tokens, according to learned probability distributions. Hence, the function of the MLLM can be expressed as

$$\hat{D} = \mathcal{F}_{MLLM}(I, T) = \arg\max_{D \in \mathcal{D}} P_M(D|I, T), \tag{1}$$

where $\mathcal{D}$ represents a potential decision set. As MLLM makes prediction directly from the visual language inputs without explicit reasoning, it acts as a black-box. This often leads to factual inaccuracies [49] and hallucinations [5, 25, 53], especially when the model is applied to novel tasks without gt annotations.

In contrast, humans justify decisions using explicit visual evidence and concrete logic rules [61]. Various modern foundation models can provide comple-

mentary strengths that support such explicit reasoning. LLMs excel at language understanding and semantic reasoning [77], MLLMs extend this capability to vision-language descriptions and VQA [4, 11, 44], and VFMs are pretrained for fundamental visual tasks such as classification, segmentation and detection [34,56]. Although these models are trained with different annotations, they share large overlapping public datasets. Therefore, they should provide consistent and complementary information when given the same same multimodal input. Hence, we propose a Logic Channel, that operates in parallel with MLLM, to support model validation, selection, justification and enhancement without requiring ground-truth. Given a VLC problem, the logic channel can be formulated as a three-step operation. First, a set of task-relevant facts and logical relations is extracted from the text expression by prompting an LLM. This can be expressed as $(\hat{F}_s, \hat{R}_s) = \mathcal{F}_{LLM}(T|\text{prompt})$, where the subscript s indicate the semantic-level representation. Next, a VFM is employed to ground each extracted fact in the query image I, and produce corresponding confidence probabilities, *i.e.*, $\hat{F}_v = \mathcal{F}_{VFM}(I|\hat{F}_s)$, where the subscript v denotes the grounded visual evidence. Finally, basic logical rules for factual, counter-factual, and relational reasoning are applied to derive the logical inference as

$$\hat{D}_L = \mathcal{F}_{LR}(D|I,T) = \arg\max_{D\in\mathcal{D}} P_{LR}(D|\hat{F}_v, \hat{R}_s), \tag{2}$$

where the subscript LR represents Logic Reasoning.

The proposed dual-channel framework is illustrated in Figure 1. The upper channel employs an MLLM to produce prediction directly as a black box, forming the *Implicit Logic Channel* (ILC). The lower channel, mimicking human logical reasoning, employs an LLM and VFM to explicitly extract and ground facts, followed by applying logical reasoning on the grounded facts for final decision, forming the *Explicit Logic Channel* (ELC). When the foundation models understand the VLC problem well, the two channels should be consistent and complementary.The logical consistency between the two channels provides principled basis for model validation, selection, justification and enhancement in zero-shot applications.

Validation: When applying an MLLM to a new VLC task, it is common to have a set of test samples without gt annotations. In such cases, the proposed dual-channel system allows us to evaluate the logical consistency between the ILC and ELC. Let $\mathcal{Q}$ represents the test set and $q \in \mathcal{Q}$ is a test sample. We define the Consistency Rate (CR) between ILC and ELC as

$$CR = \frac{1}{|\mathcal{Q}|}\sum\nolimits_{q\in\mathcal{Q}} \mathbb{I}(\hat{D}(q) = \hat{D}_L(q)), \tag{3}$$

where the indicator function $\mathbb{I}$ returns 1 if the two channels produce consistent predictions and 0 otherwise. For different VLC task, a suitable indicator function has to be implemented. A higher CR indicates that the MLLM is more reliable and logical for the new task. In practice, CR reflects three general situations:

– Well-covered knowledge. If the knowledge required for the VLC task is well represented in the shared training datasets of the foundation models, both channels tend to produce correct predictions, resulting in high CR.

- Partially covered knowledge. If only partial knowledge is presented in the shared training data, one of the two channels may fail to perform correctly, leading to lower CR.
- Novel or OOD scenarios. If both the language expressions and visual content are novel compared to the existing training datasets, e.g., out-of-distribution (OOD) scenarios, both channels may fail, and the CR score becomes low.

Therefore, CR can be used as a principled metric for zero-shot model selection when the gt annotations are not available for a new VLC task. The discrepancies between the two channels also provide cues to user to perform manual validation and diagnose the situation without the cost of gt annotations.

Enhancement: The two channels provide complementary strengths, and fusing their outputs can further improve prediction accuracy. The logical consistency serves as a guide for aligned fusion. In principle, when the ILC and ELC predictions are logically consistent, the prediction is highly likely to be correct. Our experiment results strongly support this assumption. Let $\mathcal{Q}_c \in \mathcal{Q}$ be the consistency subset, where, for each $q \in \mathcal{Q}_c$, $\mathbb{I}(\hat{D}(q) = \hat{D}_{LR}(q)) = 1$, indicating that $\hat{D}(q)$ and $\hat{D}_{LR}(q)$ are reliable predictions from the ILC and ELC channels, respectively. For all samples in the consistent subset $\mathcal{Q}_c$, we compute the mean confidence scores of the reliable predictions for both channels as

$$\mu^c_{ILC} = \frac{1}{|\mathcal{Q}_c|} \sum\nolimits_{q \in \mathcal{Q}_c} P_M(D|q) \quad \text{and} \quad \mu^c_{ELC} = \frac{1}{|\mathcal{Q}_c|} \sum\nolimits_{q \in \mathcal{Q}_c} P_{LR}(D|q). \tag{4}$$

Without gt labels, for $q \in \mathcal{Q}_c$, predictions from both channels should be trusted equally. When applied to a new test example q_n, the probability of aligned fusion can be computed as

$$P_F(D|q_n) = P_M(D|q_n) + \mu^c_{ILC}[P_{LR}(D|q_n)/\mu^c_{ELC}], \tag{5}$$

where the prediction from ELC is first normalized on μ^c_{ELC} and re-scaled to align with ILC. Aligned scores are used for final decision on maximum. Beyond prediction, the difference between the final decision and predictions from ILC and ELC would provide additional information on the reliabilities of the channels.

3.1 Logic Consistency for MC-VQA on Factual Evidence

A representative task of VLC is multiple-choice VQA (MC-VQA). Given an image I, a question T and a set of candidate answers $\mathcal{A} = [a_1, \cdots, a_K]$, the model must choose the correct answer ($\hat{c} \in [1, K]$). A related variant replaces the question with a set of candidate captions $\mathcal{T} = [T_1, \cdots, T_K]$, where the model selects the most suitable caption ($\hat{c} \in [1, K]$). It can be formulated as

$$\hat{c} = \arg\max_{c \in [1,K]} P_M(c|I, T, \mathcal{A}) \quad \text{or} \quad \hat{c} = \arg\max_{c \in [1,K]} P_M(c|I, \mathcal{T}). \tag{6}$$

In both cases, the MLLM operates as a black-box choice function.

To evaluate joint vision-language understanding, MC-VQA datasets typically design questions or captions that reference objects or concepts that are present or

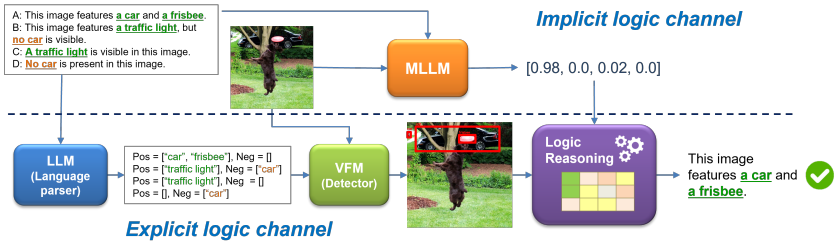

Fig. 2: ELC (Explicit Logic Channel) and logic consistency for MC-VQA task on factual and counter-factual reasoning, with an example from NegBench for illustration.

absent from the image. Logically, solving such tasks requires decisions grounded in factual evidence (presence of positive elements) and counter-factual (absence of negative elements). As modern LLMs possess powerful capabilities in language understanding, concept extraction, and semantic reasoning, for each text query, we prompt the LLM to extract positive (pos) and negative (neg) objects or concepts for explicit logical justification.

The dual-channel system for MC-VQA is presented in Figure 2, with an example from NegBench for illustration. For each question, the MLLM receives the image, the text query, and answer choices, and is prompted to output both the predicted answer and confidence values (0%-100%) for all K options (*i.e.*, $P_M()$). In the ELC, an LLM is first prompted to extract the *pos* and *neg* nouns from the text query. These nouns are then passed to a VFM. For each object category, the VFM locates all instances in the image and computes their detection probabilities. The probability of an object category is obtained as the maximum over all its detected instances. Let there be K pos object categories and L neg object categories, their presence probabilities are denoted as $\{P(O_p^k)\}_{k=1}^K$ and $\{P(O_n^l)\}_{l=1}^L$. The presence of pos objects provides factual evidence on T, while the presence of neg objects provides counter-factual evidence. The corresponding probabilities are computed logically as

$$P(pos) = \min\{P(O_p^1), \cdots, P(O_p^K)\}, \;\; P(neg) = \max\{P(O_n^1), \cdots, P(O_n^L)\}. \quad (7)$$

The final probability of factual and counter-factual evidence is computed as

$$P_{LR}(c|I,T) = \begin{cases} P(pos), & neg = \emptyset \\ 1 - P(neg), & pos = \emptyset \\ [P(pos)(1-P(neg))]^{\frac{1}{2}}, & pos \neq \emptyset \;\&\; neg \neq \emptyset \end{cases} \quad (8)$$

where the geometric mean is used for normalization as done in [47]. The ELC selects the choice with the maximum posterior $P_{LR}(T|I)$ as the correct answer.

For validation, the two channels are run independently, and CR is then computed. With a small set of logical consistent samples, the system can further enhance the performance by applying the aligned fusion strategy in Eq. (5), fusing outputs from both the ILC and ELC channels.

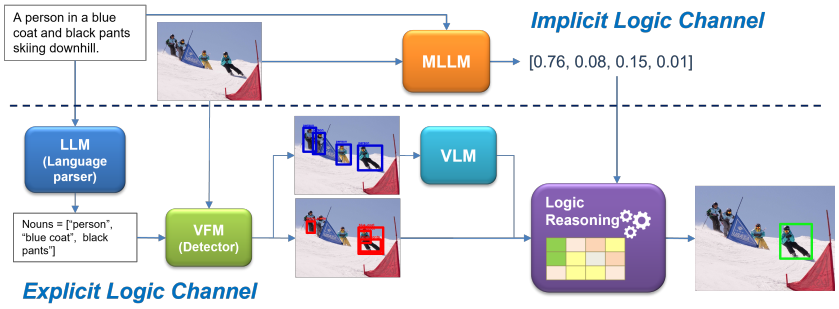

Fig. 3: ELC (Explicit Logic Channel) and logic consistency for HC-REC task on object association, with an example from HC-RefCOCOg for illustration.

3.2 Logical Consistency for HC-REC on Association

Referring Expression Comprehension (REC) is another representative task in VLC. Given an image I and a text expression T, the goal is to localize the referred object (O). Recent Human-Centric REC (HC-REC) benchmarks introduce rich or long contextual descriptions, making them challenging for VLC [65]. In HC-REC, correct prediction depends on the presence of the target person, associated objects, and understanding their visual relations in the image.

The dual-channel system for HC-REC is shown in Figure 3, with an example from HC-RefCOCOg for illustration. In ILC, the MLLM receives the image and text query, and is prompted to directly predict the bounding box coordinates of the referred person, along with its confidence score (*i.e.*, $P_M(h_m|T)$). In ELC, an LLM is prompted to extract person-related and object-related nouns from T. Then, these nouns are passed to a VFM to locate all instances of the described persons and associated objects. Each detected person is cropped, and fed to a VLM along with the referring expression to obtain a vision-language matching score. For associated objects, their presence probabilities are based on detection confidence scores. Assume that there are $(K+1)$ extracted nouns, which consist of one person H and K object concepts, denoted as $\{H, O_1, \cdots, O_K\}$. The VFM locates a set of instances for person and objects. For each detected instance, let h_i denote the ith detected person, and o_{jl} the lth instance of object O_j.

To estimate the probability of each person being the target, we adopt an evidence-accumulation approach, as inspired by the principle of perceptual decision-making [6]. Formally, it can be expressed as

$$P_{LR}(h_i|T) = \frac{1}{K+1}\left(\sum\nolimits_{k=1}^{K} P(O_k|h_i) + P(h_i|H)\right), \tag{9}$$

where $P(h_i|H)$ represents the presence probability of person h_i, and $P(O_k|h_i)$ denotes its association with object O_k. Assuming that there are L detected instances of object O_k, i.e., $\{o_{kl}\}_{l=1}^{L}$, the association rate between h_i and o_{kl} can be computed as

$$R_A(o_{kl}, h_i) = A_{int}(o_{kl}, h_i)/A(o_{kl}), \tag{10}$$

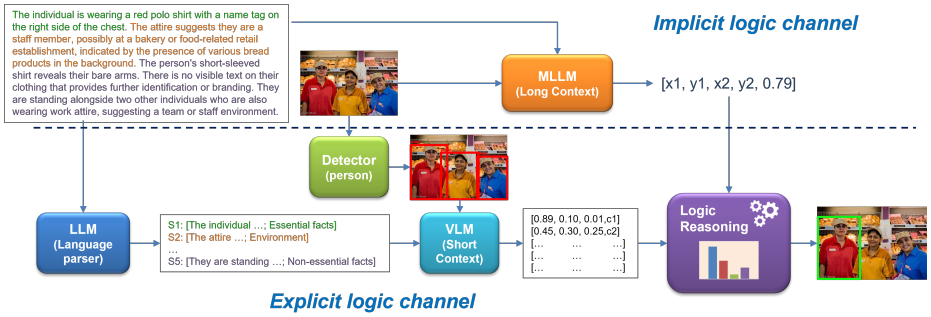

Fig. 4: ELC (Explicit Logic Channel) and logic consistency for task of HC-REC on long context text query, with an example from HC-RefLoCo for illustration.

where $A_{int}(o_{kl}, h_i)$ denotes the intersection area of o_{kl} and h_i, and $A(o_{kl})$ is the area of o_{kl}. The overall association probability between h_i and O_k becomes

$$P(O_k|h_i) = \max_{l\in[1,L]} \{R_A(o_{kl}, h_i)\}. \tag{11}$$

Finally, the person with the highest accumulated probability $P_{LR}(h_i|T)$ is selected as the grounded person in the ELC.

For model validation, both channels operate independently and the CR is computed across multiple IoU levels, *e.g.*, IoU thresholds of 0.5, 0.75, and 0.9. When fusing two channels for performance enhancement, we incorporate IoU-weighted probabilities for each candidate person, and extend the aligned fusion Eq. (5) to HC-REC as

$$\begin{cases} P_F(h_m|T) = P_M(h_m|T) + \max_{i\in[1,K]}[IoU_{mi} \cdot (\frac{\mu^c_{ILC}}{\mu^c_{ELC}} P_{LR}(h_i|T))] \\ P_F(h_i|T) \ = IoU_{im} \cdot P_M(h_m|T) + \frac{\mu^c_{ILC}}{\mu^c_{ELC}} P_{LR}(h_i|T) \end{cases} \tag{12}$$

The person with the highest $P_F()$ value is selected as the final prediction.

3.3 Logic Consistency for HC-REC on Correlation

Recent VLC tasks involving very long referring expressions have introduced new challenges for MLLM, especially in zero-shot settings. With the powerful capabilities of LLMs in language understanding, summarization and semantic categorization, we can leverage these strengths by prompting the LLM to extract essential visual facts from long text descriptions, and use them to guide explicit logical reasoning (*i.e.*, ELC) for visual-language verification.

In HC-REC with very long context input, the challenge is not only the text length, but also the presence of many non-informative or weakly relevant sentences. For VLC, sentences containing essential visual facts are far more informative for locating the target person than sentences describing general context or unrelated events [12]. Consider the following example of long referring expression for an image of human activities in a public park: "*This is a scene of*

a public park in Sunday. People play various activities in the morning. A young lady in red sportswear is sitting on a bench for a rest. A blue bottle is placed on the bench beside her." The first sentence (S_1) is a description of the environment, and does not lead the attention to the target young lady (H). The second (S_2) is not closely related and may incorrectly draw attention to every person in I. The third (S_3) provides the core description that uniquely identifies the target H. The fourth (S_4) provides secondary but relevant information. Following the recommendations in [12], we prompt the LLM to classify each sentence into one of the three categories, *i.e.*, Essential Fact, Non-Essential Fact or Environment, for the task of HC-REC. The full prompt template is presented in Suppl. Mat.

The dual-channel system for HC-REC with long context expression is presented in Figure 4, with an example from HC-RefLoCo for illustration. The ILC operates the same as in Figure 3, while the ELC is adapted to focus on the effectiveness of sentence-level information. The long context expression is first split into sentences and the LLM is prompted to classify each sentence into one of the three categories. Then, a VFM detects all persons in the image and each person box is cropped. Each cropped person patch is paired with each sentence to form a VL pair, and fed to a VLM to obtain a VL matching probability. If there are N detected persons and K sentences, the set of predicted Human-Sentence matches can be denoted as $\{\{P_{VLM}(h_n|S_k)\}_{k=1}^{K}\}_{n=1}^{N}$. Naturally, pairings between Essential Facts and the correct person would results in higher probabilities, while other pairs produce lower matching scores. Following the principle of human decision-making theory [7], the final logical prediction can obtained as a weighted function

$$P_{LR}(h_n|T) = \sum\nolimits_{k=1}^{K} P_{VLM}(h_n|S_k)u(S_k) \tag{13}$$

where $u(S_k)$ is a utility weight of S_k that reflects its informativeness for perceptual decision-making. On common-sense knowledge, the sentences of Essential Facts should receive higher utility weight, Non-Essential Facts sentences should have moderate weight, and the Environment sentences should receive very low weight. This common-sense weighting is effective for perceptual decision-making especially in zero-shot application scenarios. The person with the highest accumulated score is selected as logical prediction of ELC.

Model validation and enhancement follow the same procedure described in previous subsection 3.2 on HC-REC task.

4 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of ELC and CR for model validation and enhancement when applying MLLMs to novel VLC tasks under zero-shot settings without gt annotations. Evaluations are performed across two representative VLC tasks on three recent challenging benchmarks, with 11 frontier open-source MLLMs from four model families.

Tasks and Datasets:

NegBench [2] is a recent benchmark designed to assess visual–language understanding under factual and counter-factual reasoning. Each sample contains both

positive and negative phrases, and choices follow three linguistic templates: Affirmation, Negation, and Hybrid. An affirmation text contains only positive elements $\{pos\}$, i.e., objects present in the image. A negation text includes only negative elements $\{neg\}$, i.e., objects absent from the image but commonly associated with the present objects. A hybrid text contains both positive and negative elements. There are three natural image MCQ tasks, *i.e.*, COCO, VOC2007, and HardNeg-Syn. Our experiments are conducted on the publicly available data, i.e., 5914 MCQ questions on COCO and 5032 MCQ questions on VOC2007.

Recent HC-REC (Human-Centric Referring Expression Comprehension) benchmarks, extended from general REC task, are created for evaluating VLC capabilities in MLLMs [65]. RefCOCOg [48] is created from COCO with enriched phrases (average 8.9 words), as compared to RefCOCO/+ (3.3 and 3.4 words), providing a more challenging scenario.

HC-RefLoCo (Human-Centric Referring Expression Comprehension with Long Context) [65] is a recent challenging benchmark featuring long-context expressions with an average of 93.2 words. It is built upon various image datasets, including COCO, Objects365, OpenImage v7, and LAION-5B, covering a wide range of real-world scenes. These datasets allow us to examine the robustness of MLLMs under both moderate-length and extremely long referring expressions.

Experimental settings: ILC predictions are generated using the default prompt templates provided by each benchmark. In ELC, we employ Qwen3-4B-Instruct [59] as the LLM, EvaCLIP-8B [57] as the VLM for HC-RefCOCOg and InternVL2.0-8B [11] as the VLM for HC-RefLoCo. All models are moderate-sized, with no fine-tuning or additional optimization is applied. Extensive ablation evaluations on ELC, including reliability on model variations, critical false positives, confusion matrix, and latency analysis, can be found in the Suppl.

Experimental objectives:
Our experiments are designed to answer three key questions: (1) Is CR score meaningful without gt. To assess whether CR can serve as a reliable metric in zero-shot evaluation, we compute CR between ILC and ELC, and compute accuracy (Acc) based on gt for analysis. We measure their relationship using three representative correlation coefficients [1]: Pearson's r (r), Spearman's rho (ρ), and Kendall's tau (τ). r compares the scores of the metrics, while ρ and τ compare the rankings on the metrics. Each coefficient ranges from -1 to 1, where values near 1 indicate strong positive correlation, values between 0.5 and -0.5 indicate weak or no correlation, and values near -1 indicate strong negative correlation. (2) Is CR score helpful for model validation and selection without gt. On each of the three benchmarks, we evaluate 11 frontier MLLMs and analyze their performance, e.g., reliable predictions, strong models. (3) Effectiveness of ELC and CR in enhancing MLLM performance for VLC without gt. We evaluate the effect of aligned fusion between ILC and ELC, and examine the performance gain across benchmarks. Our results support that ELC and CR provide consistent performance improvements without requiring any additional training, confirming its practicality for zero-shot applications.

4.1 Experimental Results

MC-VQA Results on NegBench: Table 1 summarizes the performance of the 11 MLLMs on NegBench. The column Acc/Rank reports the Accuracy score and its corresponding ranking. CR/Rank presents the CR score and its ranking, and Acc_E shows the Accuracy of enhancement after applying aligned fusion. We observe that both InternVL and Qwen are strong model families, where three out of four models achieve high Acc/Rank and CR/Rank across COCO and VOC2007 datasets.

Table 1: Results on NegBench for MC-VQA task.

NegBench	COCO			VOC2007		
MLLM	Acc/Rank	CR/Rank	Acc_E	Acc/Rank	CR/Rank	Acc_E
Gemma-3-12B [58]	0.7198/6	0.6715/6	0.7201	0.8750/5	0.8006/5	0.8752
InternVL2.0-8B [11]	0.4878/9	0.3980/10	0.8434	0.5868/10	0.5198/10	0.9352
InternVL2.5-8B [11]	0.9121/2	0.7038/4	**0.9650**	0.9201/3	0.8076/4	**0.9809**
InternVL3.0-8B [11]	**0.9319**/1	**0.7259**/1	0.9557	**0.9537**/1	**0.8424**/1	0.9684
InternVL3.5-8B [11]	0.8429/3	0.6583/7	0.9324	0.9221/2	0.8088/3	0.9750
LLaVA-1.5-13B [44]	0.6206/8	0.5810/8	0.6206	0.7738/8	0.7112/8	0.7738
LLaVA-1.6-13B [44]	0.3898/10	0.4004/9	0.4499	0.6528/9	0.6096/9	0.6897
QwenVL-7B-Chat [4]	0.2426/11	0.2310/11	0.6245	0.2371/11	0.2314/11	0.5560
Qwen2.0-VL-7B [4]	0.6968/7	0.6743/5	0.7076	0.8535/6	0.7822/6	0.8764
Qwen2.5-VL-7B [4]	0.8165/4	0.7109/3	0.8356	0.9197/4	0.8374/2	0.9223
Qwen3.0-VL-8B [4]	0.7753/5	0.7134/2	0.7758	0.8408/7	0.7688/7	0.8408
ELC	0.7577			0.8684		

HC-REC Results on HC-RefCOCOg and HC-RefLoCo:
The results of 11 MLLMs on HC-RefCOCOg are presented in Table 2. The column Acc/50/75/90/R denote Accuracy scores obtained at IoU thresholds 0.5, 0.75 and 0.9, and the ranking based on Acc50. Corresponding CR is also computed at IoU thresholds 0.5, 0.75, and 0.90. CR50 scores are presented here (see full results in Suppl. Mat.). The column CR/R denotes CR50 scores and its ranking. The enhanced performance is evaluated on test set with the alignment weight obtained on val set. The Accuracy scores of enhanced performance are added in column A_E/50/75/90 of test set. Table 3 presents the results of 11 MLLMs on HC-RefLoCo. Here, the column $mAcc$ denotes the mean Accuracy over IoU from 0.1 to 0.9 with step size 0.1. Again, enhanced performance is evaluated on test set. Only Acc50 and $mAcc$50 are presented (see Suppl. Mat.).
Correlation Analysis: Based on the accuracy and CR scores across the three benchmark tables, we compute the correlation coefficients between Acc and CR for all models and presented in Table 4.
Latency Analysis: Naturally, ELC introduces additional time cost. The main time costs are spent by LLM for language parser, VFM for human and object grounding, and VLM for vision-language association. The ELC time cost information on the three benchmarks are presented in Table 5.

Table 2: Results on HC-RefCOCOg for HC-REC task.

HC-RefCOCOg	val		test		
MLLM	Acc/50/75/90/R	CR/R	Acc/50/75/90/R	CR/R	A_E/50/75/90
Gemma-3-12B	.088/.004/.001/10	.081/11	.088/.004/.001/10	.082/11	.752/.690/.587
InternVL2.0-8B	.751/.657/.386/4	.665/4	.755/.672/.408/4	.663/3	.841/.777/.637
InternVL2.5-8B	.765/.710/.517/3	.670/3	.775/.722/.533/3	.657/4	.850/.791/.659
InternVL3.0-8B	.699/.620/.403/6	.618/6	.713/.639/.427/6	.613/7	.837/.771/.633
InternVL3.5-8B	.780/.735/.597/2	.681/2	.808/.761/.610/2	.681/2	.854/.796/.670
LLaVA-1.5-13B	.219/.069/.012/9	.224/9	.236/.076/.015/9	.231/9	.768/.695/.586
LLaVA-1.6-13B	.076/.007/.001/11	.089/10	.082/.005/.002/11	.087/10	.752/.693/.591
QwenVL-7B-Chat	.391/.361/.300/8	.411/8	.391/.365/.305/8	.395/8	.780/.724/.616
Qwen2.0-VL-7B	.715/.631/.489/5	.629/5	.735/.647/.507/5	.629/5	.833/.769/.643
Qwen2.5-VL-7B	.684/.591/.395/7	.612/7	.709/.620/.390/7	.618/6	.838/.776/.642
Qwen3.0-VL-8B	.804/.749/.597/1	.703/1	.818/.777/.632/1	.692/1	**.856/.798/.673**
ELC	.807/.747/.628		.801/.743/.634		

Table 3: Results on HC-RefLoCo for HC-REC task.

HC-RefLoCo	val			test				
MLLM	Acc50	$mAcc$/R	CR/R	Acc50	$mAcc$/R	CR/R	Acc_E50	$mAcc_E$
Gemma-3-12B	.0921	.0219/10	.0860/10	.0920	.0220/10	.0860/10	.4270	.3500
InternVL2.0-8B	.7592	.5403/4	.6802/4	.7534	.5350/4	.6860/4	.8120	.6412
InternVL2.5-8B	.5039	.3048/7	.4447/7	.4970	.3041/7	.4478/7	.6618	.5144
InternVL3.0-8B	.6817	.4790/6	.6100/6	.6742	.4720/6	.6095/6	.7630	.6071
InternVL3.5-8B	.8832	.6718/2	.7691/2	.8866	.6733/2	.7746/2	.9003	.7323
LLaVA-1.5-13B	.1906	.0821/9	.1936/9	.1957	.0834/9	.1966/9	.4304	.3353
LLaVA-1.6-13B	.0626	.0140/11	.0612/11	.0621	.0144/11	.0589/11	.3335	.2797
QwenVL-7B-Chat	.3284	.2687/8	.3313/8	.3314	.2699/8	.3332/8	.4959	.4303
Qwen2.0-VL-7B	.8396	.6532/3	.7375/3	.8451	.6542/3	.7449/3	.8652	.7123
Qwen2.5-VL-7B	.7502	.5283/5	.6604/5	.7522	.5288/5	.6650/5	.8260	.6526
Qwen3.0-VL-8B	.9412	.7992/1	.8045/1	.9418	.8024/1	.8128/1	.9388	**.8143**
ELC	.804	.723		.815	.734			

4.2 Evaluation and Discussion

Effectiveness of CR: As shown in Table 4, the correlation between CR and Acc is consistently strong across all benchmarks. All the correlation measures, r, ρ and τ, exceed 0.89, except on COCO of NegBench where ρ>0.83 and τ>0.67, still well above 0.5. The average scores of the three metrics exceed 0.9, indicating very strong correlation between CR with Acc. These results confirm thatCR is a reliable indicator of model performance even without gt annotation. Consistent but Incorrect (CI) errors might cause risks, especially when CR is high. We evaluated such cases. As example, on HC-RefCOCOg with Qwen3.0-VL, CR is 0.69 (top 1), the rate of CI is 4.7%, and manual examination found 52% caused by inconsistency with gt boxes, 32% related to bias of models, and only 15% caused by ELC errors. Once CR is high, CI risk is low. Full results and visual examples can be found in Suppl. Sec 2.2.

Effectiveness for Model Validation and Selection: From the columns of CR and Acc, we observe substantial performance variation among recent frontier MLLMs. For example, CR on NegBench COCO set ranges widely from 0.23 to

Table 4: Correlation coefficients between CR and Acc.

Benchmark	NegBench		HC-RefCOCOg		HC-RefLoCo		All
Cor. Coefficient	COCO	VOC2007	val	test	val	test	mean
Pearson's r	0.9543	0.9968	0.9972	0.9973	0.9987	0.9987	0.9905
Spearman's ρ	0.8364	0.9727	0.9909	0.9727	1.0000	1.0000	0.9621
Kendall's τ	0.6727	0.9273	0.9636	0.8909	1.0000	1.0000	0.9091

Table 5: Latency analysis on ELC on the tasks of the three benchmarks.

Benchmark	LLM (Language Parser)	VFM (Object Grounding)	VLM (Visual-Text Association)
NegBench	Mistral: 1.80s	GroundingDINO: 0.76s	-
HC-RefCOCOg	Mistral: 1.77s	GroundingDINO: 0.46s	EvaCLIP: 1.45s
HC-RefLoCo	Mistral: 1.55s	GroundingDINO: 0.35s	EvaCLIP: 6.88s

0.73. Even within the InternVL family, CR varies from 0.40 to 0.73. Besides, newer version models are not guaranteed to outperform earlier ones, even from the same family, *e.g.*, InternVL-3.0 is better than InternVL-3.5 on NegBench. In additional, a model that excels on one task may perform poorly on another, *e.g.* Gemma-3 achieves strong results on MC-VQA but fails on HC-REC. These findings highlight that model selection is challenging in zero-shot scenarios due to inconsistent model performance across tasks. The proposed ELC and CR provide a principled mechanism to validate and select suitable models without requiring annotated datasets. In addition, the inconsistent samples would lead to efficient manual examination and reveal informative insights (see Suppl.).

Effectiveness for Performance Enhancement: To validate the aligned fusion assumption, we compute CR_{gt} on samples that are both consistent and correct when given ground-truth. The ratio CR_{gt}/CR is computed for every model across all tasks. Even when CR varies widely from 0.1 to 0.95, the ratio remains above 80%. The mean of ratio is 0.92 at CR=0.5508, and increases linearly with CR, indicating that higher consistency leads to higher correctness. However, even the rate is low, the cases of consistent but incorrect (not matching gt label) might raise concerns on the validation of CR. Additional investigation is performed (see Suppl. Sec 2.2).

Aligned fusion yields consistent improvements over MLLMs across all three benchmarks. Even the strongest model benefits from ELC-guided enhancement. For example, on NegBench COCO, InternVL2.5 Acc score increases from 0.912 to 0.965. All the top Acc scores establish new SOTA performance on the benchmarks, achieved on recent frontier MLLMs without fine-tuning. Fusion makes changes of the predictions from both ILC and ELC. The comparison of the changes would also verify the trustworthiness of ILC and ELC even without gt annotation, see Suppl. Sec 4.4.

Confusion matrices between two channels: With gt annotation, we categorize the matching of two channels as 4 cases: WW (win-win), WL (ILC win but ELC lost), LW (ILC lost but ELC win), and LL (lost-lost), and generate the confusion matrix (see Suppl. Sec 2.3). It is observed that, if CR score is

high, the percentage of WW is high. The relative difference between ILC and ELC leads to the difference of WL and LW. If ILC is stronger than ELC, the percentage of WL is larger than LW, and vice versa. Here, WL corresponds to FN (Inconsistent but ILC is Correct). CR is strongly correlated with WW and less affected by WL errors.

Ablation study on Lose Cascading Effects in ELC: 100 samples are randomly selected from HC-RefCOCOg val set, and then GT nouns and boxes are added manually. ELC performance is evaluated on different VFMs, on the nouns extracted by Mistrial (LLM), on the Ground-Truth nouns annotated manually, and on the Ground-Truth boxes of related objects labeled manually. Full results are presented in Suppl Sec 4.3. In summary, the errors of noun extraction by LLM may cause 10% gap and box allocation by cause another 10% gap on HC-RefCOCOg. We also performed ablation study on adversarial images. We manually mask one critical object in each image on 100 images, and test with different models. As example, on Qwen3.0-VL, Acc50 of ILC drops from 91% to 60%, while Acc50 of ELC drops from 82% to 41%, and CR drops from 0.80 to 0.57. It is observed that ELC is more sensitive to absence of referred objects.

Adaption to new tasks: For a new task, user knows critical facts and logic rules for safety of decisions. One may employ AI tools to select LLM, VFM, design prompt, and implement logic inference [21, 27] to build ELC for the task.

5 Conclusion

Facing the uncertainty on reliability when deploying frontier MLLMs to new tasks, we proposed an Explicit Logic Channel for model validation and enhancement without any re-training or fine-tuning. Leveraging foundation models and logical reasoning [21], ELC produces decisions grounded on explicit facts and relations, providing a principled way for model validation. We investigated the effectiveness of proposed framework on three challenging benchmarks using 11 frontier MLLMs. The CR serves as reliable metric for model validation, selection and enhancement, with enhanced explainability and trustworthiness. The diversity of the three VLC tasks demonstrates the generality and flexibility of our approach in real-world applications. Extending ELC to more complex multimodal CoT reasoning tasks is a promising direction for future work.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). 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. The authors would like to thank the anonymous reviewers for their constructive comments and valuable suggestions, which helped improve the quality and clarity of this paper.

References

1. Akoglu, H.: User's guide to correlation coefficients. Turkish Journal of Emergency Medicine **18**(3), 91–93 (2018). https://doi.org/https://doi.org/10.1016/j.tjem.2018.08.001, https://www.sciencedirect.com/science/article/pii/S2452247318302164
2. Alhamoud, K., Alshammari, S., Tian, Y., Li, G., Torr, P., Kim, Y., Ghassemi, M.: Vision-Language Models Do Not Understand Negation. In: CVPR-25 (2025)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond (2023), https://arxiv.org/abs/2308.12966
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
5. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey (2025), https://arxiv.org/abs/2404.18930
6. Balsdon, T., Wyart, V., Mamassian, P.: Confidence controls perceptual evidence accumulation. Nat Commun **11**(1753) (2020). https://doi.org/https://doi.org/10.1038/s41467-020-15561-w
7. Bradley, R.: Decision Theory: A Formal Philosophical Introduction. In: Hansson, S.O., Hendricks, V.F. (eds.) Introduction to Formal Philosophy, pp. 611–655. Springer International Publishing (2018)
8. Chen, C., Anjum, S., Gurari, D.: Grounding answers for visual questions asked by visually impaired people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19098–19107 (June 2022)
9. Chen, C., Anjum, S., Gurari, D.: Vqa therapy: Exploring answer differences by visually grounding answers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15315–15325 (2023)
10. Chen, Y., Su, L., Chen, L., Lin, Z.: Lcv2: a universal pretraining-free framework for grounded visual question answering. Electronics **13**(11), 2061 (2024)
11. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24185–24198 (June 2024)
12. Cheng, A., Jacovi, A., Globerson, A., Golan, B., Kwong, C., Alberti, C., Tao, C., Ben-David, E., Tomar, G.S., Haas, L., Bitton, Y., Bloniarz, A.E., Bai, A., Wang, A., Siddiqui, A., Castillo, A.B., Atias, A., Liu, C., Fry, C., Balle, D., Ghosal, D., Kukliansky, D., Marcus, D., Gribovskaya, E., Ofek, E.O., Zhuang, H., Laish, I., Ackermann, J., Wang, L., Risdal, M., Barnes, M., Fink, M., Amin, M., Ambar, M., Potikha, N., Gupta, N., Katz, N., Velan, N., Roval, O., Ram, O., Zablotskaia, P., Bang, P., Agrawal, P., Ghiya, R., Ganapathy, S., Baumgartner, S., Erell, S., Prakash, S., Sellam, T., Rao, V., Wang, X., Akulov, Y., Yang, Y., Yang, Z.B.O., Lai, Z., Wu, Z., Dragan, A., Hassidim, A., Pereira, F., Petrov, S., Venkatachary, S., Doshi, T., Matias, Y., Goldshtein, S., Das, D.: The facts leaderboard: A comprehensive benchmark for large language model factuality. ArXiv **abs/2512.10791** (2025), https://api.semanticscholar.org/CorpusID:283737312
13. Cheng, F., Li, H., Liu, F., van Rooij, R., Zhang, K., Lin, Z.: Empowering LLMs with Logical Reasoning: A Comprehensive Survey (2025), https://arxiv.org/abs/2502.15652

14. Dang, Y., Huang, K., Huo, J., Yan, Y., Huang, S., Liu, D., Gao, M., Zhang, J., Qian, C., Wang, K., Liu, Y., Shao, J., Xiong, H., Hu, X.: Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey (2024), https://arxiv.org/abs/2412.02104
15. Di, S., Xie, W.: Grounded question-answering in long egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12934–12943 (2024)
16. Diego, C., Stefano, T., Antonio, V.: Logically Consistent Language Models via Neuro-Symbolic Integration. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=7PGluppo4k
17. Feng, J., Xu, R., Hao, J., Sharma, H., Shen, Y., Zhao, D., Chen, W.: Language Models can be Deductive Solvers. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024. pp. 4026–4042. Association for Computational Linguistics, Mexico City, Mexico (June 2024). https://doi.org/10.18653/v1/2024.findings-naacl.254, https://aclanthology.org/2024.findings-naacl.254/
18. Feng, Y., Zhou, B., Lin, W., Roth, D.: BIRD: A Trustworthy Bayesian Inference Framework for Large Language Models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=fAAaT826Vv
19. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), https://arxiv.org/abs/2306.13394
20. Fu, C., Zhang, Y.F., Yin, S., Li, B., Fang, X., Zhao, S., Duan, H., Sun, X., Liu, Z., Wang, L., Shan, C., He, R.: MME-Survey: A Comprehensive Survey on Evaluation of Multimodal LLMs (2024), https://arxiv.org/abs/2411.15296
21. Gabbay, D.M., Hogger, C.J., Robinson, J.A., Siekmann, J. (eds.): Handbook of Logic in Artificial Intelligence and Logic Programming. Oxford University Press (03 1994). https://doi.org/10.1093/oso/9780198537465.001.0001, https://doi.org/10.1093/oso/9780198537465.001.0001
22. Gemini Team: Gemini: A Family of Highly Capable Multimodal Models (2025), https://arxiv.org/abs/2312.11805
23. Han, Z., Zhu, F., Lao, Q., Jiang, H.: Zero-shot referring expression comprehension via structural similarity between images and captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14364–14374 (2024)
24. Huang, J., Li, Z., Jacobs, D., Naik, M., Lim, S.N.: Laser: A neuro-symbolic framework for learning spatio-temporal scene graphs with weak supervision. In: International Conference on Learning Representations (2025), https://api.semanticscholar.org/CorpusID:258179880
25. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (Jan 2025). https://doi.org/10.1145/3703155, http://dx.doi.org/10.1145/3703155
26. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6693–6702 (2019). https://doi.org/10.1109/CVPR.2019.00686

27. Huth, M., Ryan, M.: Logic in Computer Science: Modelling and Reasoning about Systems. Cambridge University Press, 2 edn. (2004)
28. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision (2021), https://arxiv.org/abs/2102.05918
29. Kamath, A., Hsieh, C.Y., Chang, K.W., Krishna, R.: The hard positive truth about vision-language compositionality. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 37–54. Springer Nature Switzerland, Cham (2025)
30. Khan, A.U., Kuehne, H., Duarte, K., Gan, C., Lobo, N., Shah, M.: Found a reason for me? weakly-supervised grounded visual question answering using capsules. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8461–8470 (2021). https://doi.org/10.1109/CVPR46437.2021.00836
31. Khan, A.U., Kuehne, H., Gan, C., Lobo, N.D.V., Shah, M.: Weakly supervised grounding for vqa in vision-language transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 652–670. Springer Nature Switzerland, Cham (2022)
32. Khan, Z., Fu, Y.: Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10854–10863 (June 2024)
33. Kil, J., Mai, Z., Lee, J., Wang, Z., Cheng, K., Wang, L., Liu, Y., Chowdhury, A., Chao, W.L.: Mllm-compbench: A comparative reasoning benchmark for multimodal llms (2025), https://arxiv.org/abs/2407.16837
34. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: International Conference on Learning Representations (2023)
35. Li, B., Lin, Z., Peng, W., Nyandwi, J.d.D., Jiang, D., Ma, Z., Khanuja, S., Krishna, R., Neubig, G., Ramanan, D.: Naturalbench: Evaluating vision-language models on natural adversarial samples. In: Advances in Neural Information Processing Systems. vol. 37, pp. 17044–17068. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/1e69ff56d0ebff0752ff29caaddc25dd-Paper-Datasets_and_Benchmarks_Track.pdf
36. Li, J., Lu, W., Fei, H., Luo, M., Dai, M., Xia, M., Jin, Y., Gan, Z., Qi, D., Fu, C., Tai, Y., Yang, W., Wang, Y., Wang, C.: A survey on benchmarks of multimodal large language models (2024), https://arxiv.org/abs/2408.08632
37. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (2022), https://proceedings.mlr.press/v162/li22n.html
38. Li, Y., Wang, X., Xiao, J., Ji, W., Chua, T.S.: Invariant grounding for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2928–2937 (2022)
39. Li, Z., Huang, J., Naik, M.: Scallop: A Language for Neurosymbolic Programming. Proc. ACM Program. Lang **7**, 166:1–166:25 (2023)
40. Li, Z., Huang, J., Naik, M.: Scallop: A language for neurosymbolic programming (2023), https://arxiv.org/abs/2304.04812

41. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops. pp. 1587–1606 (June 2025)
42. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops. pp. 1587–1606 (June 2025)
43. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning (2023), https://arxiv.org/abs/2304.08485
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
45. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), https://arxiv.org/abs/2307.06281
46. Ma, H., Pathak, V., Wang, D.Z.: Bridging vision language models and symbolic grounding for video question answering. CoRR **abs/2509.11862** (2025). https://doi.org/10.48550/ARXIV.2509.11862, https://doi.org/10.48550/arXiv.2509.11862
47. Malinin, A., Gales, M.: Uncertainty estimation in autoregressive structured prediction. arXiv preprint arXiv:2002.07650 (2020)
48. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
49. Muhammad Ahsan Awais Muhammad Ahsan Awais, Peter Redmond, T.E.W., Healy, G.: Amber: advancing multimodal brain-computer interfaces for enhanced robustness—a dataset for naturalistic settings. Front. Neuroergonomics **4** (2023). https://doi.org/10.3389/fnrgo.2023.1216440
50. Olausson, T., Gu, A., Lipkin, B., Zhang, C., Solar-Lezama, A., Tenenbaum, J., Levy, R.: LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5153–5176. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.313, https://aclanthology.org/2023.emnlp-main.313/
51. Pournemat, M., Rezaei, K., Sriramanan, G., Zarei, A., Fu, J., Wang, Y., Eghbalzadeh, H., Feizi, S.: Reasoning under uncertainty: Exploring probabilistic reasoning capabilities of llms (2025), https://arxiv.org/abs/2509.10739
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), https://proceedings.mlr.press/v139/radford21a.html
53. Rahman, S.S., Islam, M.A., Alam, M.M., Zeba, M., Rahman, M.A., Chowa, S.S., Raiaan, M.A.K., Azam, S.: Hallucination to truth: a review of fact-checking and factuality evaluation in large language models. Artificial Intelligence Review **59**(2) (Jan 2026). https://doi.org/10.1007/s10462-025-11454-w, http://dx.doi.org/10.1007/s10462-025-11454-w
54. Reich, D., Putze, F., Schultz, T.: Visually grounded vqa by lattice-based retrieval (2022), https://arxiv.org/abs/2211.08086

55. Reich, D., Putze, F., Schultz, T.: Measuring faithful and plausible visual grounding in VQA. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 3129–3144. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.206, https://aclanthology.org/2023.findings-emnlp.206/
56. Shen, H., Zhao, T., Zhu, M., Yin, J.: Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4766–4775 (2024)
57. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2023)
58. Team, G.: Gemma 3 (2025), https://goo.gle/Gemma3Report
59. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
60. Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., Liu, M., Gu, P., Xia, S., Li, W., Zhang, Y., Wu, Z., Liu, Z., Zhong, T., Ge, B., Zhang, T., Qiang, N., Hu, X., Jiang, X., Zhang, X., Zhang, W., Shen, D., Liu, T., Zhang, S.: A comprehensive review of multimodal large language models: Performance and challenges across different tasks (2024), https://arxiv.org/abs/2408.01319
61. Wang, P., Hahm, C., Hammer, P.: A Model of Unified Perception and Cognition. Frontiers in Artificial Intelligence **Volume 5 - 2022** (2022). https://doi.org/10.3389/frai.2022.806403, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.806403
62. Wang, T., Lin, K., Li, L., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Equivariant similarity for vision-language foundation models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 11964–11974 (2023), https://api.semanticscholar.org/CorpusID:257766245
63. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models (2023), https://arxiv.org/abs/2203.11171
64. Wang, Y., Wang, M., Manzoor, M.A., Liu, F., Georgiev, G., Das, R.J., Nakov, P.: Factuality of large language models: A survey (2024), https://arxiv.org/abs/2402.02420
65. Wei, F., Zhao, J., Yan, K., Zhang, H., Xu, C.: A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 69566–69587. Curran Associates, Inc. (2024)
66. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13204–13214 (June 2024)
67. Xiao, L., Yang, X., Lan, X., Wang, Y., Xu, C.: Towards Visual Grounding: A Survey (2024), https://arxiv.org/abs/2412.20206
68. Yang, X., Liu, F., Lin, G.: Neural Logic Vision Language Explainer. IEEE Transactions on Multimedia **26**, 3331–3340 (2024)
69. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision) (2023), https://arxiv.org/abs/2309.17421
70. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What you see is what you read? improving text-image alignment

evaluation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
71. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12) (2024). https://doi.org/10.1093/nsr/nwae403, http://dx.doi.org/10.1093/nsr/nwae403
72. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016)
73. Yuan, J., Peng, T., Jiang, Y., Lu, Y., Zhang, R., Feng, K., Fu, C., Chen, T., Bai, L., Zhang, B., Yue, X.: Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms (2025), https://arxiv.org/abs/2505.21327
74. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
75. Zhao, J.X., Xie, Y., Kawaguchi, K., He, J., Xie, M.Q.: Automatic model selection with large language models for reasoning (2023), https://arxiv.org/abs/2305.14333
76. Zhao, T.Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models (2021), https://arxiv.org/abs/2102.09690
77. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.Y., Wen, J.R.: A Survey of Large Language Models (2025), https://arxiv.org/abs/2303.18223
78. Zhi, Z., Feng, C., Daneshmend, A., Orlu, M., Demosthenous, A., Yin, L., Li, D., Liu, Z., Rodrigues, M.R.D.: TFAR: A training-free framework for autonomous reliable reasoning in visual question answering. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=cBAKeZN3jy
79. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4995–5004 (2016). https://doi.org/10.1109/CVPR.2016.540