

000 001 LOGIC CHANNEL VALIDATION AND ENHANCEMENT 002 OF ZERO-SHOT VISION-LANGUAGE COMPREHENSION 003 ON VISION LANGUAGE MODELS 004 005

006 **Anonymous authors**
007 Paper under double-blind review
008
009
010
011

ABSTRACT

013 Frontier Large Vision-Language Models (LVLMs) exhibit remarkable capabilities
014 in Visual-Language Comprehension (VLC) tasks, enabled by pretraining on
015 vast visual-textual corpus. However, they are often deployed as zero-shot solution
016 in a black-box manner, as retraining challenges remain due to data privacy
017 or model inaccessibility. Validating and understanding the behavior of the models
018 become important for generalization to new task. We propose a Logic Channel,
019 in parallel with the black-box model channel, to perform explicit logic reasoning
020 for validation and enhancement. The frontier LVLM, encapsulating latent vision-
021 language knowledge, can be considered as an Implicit Logic Channel. The pro-
022 posed Explicit Logic Channel, mimicking human logic reasoning, incorporates a
023 Large Language Model (LLM), a Visual Foundation Model (VFM), and a logical
024 reasoning module involving novel probabilistic inference for factual, counterfac-
025 tual, relational, and causal condition reasoning over the extracted and grounded
026 visual-textual facts. Cross-channel logic consistency analysis enables model vali-
027 dation and selection, even without ground-truth annotations. Additionally, cross-
028 channel integration further improves performance in zero-shot tasks over SOTA
029 models. Our experiments on three recent challenging VLC benchmarks, Neg-
030 Bench, HC-RefCOCOg, and HC-RefLoCo, demonstrate the effectiveness of the
031 proposed Logic Channel for logic-based model validation, selection and improve-
032 ment on LVLM with enhanced explainability and trustworthiness.
033

1 INTRODUCTION

035 Large Language Models (LLMs) have demonstrated remarkable general intelligence in text-based
036 tasks Zhao et al. (2025). Recently, LLMs have been extended as Multimodal LLMs (MLLMs) by
037 incorporating visual inputs. LVLMs have advanced significantly, as evidenced by the frequent re-
038 lease of frontier models by leading AI organizations and the rapid growth of vision-language (VL)
039 benchmarks in both data diversity and quantity Li et al. (2025a). While LVLMs have achieved ad-
040 vanced capabilities in visual language comprehension (VLC) across a wide range of multi-modal
041 tasks, recent research increasingly highlight the limitations of LVLMs in terms of reliability, fac-
042 tuality, explainability, and logic reasoning Dang et al. (2024). Frontier LVLMs are often deployed
043 for new tasks as black-boxes in zero-shot learning, due to data privacy concern, large model size, as
044 well as inaccessibility of closed-source model. It is crucial to be able to identify reliable model and
045 enhance the accuracy of predictions without re-training the model for new tasks Khan & Fu (2024).

046 Human make decision and judgment on facts and logical rules Wang et al. (2022). The faithfulness,
047 reasonability, and trustworthiness of a decision are justified based on explicit and concrete facts,
048 relations, and logic causation. On this observation, we propose a Logic Channel, in parallel with
049 LVLM, to perform a logic validation of LVLM’s prediction on logic consistency, as well as enhance
050 the accuracy of LVLM’s prediction. Since a LVLM has been pre-trained on vast vision-language
051 datasets, it is assumed to have learned human logic on visual-language problems. It can be con-
052 sidered as an Implicit Logic Channel (ILC). However, human natural language is not a complete
053 logic language Gomes (2024). The ground truth (gt) annotations based on natural language may be
054 ambiguous, uncertain, or logically incomplete, leading to logically inconsistent predictions. To com-
055 plement ILC, we propose an Explicit Logic Channel (ELC), where we prompt a LLM as Language

054 Parser to extract concept-level logic conditions from the input text, employ a VFM to ground the the
 055 facts explicitly in the image, and design a Logic Reasoning module with novel probabilistic infer-
 056 ence approach to make decisions explicitly on the grounded facts, relations, and causal conditions.
 057 The consistency between the two logic channels can be exploited for model validation, justification,
 058 and selection, even when gt annotations are not available for evaluation. In addition, the combination
 059 of both channels further enhance the accuracy of prediction without additional fine-tuning.

060 To investigate the effectiveness of the proposed Logic Channel for VLC validation and accuracy
 061 enhancement under zero-shot setting, we perform experiments on three recent VLC challenges, i.e.,
 062 NegBench, HC-RefCOCOg and HC-RefLoCo. Our experimental results show that, (a) the Logic
 063 Channel with probabilistic inference is able to produce accurate and robust predictions with explicit
 064 visual justification; (b) logic consistency between the ILC and ELC is effective for model validation
 065 and selection even without gt annotation; (c) the combination of ILC and ELC can further improve
 066 SOTA performance with enhanced explainability and trustworthiness.

067 The main contributions can be summarized as: (1) A Logic Channel on Foundation Models and
 068 Logic Reasoning for VLC tasks, producing concrete prediction on explicit facts and logic relations
 069 without training; (2) Logic consistency between ILC and ELC for model validation without the
 070 need of gt annotation; (3) Effective LC solutions with novel probabilistic inference approach for
 071 VLC with explicit logic reasoning on factual and counterfactual evidence, relations, and conditions
 072 of causation, achieving significant improvement over SOTA performance on three VLC benchmarks.

074 2 RELATED WORK

075 **LVLMs.** There is a growing number of LVLMs released in recent years Li et al. (2025b). One
 076 cluster are jointly embedding-based Vision Language Models (VLMs), where the visual and textual
 077 inputs are encoded separately, and then embedded into a shared latent space. Contrastive learning
 078 is then used to align the embedded image-text pairs by optimizing an objective function. Repre-
 079 sentative VLMs, particularly CLIP cluster models Radford et al. (2021), and those developed on
 080 BLIP Li et al. (2022) and ALIGN Jia et al. (2021), are pre-trained on large-scale VL datasets. An-
 081 other large group are Multimodal LLM (MLLM), which employs pre-trained LLM as backbone Yin
 082 et al. (2024). The visual features are extracted by a vision encoder, projected into a shared embed-
 083 ding space of text tokens, and inserted into text sequence as input to LLM to predict next tokens
 084 auto-regressively. Recent MLLMs are LLaVA Liu et al. (2023), GPT-4V Yang et al. (2023), Gemini
 085 Gemini Team (2025), InternVL Chen et al. (2024), and Qwen-VL Bai et al. (2023b). In this
 086 study, we employ EvaCLIP and InternVL2(8b) as representative VLM and MLLM, respectively.
 087 These models are less than 10B, and are more portable and practical for downstream applications.

088 **VLC benchmarks.** The benchmarks for evaluating the capabilities of latest LVLM on VLC are
 089 growing rapidly Li et al. (2025b); Fu et al. (2024). VLC capabilities are mostly evaluated us-
 090 ing Visual Question Answering (VQA) and Referring Expression Comprehension (REC) or Visual
 091 Grounding. Traditional VQA benchmarks typically involved multiple choice formats or answers
 092 with limited text length. Subsequent VL benchmarks have been extended to include a wider range
 093 of question types, field of expertise, and tasks such as math reasoning and chart understanding. The
 094 VQA datasets have also been extended to investigate bias in dataset distributions Li et al. (2024)
 095 and negations Alhamoud et al. (2025). Standard VQA datasets are considered less challenging to
 096 frontier models Li et al. (2025b) due to vast and extended pre-training datasets. We perform ex-
 097 periments on a latest benchmark with negations to evaluate the effectiveness on zero-shot setting.
 098 The REC task aims to identify and localize a specific image region of the referred object on text
 099 description Xiao et al. (2024). RefCOCO/+g are standard datasets for REC, built on top of MS
 100 COCO dataset. RefCOCO/+ might not be adequate for evaluating LVLM due to their limitations on
 101 overly concise referring phrases and limited vocabulary. On the other hand, zero-shot REC is still a
 102 very challenging task Han et al. (2024). In this study, we conduct experiments on HC-RefCOCOg
 103 with enriched phrase of 8.9 words in average, wrt 3.3 and 3.4 of RefCOCO/+, and a recent challenge
 104 HC-RefLoCo with long context expression of 93 words in average Wei et al. (2024).

105 **Logic reasoning on LLMs and VLM.** Efforts to enhance LLM with logical reasoning abilities have
 106 attracted researches' attention Cheng et al. (2025); Pournemat et al. (2025). The goal is to improve
 107 the logical accuracy and consistency of the generated text answers and expressions. The approaches
 include introducing an external logic solver Olausson et al. (2023); Diego et al. (2025) or building

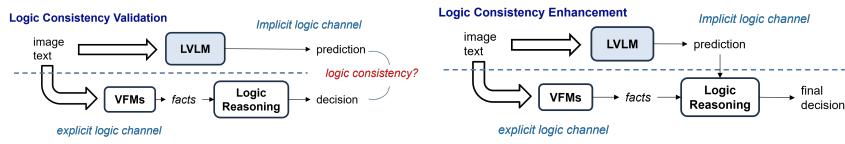


Figure 1: The Logic Channel for validation (left) and enhancement (right).

additional dataset with logically consistent annotations to fine-tune a model Feng et al. (2024). In BIRD Feng et al. (2025), probabilistic inference is used to improve the trustworthiness of LLM. Existing NeuroSymbolic framework and programming require additional learning to perform logic reasoning Li et al. (2023); Huang et al. (2025); Yang et al. (2024).

3 METHODOLOGY

Typical VLC tasks could be: (A) given a query image and a text expression to predict a right answer, or (B) given a query image and a text prompt to localize a target object in the image. They can be viewed as a logic problem of *if I (image) and T (text), then D (decision)*. We propose a Logic Channel in parallel with LVLM, as shown in Figure 1. The upper channel employs a LVLM to make prediction directly as a black box, which can be viewed as an *Implicit Logic Channel*. The lower channel, mimicking human logic reasoning, employs LLM and VFM to ground the facts explicitly, and performs logic reasoning on the grounded facts for final decision, which can be considered as an *Explicit Logic Channel*.

The two channels are complementary as LVLM, LLM and VFM are trained on different purposes. ELC provides explicit facts and relations to support the prediction. The logic consistency between the two channels can be used for model validation and selection, while the combination of the two predictions would further enhance the accuracy of final prediction.

Validation: When applying a LVLM for a new task, we may just have some test examples without gt annotation. We can run the model with Logic Channel to evaluate the logic consistency between the two channels. We define a Consistency Rate (*CR*) as

$$CR = \frac{\text{Number of Aligned Predictions}}{\text{Total Number of Test Samples}}, \quad (1)$$

where an aligned prediction means that the both channels give the same prediction. Naturally, if *CR* is higher, the LVLM is more logically reliable and trustworthy for the new task. The metric *CR* can be used to select a most suitable model for a new VLC task even the gt annotations are not ready for evaluation. The difference between the two channels may also guide user to perform a manual validation effectively on grounded visual evidence.

Enhancement: Fusing the outputs of the two complementary channels could also enhance the accuracy of the final prediction without additional re-training or fine-tuning, and the explicit visual evidence from ELC can provide a concrete logic justification of the prediction.

In this study, we investigate the effectiveness of Logic Channel with three general logic capabilities on VLC tasks, *i.e.*, (a) logical decision on factual and counter-factual evidence, (b) logical decision on evidence of facts and relations, and (c) logical decision on conditions of causation. We propose three LC solutions on these VLC requirements and evaluate their effectiveness on three recent VLC challenges on zero-shot setting.

3.1 LOGICAL DECISION ON FACTS AND NEGATIONS

A typical task of vision-language understanding is that, given an image (*I*), the model is required to predict a correct text description (*T*). A latest challenging benchmark, NegBench Alhamoud et al. (2025), involves both positive and negative phrases. Four choices of text descriptions are provided for each image. The choices are classified into three linguistic templates: Affirmation, Negation, and Hybrid. An affirmation text contains only positive elements $\{pos\}$, *i.e.*, objects or concepts present in the image. A negation text includes only negative elements $\{neg\}$, which are absent from the image but commonly associated with the present objects. A hybrid text contains both positive and

negative elements. The SOTA performance shows the difficulty of existing LVLMs to understand both positive and negative descriptions on a query image.

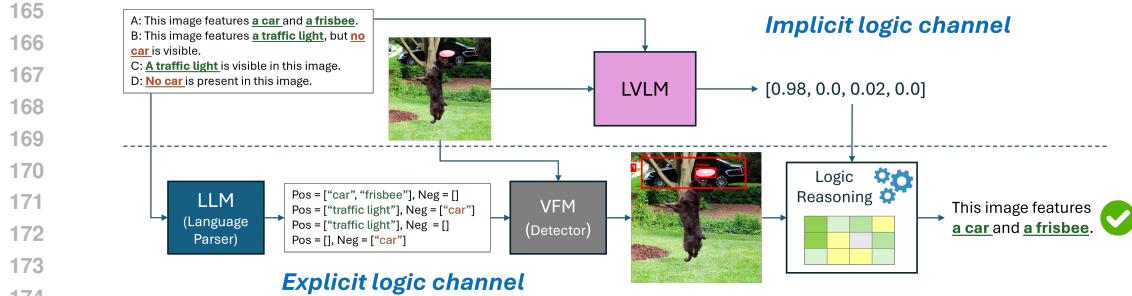


Figure 2: Logic Channel for NegBench.

Logically, the NegBench task requires decision to be made and justified on concrete factual and counter-factual evidence, i.e., the presence of pos elements (facts) and absence of neg elements (counter-facts). We implement a LC solution for the MCQ task of the NegBench benchmark, as illustrated in Figure 2. A representative LVLM is employed for the ILC. Following the protocol in Alhamoud et al. (2025), for each problem, the query image is presented to the model alongside four choice texts. The model is prompted to select the correct answer as well as provide the confidence-level of correctness (0%-100%) for all four choices. In the ELC, first, a SOTA LLM is employed as the Language Parser to extract the pos and neg nouns. The nouns are then fed to a VFM. For each object category, the VFM locates all instances in the image with the probabilities of their presence. The probability of an object category is obtained as the maximum of all the instances. Assuming there are K pos object categories and L neg object categories, the presence of the object categories in the image are denoted as $\{P(O_p^k)\}_{k=1}^K$ and $\{P(O_n^l)\}_{l=1}^L$. The presence of pos objects gives a factual evidence on T and the presence of neg objects gives a counter-factual evidence on T can be computed logically as

$$P(pos) = \min\{P(O_p^1), \dots, P(O_p^K)\}, \quad P(neg) = \max\{P(O_n^1), \dots, P(O_n^L)\}. \quad (2)$$

For a pair of image I and text T , the probability of factual and counter-factual evidence can be computed as

$$P(T|I) = \begin{cases} P(pos), & \text{affirmation} \\ 1 - P(neg), & \text{negation} \\ [P(pos)(1 - P(neg))]^{\frac{1}{2}}, & \text{hybrid} \end{cases} \quad (3)$$

where the geometric mean is used for normalization for the hybrid case. The choice of maximum $P(T|I)$, out of the four given choices, is selected as the correct answer.

When performing logic consistency validation, the two channels are run independently, and the CR is then computed. For enhancement, the combined probability is obtained as the sum of the two channels' outputs, and the best choice is selected as the final prediction.

3.2 LOGICAL DECISION ON FACTS AND RELATIONS

Recent HC-REC (Human-Centric Referring Expression Comprehension) benchmarks, sourced from general REC task, are created for the evaluation of large multi-modal models on VLC capabilities Wei et al. (2024). Given a query image (I) and a text expression (T), the model is required to predict and localize the referred person (H) within the image. The prediction should be made on the presence of the person and relevant objects, as well as the evidence of their relations in the image. The presence of a person is the core factual evidence, and his/her relations with the mentioned objects provide the evidence of association on logic relations for prediction.

The Logic Channel solution for HC-REC is shown in Figure 3. In the ILC, we first employ a VFM to locate every person in the image. Then, each person is cropped, and fed to a LVLM along with the referring expression. The LVLM makes a prediction on each person with a confidence value on its learned VLC knowledge. In the ELC, the Language Parser is prompted to extract the nouns of

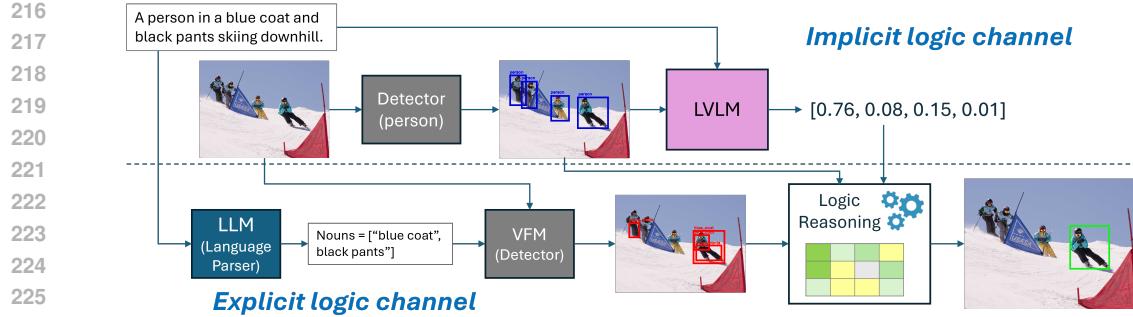


Figure 3: Logic Channel for HC-REC task.

persons and objects from T . The nouns are then fed to a VFM to locate all instances of persons and relevant objects with the probabilities of their presence in the image.

Assuming that there are $(K + 1)$ nouns extracted from T , of which the first is human and the rest are object names, denoted as $\{H, O_1, \dots, O_K\}$. The VFM locates a set of instances of human and objects. There may be multiple instances of one object category. In addition, due to errors of zero-shot detection, there may be missed detection and false positives. Let us denote the box of the i th detected person as h_i , and the box of the l th instance of object O_j as o_{jl} .

We compute the probability of the targeted person on accumulated evidence, as inspired by the principle of perceptual decision-making on evidence accumulation Balsdon et al. (2020). Formally, it can be expressed as

$$P(h_i|T) = \frac{1}{K+1} \left(\sum_{k=1}^K P(O_k|h_i) + P(h_i|H) \right), \quad (4)$$

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

$$R_A(o_{kl}, h_i) = A_{int}(o_{kl}, h_i) / A(o_{kl}), \quad (5)$$

where $A_{int}(o_{kl}, h_i)$ means the area of intersection of boxes o_{kl} and h_i , and $A(o_{kl})$ is the area of box o_{kl} . The probability of association of h_i with an object named O_k is obtained as

$$P(O_k|h_i) = \max_{l \in [1, L]} \{R_A(o_{kl}, h_i)\}. \quad (6)$$

Finally, the person of maximum $P(h_i|T)$ is selected as the grounded person referred by text expression on explicit visual evidence.

For model validation, the two channels are run independently. In ELC, $P(h_i|H) = 1.0$ is used in (4), which means every detected person is assumed as a potential target person. Then, the CR is computed. For enhancement, the confidence value of LVLM's prediction of h_i on T is used as $P(h_i|H)$ in (4), which integrates the power of LVLM in logic reasoning.

3.3 LOGICAL DECISION ON CONDITIONS OF CAUSATION

A recent challenge of REC is HC-RefLoCo, i.e., Human-Centric Referring Expression Comprehension with Long Context Wei et al. (2024). The very long text expression, which contains multiple sentences about the scene, global events and activities, to individual appearance and action, cause difficulty and even failure of many LVLMs for zero-shot task.

The challenges are not simply caused by long context expression, but also by sentences that are not logically complete for the cause-effect prediction. On logic cause-effect relation Gomes (2024), given image I , the text T should be sufficient and necessary to cause the localization of H in I . However, many sentences in T are neither sufficient nor necessary to lead to the visual attention to H in I . This can be explained in the following example. Suppose a long context expression on an image of human activities in a public park in a weekend is given as: “*This is a scene of a public park*

270 *in Sunday. People play various activities in the morning. A young lady in red sportswear is sitting*
 271 *on a bench for a rest. A blue bottle is placed on the bench beside her.*” The first sentence (S_1) is
 272 neither sufficient nor necessary to lead to the attention to the young lady (H). The second (S_2) is a
 273 sufficient but not necessary cause to H since it may lead to attention to every person in I , not just
 274 H . The third (S_3) is a sufficient and necessary cause to H , if there is no other young lady in red
 275 sportswear sitting on a bench in I . The fourth (S_4) is necessary but may not be sufficient cause as it
 276 does not focus on person.

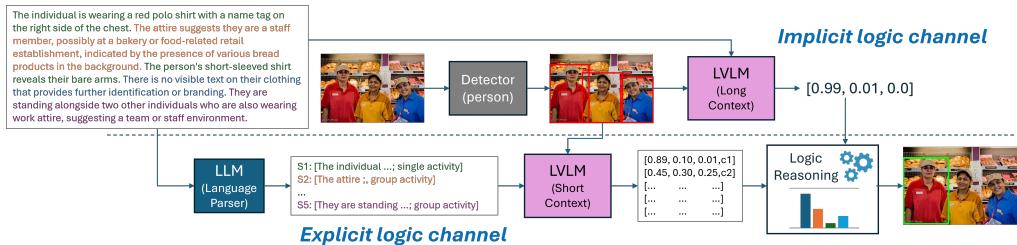


Figure 4: Logic Channel for HC-RefLoCo.

288 A LC solution for HC-RefLoCo is developed, as shown in Figure 4. On this challenge, we focus on
 289 the effectiveness of understanding the logic causation conditions on long context expressions with
 290 LVLMs. We employ LVLMs in both channels, one works on the long text of full expression, and
 291 another works on short text of sentence. The persons detected by VFM are used in both channels as
 292 potential visual grounds. In ILC, each person is cropped from the image, and fed to a LVLM along
 293 with the full expression. The LVLM produces predictions on its learned VLC knowledge, each with
 294 a confidence value of its prediction. In ELC, a logic reasoning approach on LLM and LVLM is
 295 developed. First, each sentence is extracted from the full expression and a LLM is prompted to
 296 classify each sentence into four categories: single activity (e.g., S_3), object (e.g., S_4), group activity
 297 (e.g., S_2) and scene (e.g., S_1). On logic formation, they could be assumed to correspond to different
 298 logic conditions of causation, i.e., (a) Single Activity: a sufficient and necessary cause; (b) Object:
 299 a necessary but not sufficient cause; (c) Group Activity: a sufficient but not necessary cause; and (d)
 300 Scene: a neither sufficient nor necessary cause.

301 Then, for a detected person, the cropped patch and each sentence form a VL pair. It is fed to a
 302 lightweight LVLM. Hence, for each person and each sentence, we obtain a probability of LVLM’s
 303 prediction on the VL association. If there are N persons detected and K sentences extracted from
 304 T , the predicted HS matches can be denoted as $\{\{P_{LVLM}(h_n|S_k)\}_{k=1}^K\}_{n=1}^N$. On the principles of
 305 probabilistic logic causation Gomes (2024), probabilistic inference framework on factors Feng et al.
 306 (2025), and human decision-making theory Bradley (2018), the logical prediction is obtained as

$$P(h_n|T) = \sum_{k=1}^K P_{LVLM}(h_n|S_k)u(S_k) \quad (7)$$

307 where $u(S_k)$ can be considered as a utility weight of S_k for perceptual decision-making, which can
 308 be selected on sentence category (see Sec. Experiments and Supplementary Material). The h_n of
 309 maximum $P(h_n|T)$ is selected as H .

310 Different LVLMs can be used for the Implicit and Explicit Logic Channels, on their performance
 311 on long/short text expressions. For model validation, the two channels are run independently, and the
 312 results are compared for logic validation and model selection. For enhancement, a simple weighted
 313 sum can further improve the accuracy of the final prediction.

4 EXPERIMENTS

314 In this section, we present the quantitative results and analysis on the effectiveness of Logic Channel
 315 for model validation, justification, selection, and enhancement of prediction when applying LVLM
 316 to challenging VLC tasks in zero-shot setting. In our experiments, we employ two LVLMs, i.e.,
 317 EvaCLIP and InternVL2(8b), as representative models of VLM and MLLM, respectively. The LLM
 318 used as Language Parser is Mistral Jiang et al. (2023), and the VFM for person and object detection
 319 is GroundingDINO Liu et al. (2024). In the following, ILC-LVLM denotes the results from ILC,

324 and ELC-LVLM for the results from ELC, where $\text{LVLM} \in \{\text{EvaCLIP}, \text{InternVL2(8b)}\}$, referring to
 325 the corresponding LVLM used in the channel.
 326

327 4.1 EVALUATIONS ON NEG BENCH BENCHMARK

329 NegBench Alhamoud et al. (2025) is one of most recent VLC challenges. There are three natural
 330 image MCQ tasks, i.e., COCO, VOC2007, and HardNeg-Syn. Our experiments are conducted
 331 on the publicly available data, i.e., 5914 MCQ questions on COCO and 5032 MCQ questions on
 332 VOC2007. The SOTA accuracies are extracted from Table 2 and Figure 7 in the Appendix of the
 333 paper Alhamoud et al. (2025).

334 **Validation.** First, we evaluate the CR without gt annotation. The results of the ILC on EvaCLIP
 335 and InternVL2(8b) are presented in Table 1. The CR of EvaCLIP is around 27%, slightly over 25%
 336 of random chance. Even without gt annotation, one can predict that EvaCLIP is poor for the task,
 337 according to the logic validation on the factual and counter-factual evidence. On the other hand,
 338 the CR of InternVL2(8b) is around 70% across the two datasets, much higher than EvaCLIP, which
 339 provides a logic validation that InternVL2(8b) may be more suitable for the task.

Benchmark	Dataset	EvaCLIP	InternVL2(8b)
NegBench	COCO	27.05	62.80
	VOC	27.25	79.25
	ALL	27.14	70.36

345 Table 1: Logic Consistency Rates on NegBench.

Model	COCO	VOC2007
CLIP-Laion400M (2022)	24.26%	27.01%
CLIP-L14(2021)	22.44%	23.69%
CLIP-H14(2021)	32.14%	38.26%
LLaVA(7b)(2023)	48.00%	56.00%
LLaVA(13b)(2023)	54.00%	63.00%
ILC-EvaCLIP	24.82%	26.60%
ILC-InternVL2(8b)	63.26%	85.41%
ELC	75.77%	86.84%
ILC-InternVL2+ELC	70.34%	88.25%

356 Table 2: MCQ Total Accuracy (%) on NegBench.

357 When the gt annotations are used for performance evaluation, one can observe the results of individual
 358 ILC or ELC in Table 2. The accuracy of EvaCLIP is around 25%, agreeing with CR validation
 359 and the results of CLIP cluster models reported in Alhamoud et al. (2025). As indicated by CR
 360 validation, InternVL2(8b) produces much better performance than EvaCLIP, and achieves a large
 361 jump over the SOTA by LLaVA(13b), one of the MLLMs.
 362

363 **Enhancement.** When we simply sum up the predictions from both ILC and ELC, we obtain per-
 364 formance better than the end-to-end model on COCO, but lower than the best single channel due to
 365 a large gap between the two channels. Nonetheless, we obtain further improvement on VOC2007,
 366 increasing SOTA accuracy from 63% to over 88%.

367 4.2 EVALUATIONS ON HC-REFCOCOG BENCHMARK

369 In this experiment, we focus on HC-RefCOCOG with enriched phrase. As there is still a lack of
 370 benchmarks on HC-RefCOCOG, we download publicly available SOTA models on RefCOCOG and
 371 run on HC-RefCOCOG. In Table 3, to compare with SOTA of zero-shot REC, we list the published
 372 results on RefCOCOG, as well as the corresponding results on HC-RefCOCOG obtained by us.

373 **Validation.** First, we evaluate the logic consistency on the val set without gt annotation. The CR on
 374 EvaCLIP and InternVL2(8b) are 63.1% and 40.8% respectively. The former is around 20% higher
 375 than the latter, which may indicate that EvaCLIP may be more suitable for the task than InternVL2.

377 When the gt annotations are used for performance evaluation, one can observe the results of individual
 378 ILC with EvaCLIP and InternVL2(8b) in Table 3. As indicated by CR , the performance

Model	RefCOCOG		
	Val	Test	Val
ReCLIP(2022)	59.33	59.01	
RelVLA(2024)	57.60	56.64	
GroundVLP(2024)	64.30	63.54	
Model	HC-RefCOCOG		
	Val	Test	Val
ReCLIP(2022)	68.36	67.33	
RelVLA(2024)	65.65	64.57	
GroundVLP(2024)	74.44	74.81	
ILC-EvaCLIP	78.40	78.66	
ILC-InternVL2(8b)	47.80	51.66	
ELC	67.68	67.52	
ILC-EvaCLIP+ELC	80.70	80.05	

356 Table 3: Performance evaluation on HC-RefCOCOG.

378 of EvaCLIP is more than 20% over InternVL2, and achieves an improvement of around 4% over
 379 the SOTA. The ELC obtains performance of over 67%, close to the performance of RelVLA and
 380 ReCLIP, indicating that it can be used as an effective and reliable tool for model validation with
 381 explicit visual evidence.

382 **Enhancement.** By introducing the predictions of ILC into Eq. (4), the integrated LCR model (last
 383 row) produces new SOTA performance of over 80%, around 6~15% over previous SOTA and 1.8%
 384 over ILC with EvaCLIP, demonstrating the effectiveness of Logic Channel for enhancing the accu-
 385 racy of the final prediction.

387 4.3 EVALUATIONS ON HC-REFLOCO BENCHMARK

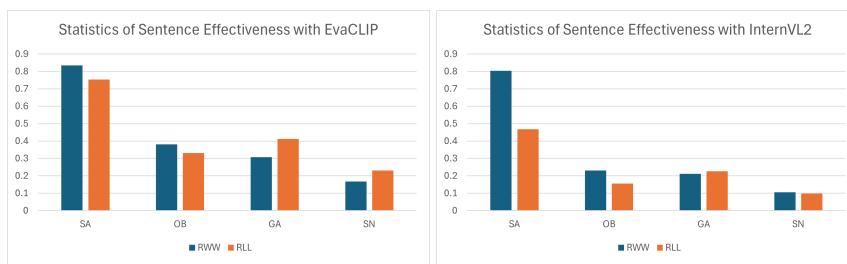
389 The recent challenge HC-RefLoCo Wei et al. (2024) is built on various image datasets, e.g., COCO,
 390 Objects365, OpenImage v7, and LAION-5B. On a query image, the long context annotation is
 391 generated by GPT-4V with user prompts and reviewed manually. The long text expressions form a
 392 great challenge to most LVLMs not pre-trained on long texts. Here, we show that it is possible to
 393 develop a LC solution with lightweight LVLMs and FMs (Figure 4) for the HC-RefLoCo challenge.

Model	Val+Test				Val	Test
	Acc _{0.5}	Acc _{0.75}	Acc _{0.9}	mAcc	mAcc	mAcc
Qwen-VL Bai et al. (2023a)	67.9	56.8	34.8	52.8	53.1	52.6
CogVLM Wang et al. (2024)	66.0	59.6	43.8	55.8	56.3	55.5
ONE PEACE Wang et al. (2023)	79.3	69.0	43.8	63.1	63.4	62.9
SPHINX-MoE-1k Lin et al. (2023)	85.8	77.3	53.7	71.4	71.5	71.4
SPHINX-v2-1k Lin et al. (2023)	84.1	77.1	56.2	71.7	71.6	71.7
ILC-EvaCLIP	75.3	70.5	59.1	67.6	67.2	67.8
ILC-InternVL2(8b)	33.5	31.4	24.9	29.6	29.1	29.8
ELC-EvaCLIP	67.7	63.4	52.5	60.6	60.2	60.8
ELC-InternVL2(8b)	80.8	75.9	63.4	72.7	71.8	73.1
ILC-EvaCLIP+ELC-InternVL2	83.4	78.3	65.2	75.0	74.5	75.2

406 Table 4: Performance evaluation on HC-RefLoCo benchmark.

408 **Validation.** First, we evaluate if a LVLm is effective on long context expressions on the logic
 409 consistency without gt annotation. We use EvaCLIP or InternVL2(8b) in both the ILC and ELC
 410 respectively. In the ELC, we simply sum up the probabilities on sentences. The *CR* between the
 411 two channels on EvaCLIP and InternVL(8b) are 72.8% and 15.8% respectively. The results indicate
 412 that, EvaCLIP might be less sensitive to the truncated long context expressions, while InternVL2(8b)
 413 might be too sensitive to the truncation of long context expressions.

414 The ILC with EvaCLIP and InternVL2(8b) are evaluated directly on the full expressions. The results
 415 are listed in Table 4 (i.e., ILC-EvaCLIP and ILC-InternVL2(8b)). Both LVLMs may truncate the
 416 long text input. However, in EvaCLIP, the truncated text input is embedded as a text vector of a
 417 fixed length and compared with the visual feature vector embedded as the same length. The result
 418 shows that EvaCLIP still lags behind the SOTA (SPHINX) but it achieves better result of Acc_{0.9}
 419 than SOTA, with the localization by GroundingDINO. On the other hand, InternVL2(8b) interprets
 420 the input text tokens one-by-one for next prediction, the truncation of the long text input causes poor
 421 performance on the HC-RefLoCo task.



431 Figure 5: The statistics of R_{WW} and R_{LL} on sentence categories on HC-RefLoCo val set.

In the ELC, a full expression is first separated as sentences, and the sentences are classified into four categories, labeled as SA, OB, GA and SN, denoting single activity, object, group activity, and scene respectively. Each sentence represents a degree of necessity or sufficiency, or a probability of the causation. Suppose there are K sentences and N persons in the image. Each sentence, together with all patches of cropped persons, are fed to a LVLM, and it is prompted to find the best match, with confidence values for its predictions. All together, we obtain LVLM predictions $\{\{P_{LVLM}^{n,k}\}_{k=1}^K\}_{n=1}^N$. If each sentence provides appropriate information and the LVLM is able to understand the VL information correctly, when we sum up the probabilities over all sentences for each person, the gt person would have the maximum probability, or labeled as ‘GT-Win’. For each person, we can also sum up the probabilities of each category, e.g., P_{SA}^n . If P_{SA}^n of the gt person is also the maximum of SA cluster (denoted as ‘SA-Win’), SA contributes to GT-Win. We compute a Rate of Win-Win for SA and a Rate of Lose-Lose for SA as

$$R_{WW}(SA) = \frac{Num(SA\text{-}Win\>\text{-}Win)}{Num(GT\text{-}Win)}, \quad R_{LL}(SA) = \frac{Num(SA\text{-}Lose\>\text{-}Lose)}{Num(GT\text{-}Lose)}. \quad (8)$$

In the same way, we can obtain the Rates for OB, GA, and SN categories. The statistics of the Rates with EvaCLIP and InternVL2(8b) on val set are shown in Figure 5. Logically, if SA-Win and GT-Win occur, we may view SA-Win as a sufficient condition of GT-Win, i.e., *if SA-Win, then GT-Win*. On the other hand, if SA-Lose and GT-Lose occur, we may view SA-Win is a necessary condition for GT-Win, i.e., *if not SA-Win, then not GT-Win*, or *if SA-Lose, then GT-Lose*. When examining the bars in Figure 5, one can observe that SA has strong relation with gt, while OB has weaker relation with gt, but still has more positive effect (R_{WW}) than negative effect (R_{LL}), and the remaining two categories have less relation with gt. On such statistics, we obtain the utility weights of SA, OB, GA, and SN (see Supplementary Material), and apply them in Eq. (7) for logic prediction.

The results of Logic Channel on sentences using EvaCLIP and InternVL2(8b) in ELC are shown in Table 4 (i.e., ELC-EvaCLIP and ELC-InternVL(8b)). The performance of ELC-EvaCLIP is poorer than ILC-EvaCLIP on full annotation, maybe due to missing information in separated sentences, but it is still comparable with most SOTA models except SPHINX. ELC-InternVL2(8b) produces much better performance on sentences, generates new SOTA results except $Acc_{0.5}$, especially, increases $Acc_{0.9}$ from 56.2% to 63.4%, significantly improving the accuracy of localization.

Enhancement. The model validation results show that EvaCLIP is effective on full annotation, and InternVL2(8b) is effective on sentences. By combining the two channels with a weighted sum, we obtain new SOTA performance except $Acc_{0.5}$, improving the mAcc from 71.7% to 75.0%, $Acc_{0.75}$ from 77.3% to 78.3%, and $Acc_{0.9}$ from 56.2% to 65.2%.

4.4 VISUAL VALIDATION AND JUSTIFICATION

One critical limitation of LVLM is the lack of explicit explanation of its decision making. The uncertainty may be caused by visual hallucination or ambiguity of gt annotation. With Logic Channel validation, we may effectively address such concerns by manually examining the inconsistent results between LVLM and ELC even without gt annotation. The inconsistencies may be caused by: (a) Hallucination: LVLM makes a prediction on nonexistent object or fails to see the referred object; (b) Ambiguity: LVLM makes a decision on neither sufficient nor necessary condition, or gt annotation is incorrect; or (c) Weakness of FMs: LLM fails to extract correct concept, or VFM fails to detect the relevant objects. The visual examples of such errors observed on explicit visual evidence in logic reasoning are presented in the Supplementary Material with further discussions.

5 CONCLUSION

Facing the uncertainty on reliability when deploying frontier LVLM to new tasks, we proposed a Logical Channel for model validation and enhancement w/o re-training or fine-tuning. Based on Foundation Models and logic reasoning, the Logic Channel make decision on explicit facts and relations, which provides a way for model validation on human logic. We investigated the effectiveness on three recent benchmark challenges in VLC. This Logic Channel enables diagnosis for model validation, selection and improvement, as well as enhancement of explainability and trustworthiness. The diversity of the three challenges demonstrates the generalization of our approach. In the future, it is worth to investigate the LC for validation of more complex multimodal CoT reasoning tasks.

486 REFERENCES
487

488 Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and
489 Marzyeh Ghassemi. Vision-Language Models Do Not Understand Negation. In *CVPR-25*, 2025.

490 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang
491 Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, local-
492 ization, text reading, and beyond, 2023a.

493 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang
494 Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding,
495 Localization, Text Reading, and Beyond, 2023b. URL <https://arxiv.org/abs/2308.12966>.

496 T. Balsdon, V. Wyart, and P. Mamassian. Confidence controls perceptual evidence accumulation.
497 *Nat Commun*, 11(1753), 2020. doi: <https://doi.org/10.1038/s41467-020-15561-w>.

500 Richard Bradley. Decision Theory: A Formal Philosophical Introduction. In Sven Ove Hansson and
501 Vincent F. Hendricks (eds.), *Introduction to Formal Philosophy*, pp. 611–655. Springer Interna-
502 tional Publishing, 2018.

503 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong
504 Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL:
505 Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Pro-
506 ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.
507 24185–24198, June 2024.

508 Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin.
509 Empowering LLMs with Logical Reasoning: A Comprehensive Survey, 2025. URL <https://arxiv.org/abs/2502.15652>.

510 Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie
511 Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. Explainable
512 and Interpretable Multimodal Large Language Models: A Comprehensive Survey, 2024. URL
513 <https://arxiv.org/abs/2412.02104>.

514 Calanzone Diego, Teso Stefano, and Vergari Antonio. Logically Consistent Language Models via
515 Neuro-Symbolic Integration. In *The Thirteenth International Conference on Learning Repre-
516 sentations*, 2025. URL <https://openreview.net/forum?id=7PGluppo4k>.

517 Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, Dongyan Zhao, and
518 Weizhu Chen. Language Models can be Deductive Solvers. In Kevin Duh, Helena Gomez,
519 and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL*
520 2024, pp. 4026–4042, Mexico City, Mexico, June 2024. Association for Computational Lin-
521 guistics. doi: 10.18653/v1/2024.findings-naacl.254. URL <https://aclanthology.org/2024.findings-naacl.254>.

522 Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. BIRD: A Trustworthy Bayesian Inference Frame-
523 work for Large Language Models. In *The Thirteenth International Conference on Learning Repre-
524 sentations*, 2025. URL <https://openreview.net/forum?id=fAAaT826Vv>.

525 Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing
526 Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. MME-Survey: A Comprehensive
527 Survey on Evaluation of Multimodal LLMs, 2024. URL <https://arxiv.org/abs/2411.15296>.

528 Gemini Team. Gemini: A Family of Highly Capable Multimodal Models, 2025. URL <https://arxiv.org/abs/2312.11805>.

529 Gilberto Gomes. Necessary and Sufficient Conditions, Counterfactuals and Causal Explanations.
530 *Erkenntnis*, 89:3085–3108, 2024. doi: <https://doi.org/10.1007/s10670-023-00668-5>.

540 Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaiyu Jiang. 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.

541

542

543 Jiani Huang, Ziyang Li, David Jacobs, M. Naik, and Ser Nam Lim. Laser: A neuro-symbolic framework for learning spatio-temporal scene graphs with weak supervision. In *International Conference on Learning Representations*, 2025. URL <https://api.semanticscholar.org/CorpusID:258179880>.

544

545

546

547 Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, 2021. URL <https://arxiv.org/abs/2102.05918>.

548

549

550

551 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL <https://arxiv.org/abs/2310.06825>.

552

553

554

555

556 Zaid Khan and Yun Fu. 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.

557

558

559

560

561 Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems*, volume 37, pp. 17044–17068. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1e69ff56d0ebff0752ff29caaddc25dd-Paper-Datasets_and_Benchmarks_Track.pdf.

562

563

564

565

566

567

568 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL <https://proceedings.mlr.press/v162/li22n.html>.

569

570

571

572

573

574 Ziyang Li, Jiani Huang, and Mayur Naik. Scallop: A Language for Neurosymbolic Programming. *Proc. ACM Program. Lang.*, 7:166:1–166:25, 2023.

575

576

577 Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 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 2025a.

578

579

580

581 Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 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 2025b.

582

583

584

585 Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

586

587

588 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, 2023. URL <https://arxiv.org/abs/2304.08485>.

589

590

591 Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *ECCV 2024*, pp. 38–55. Springer Nature Switzerland, 2024.

594 Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenen-
 595 baum, and Roger Levy. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining
 596 Language Models with First-Order Logic Provers. In Houda Bouamor, Juan Pino, and Kalika
 597 Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language
 598 Processing*, pp. 5153–5176, Singapore, December 2023. Association for Computational Linguis-
 599 tics. doi: 10.18653/v1/2023.emnlp-main.313. URL <https://aclanthology.org/2023.emnlp-main.313/>.

600

601 Mobina Pournemat, Keivan Rezaei, Gaurang Sriramanan, Arman Zarei, Jiaxiang Fu, Yang Wang,
 602 Hamid Eghbalzadeh, and Soheil Feizi. Reasoning under uncertainty: Exploring probabilistic
 603 reasoning capabilities of llms, 2025. URL <https://arxiv.org/abs/2509.10739>.

604

605 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
 606 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya
 607 Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Ma-
 608 rina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine
 609 Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR,
 610 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.

611

612 Pei Wang, Christian Hahm, and Patrick Hammer. A Model of Unified Perception and Cog-
 613 nition. *Frontiers in Artificial Intelligence*, Volume 5 - 2022, 2022. ISSN 2624-8212.
 614 doi: 10.3389/frai.2022.806403. URL <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.806403>.

615

616 Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang,
 617 and Chang Zhou. One-peace: Exploring one general representation model toward unlimited
 618 modalities. *arXiv preprint arXiv:2305.11172*, 2023.

619

620 Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,
 621 Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances
 622 in Neural Information Processing Systems*, 37:121475–121499, 2024.

623

624 Fangyun Wei, Jinjing Zhao, Kun Yan, Hongyang Zhang, and Chang Xu. A Large-Scale Human-
 625 Centric Benchmark for Referring Expression Comprehension in the LMM Era. In A. Globerson,
 626 L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural
 627 Information Processing Systems*, volume 37, pp. 69566–69587. Curran Associates, Inc., 2024.

628

629 Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards Visual
 630 Grounding: A Survey, 2024. URL <https://arxiv.org/abs/2412.20206>.

631

632 Xiaofeng Yang, Fayao Liu, and Guosheng Lin. Neural Logic Vision Language Explainer. *IEEE
 633 Transactions on Multimedia*, 26:3331–3340, 2024.

634

635 Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan
 636 Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), 2023. URL <https://arxiv.org/abs/2309.17421>.

637

638 Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on
 639 multimodal large language models. *National Science Review*, 11(12), 2024. ISSN 2053-714X.
 640 doi: 10.1093/nsr/nwae403. URL <http://dx.doi.org/10.1093/nsr/nwae403>.

641

642 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min,
 643 Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen,
 644 Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-
 645 Rong Wen. A Survey of Large Language Models, 2025. URL <https://arxiv.org/abs/2303.18223>.

646

647