Instruction-Guided Visual Masking

Jinliang Zheng [*1,2] **Jianxiong Li** [*1] **Sijie Cheng** [1] **Yinan Zheng** [1] **Jihao Liu** [2,3] **Yu Liu** [2] **Jingjing Liu** [1] **Xianyuan Zhan** [1,4]

Abstract

Instruction following is crucial in contemporary LLM. However, when extended to multimodal setting, it often suffers from misalignment between specific textual instruction and targeted local region of an image. To achieve more accurate and nuanced multimodal instruction following, we introduce *Instruction-guided Visual Masking* (IVM), a new versatile visual grounding model that is compatible with diverse multimodal models, such as LMM and robot model. By constructing visual masks for instruction-irrelevant regions, IVM-enhanced multimodal models can effectively focus on task-relevant image regions to better align with complex instructions. Specifically, we design a visual masking data generation pipeline and create an IVM-Mix-1M dataset with 1 million image-instruction pairs. We further introduce a new learning technique, *Discriminator Weighted Supervised Learning* (DWSL) for preferential IVM training that prioritizes high-quality data samples. Experimental results on generic multimodal tasks such as VQA and embodied robotic control demonstrate the versatility of IVM, which as a plug-and-play tool, significantly boosts the performance of diverse multimodal models, yielding new state-of-the-art results across challenging multimodal benchmarks. Code is available at https://github.com/2toinf/IVM.

1. Introduction

Multimodal instruction following is a fundamental multimodal task, powering a wide-range of applications such as visual question answering (VQA) (Goyal et al., 2017), visual captioning (Achiam et al., 2023; Liu et al., 2024c), and embodied robotic control (Driess et al., 2023). To effectively solve this task, one critical capability required is nuanced image-language grounding, which current multimodal models grow implicitly and slowly through data-intensive end-to-end training without explicit grounding supervisions. Two challenges emerge in this indirect learning of image-instruction alignment: 1) How to accurately localize targeted image regions that corresponds to a specific textual instruction, as illustrated in Figure 1. 2) How to generalize to diverse visual representations (*e.g.*, same object with different colors, compositions, or backgrounds) that reflect similar textual instruction (e.g., Q3 in Figure 1). Lacking an effective and direct solution to these challenges, the most advanced Large Multimodal Models (LMMs) (Achiam et al., 2023; Bai et al., 2023; Liu et al., 2024c; Driess et al., 2023) still suffer from hallucinations even when trained with high-quality data in the magnitude of billions (Li et al., 2023c).

We introduce *Instruction-guided Visual Masking* (IVM), a versatile plug-and-play model designed to enhance multimodal instruction following via nuanced surgical visual grounding. To eliminate the distraction of instruction-irrelevant visual regions, IVM automatically masks out these regions to sharpen the focus of instruction following, and meticulously crops visual input to tailor for a specific instruction and enforce multimodal models to zoom in on task-related visual content. Existing visual grounding methods are limited either to predefined object categories, which cannot cover diverse instruction-related visual content; or they subscribe to a fixed instruction format, which restricts the expressiveness of instructions. As shown in Figure 2, such simplistic grounding techniques often fail to comprehend complex instruction-following tasks. Learning an IVM model requires pixel-level, fine-grained, instruction-guided mask annotations that provide explicit grounding supervisions. To create such a dataset, we build a LLM-empowered Mixture of Expert pipeline with SOTA visual grounding models (Sun et al., 2023; Shao et al., 2024; Lai et al., 2024; Gupta et al., 2022) to efficiently create abundant reliable labels. To compensate the noises in auto-generated labels, we further manually label a smaller dataset with clean annotations, and integrate the two into an IVM-Mix-1M dataset that contains 1 million image-instruction pairs. To reduce demand on costly human labels and ensure optimized utility

[1] AIR, Tsinghua University [2]Sensetime Research [3]MMLab, CUHK [4]Shanghai AI Lab. Correspondence to: Jinliang Zheng <zhengjl23@mails.tsinghua.edu.cn>, Jianxiong Li <li-jx21@mails.tsinghua.edu.cn>, Xianyuan Zhan <zhanxianyuan@air.tsinghua.edu.cn>.

Multi-modal Foundation Model meets Embodied AI Workshop @ ICML2024, Vienna, Austria.

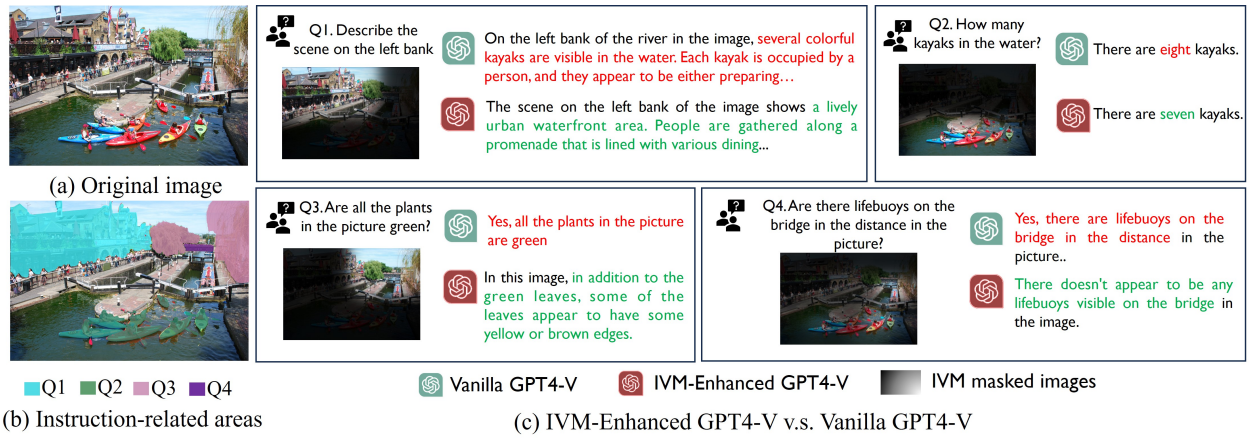


Figure 1. The most advanced LMMs (e.g. GPT4-V) still fail on complex instruction following tasks. With IVM assistance to simplify visual inputs, existing LMMs can gain significant improvement.

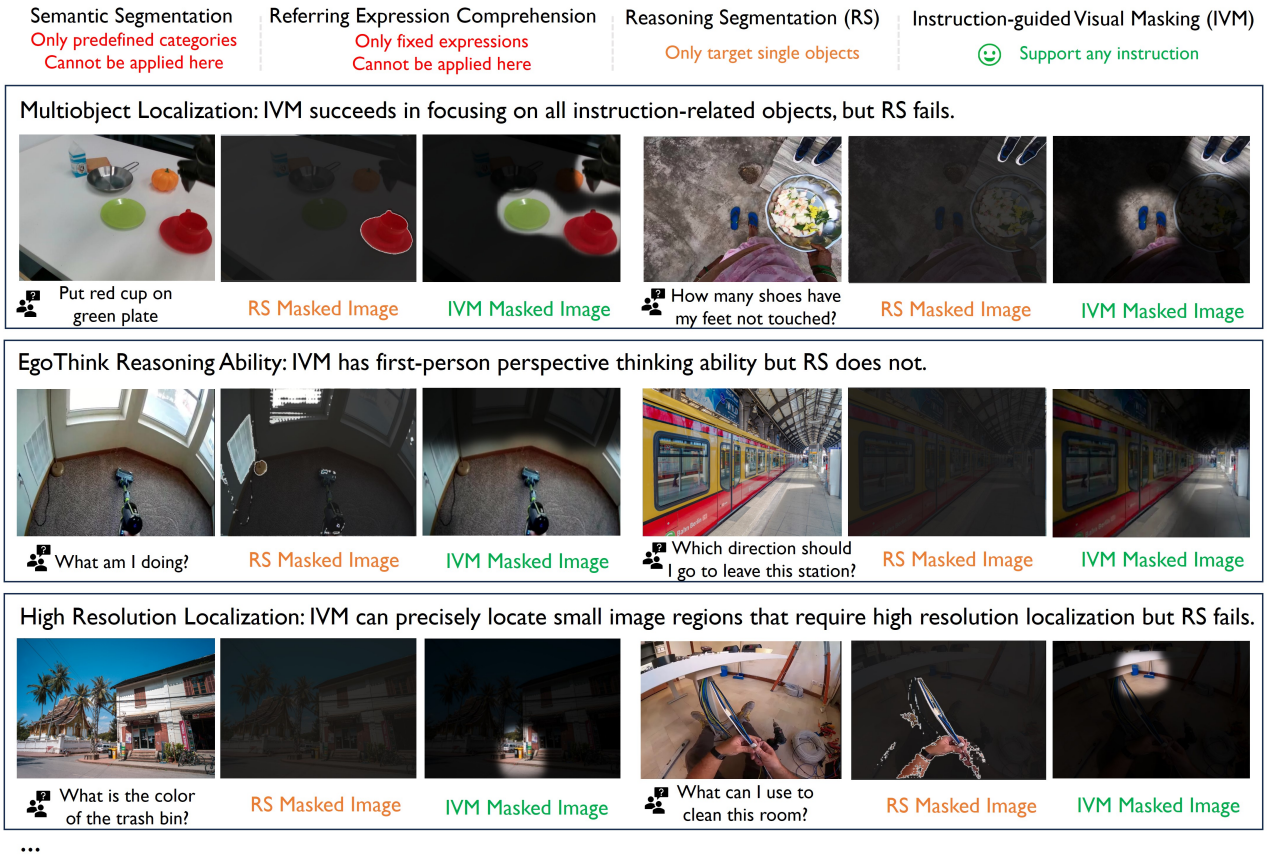


Figure 2. Comparison between IVM and Reasoning Segmentation (RS) (Lai et al., 2024). Traditional methods such as semantic segmentation (Zhou et al., 2017) and referring expression comprehension (Yu et al., 2016) are limited to fixed categories or fixed instruction formats, thus inapplicable to complex instruction following tasks. RS has reasoning ability, but only allows single object localization. IVM, instead, is universally applicable to any instruction.

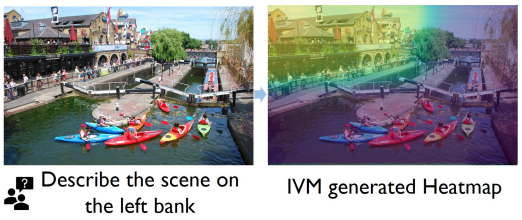


Figure 3. Instruction-guided Visual Masking.

of machine-generated labels, we employ a Discriminator-Weighted Supervised Learning (DWSL) framework for IVM training, inspired by recent advances in offline imitation learning (Xu et al., 2022). Specifically, we introduce a discriminator to assign weights to masks, where high values are assigned to high-quality annotations and vice versa. Thus, these weights generated by the discriminator can naturally act as a weighting function for the IVM training objective, allowing for a preferential training process that prioritizes learning from reliable samples and discards misleading ones. Extensive experiments demonstrate great versatility of the IVM model when integrated into existing multimodal chatbots (commercial and open-sourced) without fine-tuning. Our IVM-enhanced LMMs gain significant performance improvement across new challenging benchmarks such as V*Bench (Wu & Xie, 2023), EgoThink (Cheng et al., 2024) and POPE (Li et al., 2023c), achieving new state of the art. IVM model also proves valuable in vision-language robotic manipulation tasks, where data collection is notoriously challenging and generalization is a major concern (Li et al., 2024). With the integration of IVM, our enhanced robot model exhibits boosted performance and better generalization capabilities.

Our contributions are summarized as follows: 1) We propose Instruction-guided Visual Masking (IVM), a novel approach that serves as a versatile plug-and-play module to enhance multimodal models through visual grounding. 2) We introduce the IVM-Mix-1M dataset and propose an LLM-empowered Mixture of Expert pipeline to create visual grounding labels. 3) We present the DWSL algorithm for IVM training that automatically prioritizes high-quality training samples.

2. Instruction-Guided Visual Masking

To help multimodal models focus on instruction-sensitive image regions without distractions from irrelevant visual elements, we introduce Instruction-guided Visual Masking (IVM), a versatile plug-and-play model that enhances multimodal instruction following via surgical targeted visual grounding.

IVM aims to produce a heatmap $\mathbf{H}$, given an image $\mathbf{x}_{\text{img}}$ and a textual instruction $\mathbf{x}_{\text{txt}}$. The heatmap $\mathbf{H}$ identifies the critical image region to follow the instructions, as illustrated in Figure 3, allowing multimodal models to easily zoom in on targeted image regions while ignoring neighboring areas.

This formulation evokes the problem definition of Reasoning Segmentation (RS) (Lai et al., 2024). There are two main differences: 1) IVM addresses a more challenging problem. RS tries to target single objects from simple instructions, *e.g.*, "what is.., where is.., who is...", while IVM aims to include all instruction-related visual regions within the image given any instruction, which demands advanced and nuanced image-language grounding ability (as illustrated in Figure 2). 2) RS has clear ground truths but IVM does not. The instructions in RS primarily correspond to simple and semantic-meaningful objects that are straightforward for human annotations. IVM, however, deals with broader and more ambiguous instruction-related regions (*e.g.*, the left bank regions in Figure 3), making the training and annotating much more challenging.

For more details of data preparation and model training, please refer to Appendix D E

3. Experiments

In this work, we employ *LLaVA-7B* (Liu et al., 2024b) as the LMM and *SAM-H* (Kirillov et al., 2023) as the vision backbone for our IVM model (Figure 8), which is trained on the IVM-Mix-1M dataset using the proposed DWSL algorithm. More details on the architecture and training can be found in Appendix E. We conduct extensive experiments to assess the effectiveness of the IVM model. Specifically, we utilize the heatmap generated by the IVM for image post-processing. These processed images can then be seamlessly fed into downstream multimodal models for diverse tasks. Unless otherwise specified, we use the image post-processing method of overlaying and cropping to discard instruction-irrelevant image content. A detailed discussion on post-processing methods is presented in Section F.1. We also provide more evaluation results and analysis in Appendix F.

3.1. Main Results

Integration with Commercial Chatbot. We use GPT4-V (Achiam et al., 2023) as the base model. Considering the superior perception and reasoning capability of GPT4-V, we evaluate IVM-enhanced GPT4-V on V*bench (Wu & Xie, 2023), a recently proposed challenging VQA-type benchmark characterized by images with abundant redundancies. Results are presented in Table 1. The accuracy of the vanilla GPT4-V is mediocre (55.0%). Our IVM model, however, can significantly improve the performance (+26.2%) and establish a new state of the art on this benchmark, even surpassing the task-specialized SEAL (Wu & Xie, 2023)

Table 1. V* bench results.

LMMs	Attribute(%)	Spatial(%)	Overall(%)
Open-Sourced LMMs			
BLIP2 (Li et al., 2023b)	27.0	53.9	37.7
MiniGPT-4 (Zhu et al., 2023)	30.4	50.0	38.2
InstructBLIP (Dai et al., 2023)	25.2	47.4	34.0
Otter (Li et al., 2023a)	27.0	56.6	38.7
LLaVA-1.5 (Liu et al., 2023a)	43.5	56.6	48.7
Commercial Chatbots			
Bard (Manyika & Hsiao, 2023)	31.3	46.1	37.2
Gemini-Pro (Team et al., 2023)	40.9	59.2	48.2
GPT4-V (Achiam et al., 2023)	51.3	60.5	55.0
Specific Visual Search Models			
SEAL (Wu & Xie, 2023)	74.8 (+23.5)	76.3 (+15.8)	75.4 (+20.4)
IVM-Enhanced GPT4-V	**87.0 (+35.7)**	**72.4 (+11.9)**	**81.2 (+26.2)**

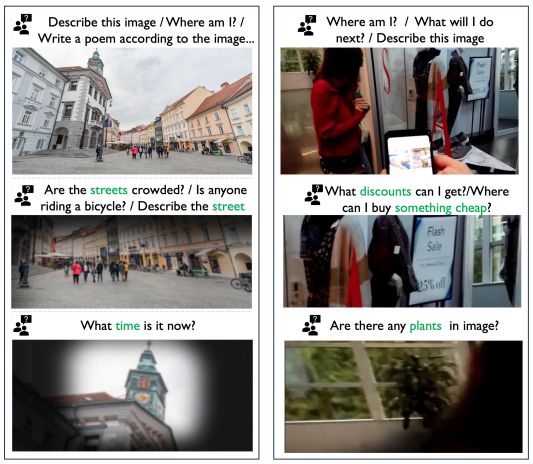


Figure 4. IVM can handle various instructions, ranging from retaining entire images for captioning (row 1) to localizing unique objects (row 2 and 3).

Table 2. Results on other multimodal benchmarks. MME* denotes the aggregate of scores from -p and -c.

LMMs	#Param	EgoThink	POPE	MME*	GQA	SQA	VQAv2
InstructBLIP (Dai et al., 2023)	13B	-	78.9	1212.8	49.5	60.5	-
Qwen-7B (Bai et al., 2023)	7B	-	-	-	58.3	67.1	78.8
SEAL-7B (Wu & Xie, 2023)	7B	-	82.4	1129	-	-	-
LLaVA-7B (Liu et al., 2023a)	7B	51.1	85.9	1748	62.0	70.2	78.5
LLaVA-13B (Liu et al., 2023a)	13B	55.2	85.9	1834	67.1	71.6	80.0
LISA (Lai et al., 2024)-Enhanced LLaVA-7B	20B	47.9 (-3.2)	80.0 (-5.9)	1560 (-188)	56.6 (-5.4)	69.3 (-0.9)	78.2 (-0.3)
IVM-Enhanced LLaVA-7B	14B	**54.5 (+3.4)**	**87.2 (+1.3)**	**1806 (+58)**	62.2 (+0.2)	70.2 (-)	79.0 (+0.5)

that requires a complex heuristic visual search pipeline.

Integration with Open-sourced LMMs. To demonstrate the versatility of our IVM model, we further integrate it into an open-sourced LMM, LLaVA-7B (Liu et al., 2023a). We conduct extensive experiments across various benchmarks, including EgoThink (Cheng et al., 2024), POPE (Li et al., 2023c), MME (Fu et al., 2023), GQA (Hudson & Manning, 2019), SQA (Lu et al., 2022), and VQAv2 (Goyal et al., 2017). As shown in Table 2, our IVM-enhanced LLava-7B gains consistent performance improvements, achieving comparable performance to (even surpassing) LLaVA-13B on EgoThink, POPE and MME. Although IVM-enhanced LLaVA-7B and LLaVA-13B (Liu et al., 2023a) have roughly the same number of parameters, the latter integrates more powerful pretrained foundation models. In contrast, our IVM model allows the 7B model to outperform the 13B model by merely simplifying visual input, further validating the power of visual masking.

Meanwhile, IVM-enhanced LLaVA-7B does not show significant gains on GQA, SQA and VQAv2, which is expected, as these benchmarks do not heavily rely on grounding capabilities: VQAv2 and GQA contain relatively simple visual input where most regions of the images are instruction-relevant, while SQA primarily focuses on assessing model reasoning capability.

Comparison with Reasoning Segmentation Model. We also compare against LISA (Lai et al., 2024), which is most analogous to IVM. We provide carefully tailored prompts like *"what should we focus on the image to follow the given instruction? Give me the seg"* to extend LISA into visual masking task. However, even with larger 13B model and extensive tuning of input prompt, masks generated by LISA consistently result in severe performance degradation on all tasks.

Evaluation on Real Robotic Control. We also plug the IVM model into robot control tasks to help robot model improve generalization. Specifically, we evaluate a language-conditioned behavior cloning (LCBC) robot agent trained with or without IVM masked images. Figure 5 clearly demonstrates that without IVM assistance, the LCBC robot agent suffers from severe performance drop when noticeable distractions are applied. With IVM assistance, however, the agent consistently pays close attention to correct instruction-related image regions, performing robustly against diverse distractions such as human disturbances and numerous task-irrelevant objects of various colors and shapes. This demonstrates promising potentials of using IVM to enhance embodied agents to follow complex instructions in unseen scenes

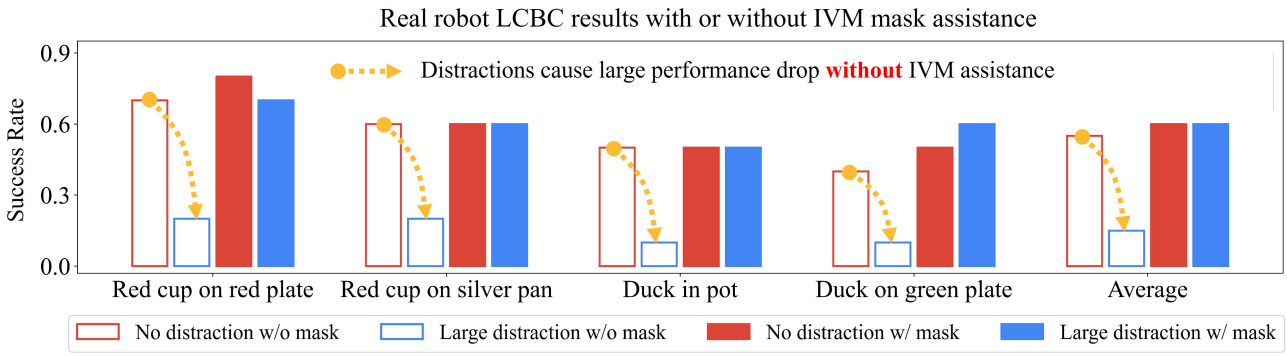


Figure 5. Real robot results with or without IVM assistance. IVM greatly helps LCBC agent to overcome major distractions, enjoying better robustness and generalization. See Appendix E.6 for experiment setups.

with plenty distractions.

4. Conclusion

In this paper, we introduce Instruction-guided Visual Masking (IVM), a generic and powerful visual grounding method that enhances broad multimodal instruction following tasks in a plug-and-play way. By masking out all instruction-irrelevant image regions, IVM effectively injects superior visual grounding ability to downstream LMMs non-intrusively, significantly boosting both commercial and open-sourced LMMs and achieving state-of-the-art results across numerous challenging multimodal benchmarks. Real robot experiments further demonstrate the versatility of IVM, showcasing the potential to deploy IVM to embodied robotic tasks where failures caused by distractions are long-standing challenges. For further improvement, one promising direction is to finetune LMMs using IVM-generated heatmap as an additional input channel to reduce suboptimal heuristics caused by mask deployment methods. However, due to resource limitation, we leave this for future work and will release our IVM checkpoint as well as the IVM-Mix-1M dataset to help the community further explore relevant directions. More discussion on limitations and future directions can be found in Appendix B.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3, 4, 11, 14

Ajay, A., Du, Y., Gupta, A., Tenenbaum, J., Jaakkola, T., and Agrawal, P. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022. 14

Alayrac, J.-B., Carreira, J., and Zisserman, A. The visual centrifuge: Model-free layered video representations. In *CVPR*, 2019. 11

Aminabadi, R. Y., Rajbhandari, S., Zhang, M., Awan, A. A., Li, C., Li, D., Zheng, E., Rasley, J., Smith, S., Ruwase, O., and He, Y. Deepspeed inference: Enabling efficient inference of transformer models at unprecedented scale, 2022. 14

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 4, 9

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 10

Black, K., Nakamoto, M., Atreya, P., Walke, H. R., Finn, C., Kumar, A., and Levine, S. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=c0chJTSbci. 9

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 9

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 9

Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., and Liu, Y. Egothink: Evaluating first-person perspective thinking capability of vision-language models, 2024. 3, 4, 14

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. 14

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B. A., Fung, P., and Hoi, S. C. H. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. 4

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 14, 15

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1, 9

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J. B., Schuurmans, D., and Abbeel, P. Learning universal policies via text-guided video generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bo8q5MRcwy. 9

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 4, 14

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 13

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017. 1, 4, 9, 14

Gupta, A., Dollár, P., and Girshick, R. Lvis: A dataset for large vocabulary instance segmentation, 2019. 9

Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F. S., and Shah, M. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9235–9244, 2022. 1, 11, 12

Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G., and Levine, S. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023. 14

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 14, 15

He, R., Cascante-Bonilla, P., Yang, Z., Berg, A. C., and Ordonez, V. Learning from models and data for visual grounding, 2024. 9

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 14

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9. 13

Hu, X., Li, J., Zhan, X., Jia, Q.-S., and Zhang, Y.-Q. Query-policy misalignment in preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UoBymIwPJR. 10

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 4, 14

Jadon, S. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pp. 1–7. IEEE, 2020. 13

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 13

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128 (7):1956–1981, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL http://dx.doi.org/10.1007/s11263-020-01316-z. 11

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J. Lisa: Reasoning segmentation via large language model, 2024. 1, 2, 3, 4, 9, 10, 11, 12, 13

Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a. 4

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b. 4, 12

Li, J., Zheng, J., Zheng, Y., Mao, L., Hu, X., Cheng, S., Niu, H., Liu, J., Liu, Y., Liu, J., et al. Decisionnce: Embodied multimodal representations via implicit preference learning. *arXiv preprint arXiv:2402.18137*, 2024. 3, 14

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c. 1, 3, 4, 14

Liang, Y., Bai, Y., Zhang, W., Qian, X., Zhu, L., and Mei, T. Vrr-vg: Refocusing visually-relevant relationships, 2019. 9, 11

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014. 11

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a. 4, 9, 14

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/. 9

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b. 3

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024c. 1, 9, 11, 17

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023b. 17

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2017. 14, 15

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. 4, 14

Manyika, J. and Hsiao, S. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2, 2023. 4

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 10

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 14

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020. 14, 15

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 11, 12

Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., and Li, H. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024. 1, 9

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 15

Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., and Wang, J. Alpha-clip: A clip model focusing on wherever you want, 2023. 1, 10, 11, 12

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. 9

Vuong, Q., Levine, S., Walke, H. R., Pertsch, K., Singh, A., Doshi, R., Xu, C., Luo, J., Tan, L., Shah, D., Finn, C., Du, M., Kim, M. J., Khazatsky, A., Yang, J. H., Zhao, T. Z., Goldberg, K., Hoque, R., Chen, L. Y., Adebola, S., Sukhatme, G. S., Salhotra, G., Dass, S., Pinto, L., Cui, Z. J., Haldar, S., Rai, A., Shafiullah, N. M. M., Zhu, Y., Zhu, Y., Nasiriany, S., Song, S., Chi, C., Pan, C., Burgard, W., Mees, O., Huang, C., Pathak, D., Bahl, S., Mendonca, R., Zhou, G., Srirama, M. K., Dasari, S., Lu, C., Fang, H.-S., Fang, H., Christensen, H. I., Tomizuka, M., Zhan, W.,

Ding, M., Xu, C., Zhu, X., Tian, R., Lee, Y., Sadigh, D., Cui, Y., Belkhale, S., Sundaresan, P., Darrell, T., Malik, J., Radosavovic, I., Bohg, J., Srinivasan, K., Wang, X., Hansen, N., Wu, Y.-H., Yan, G., Su, H., Gu, J., Li, X., Suenderhauf, N., Rana, K., Burgess-Limerick, B., Ceola, F., Kawaharazuka, K., Kanazawa, N., Matsushima, T., Matsuo, Y., Iwasawa, Y., Furuta, H., Oh, J., Harada, T., Osa, T., Tang, Y., Kroemer, O., Sharma, M., Zhang, K. L., Kim, B., Cho, Y., Han, J., Kim, J., Lim, J. J., Johns, E., Palo, N. D., Stulp, F., Raffin, A., Bustamante, S., Silvério, J., Padalkar, A., Peters, J., Schölkopf, B., Büchler, D., Schneider, J., Guist, S., Wu, J., Tian, S., Shi, H., Li, Y., Wang, Y., Zhang, M., Amor, H. B., Zhou, Y., Majd, K., Ott, L., Schiavi, G., Martín-Martín, R., Shah, R., Bisk, Y., Bingham, J. T., Yu, T., Jain, V., Xiao, T., Hausman, K., Chan, C., Herzog, A., Xu, Z., Kirmani, S., Vanhoucke, V., Julian, R., Lee, L., Ding, T., Chebotar, Y., Tan, J., Liang, J., Mordatch, I., Rao, K., Lu, Y., Gopalakrishnan, K., Welker, S., Joshi, N. J., Devin, C. M., Irpan, A., Moore, S., Wahid, A., Wu, J., Chen, X., Wohlhart, P., Bewley, A., Zhou, W., Leal, I., Kalashnikov, D., Sanketi, P. R., Fu, C., Xu, Y., Xu, S., brian ichter, Hsu, J., Xu, P., Brohan, A., Sermanet, P., Heess, N., Ahn, M., Rafailov, R., Pooley, A., Byrne, K., Davchev, T., Oslund, K., Schaal, S., Jain, A., Go, K., Xia, F., Tompson, J., Armstrong, T., and Driess, D. Open x-embodiment: Robotic learning datasets and RT-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023*, 2023. URL https://openreview.net/forum?id=zraBtFgxT0. 9, 11

Walke, H. R., Black, K., Zhao, T. Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A. W., Myers, V., Kim, M. J., Du, M., et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023. 14, 15

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 9

Wu, P. and Xie, S. V*: Guided visual search as a core mechanism in multimodal llms, 2023. 3, 4, 9, 10, 14, 15, 16, 18

Wu, T.-H., Biamby, G., Chan, D., Dunlap, L., Gupta, R., Wang, X., Gonzalez, J. E., and Darrell, T. See, say, and segment: Teaching lmms to overcome false premises, 2023. 9

Xu, H., Zhan, X., Yin, H., and Qin, H. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *International Conference on Machine Learning*, pp. 24725–24742. PMLR, 2022. 3, 10, 12

Yan, S., Bai, M., Chen, W., Zhou, X., Huang, Q., and Li, L. E. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling, 2024. 9

Yang, S., Du, Y., Ghasemipour, S. K. S., Tompson, J., Kaelbling, L. P., Schuurmans, D., and Abbeel, P. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sFyTZEqmUY. 9

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 9, 11

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016. 2, 9, 10, 11, 16

Zhang, W., Xu, H., Niu, H., Cheng, P., Li, M., Zhang, H., Zhou, G., and Zhan, X. Discriminator-guided model-based offline imitation learning. In *Conference on Robot Learning*, pp. 1266–1276. PMLR, 2023. 12

Zhao, T. Z., Kumar, V., Levine, S., and Finn, C. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 14

Zheng, Y., Li, J., Yu, D., Yang, Y., Li, S. E., Zhan, X., and Liu, J. Safe offline reinforcement learning with feasibility-guided diffusion model. *arXiv preprint arXiv:2401.10700*, 2024. 14

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 4, 9

A. Related Work

Large Multimodal Models. LLaVA (Liu et al., 2024c) first demonstrates promising capabilities in following complex instructions. Subsequent works such as LLaVA-1.5 (Liu et al., 2023a), MiniGPT4 (Zhu et al., 2023) Qwen-VL (Bai et al., 2023) and CogVLM (Wang et al., 2023), further enhance LMMs via refined model design and enriching the quality of training data, achieving state-of-the-art performance on diverse downstream tasks including visual grounding (Liang et al., 2019), visual reasoning (Thrush et al., 2022), visual question and answering (Goyal et al., 2017). Moreover, by integrating the robotics action modality, LMMs perform versatile planning and manipulation in instruction-driven robotics tasks. Notable studies in this line of inquiry include PaLM-E (Driess et al., 2023), the series of RT models (Brohan et al., 2022; 2023; Vuong et al., 2023), and text-guided video planning diffusion models (Du et al., 2023; Yang et al., 2024; Black et al., 2024). Despite the success, LMMs still struggle with complex visual grounding challenges, often misreading instruction-irrelevant visual contents (Figure 1). To address this, researchers have tried to adapt existing visual modules to higher-resolution images to obtain better perception (Liu et al., 2024a), but with limited improvement.

Visual Grounding Tasks. Visual grounding requires precisely localizing image regions corresponding to a referring expression, among which the RefCOCO series (Yu et al., 2016) is the most well-known benchmark, and numerous public visual grounding data are available (Liang et al., 2019; Young et al., 2014; Gupta et al., 2019). Recently, LMMs incorporate these visual grounding data via visual-instruction tuning (Liu et al., 2024c; 2023a; Zhu et al., 2023), establishing new SOTA in this area (Shao et al., 2024). To further broaden the reasoning ability of visual grounding, LISA (Lai et al., 2024) introduces a new task, reasoning segmentation, which demands higher capabilities in instruction comprehension. However, visual grounding is still limited to align simple instruction with specific objects, which cannot adapt to more complex instruction following tasks (*e.g.* Figure 2).

Visual Grounding Augmented LMMs. Recently, a series of visual grounding methods emerged to enhance the performance of LMMs in complex visual scenes. V* (Wu & Xie, 2023) employs a heuristic search strategy to search, locate, and crop image areas relevant to instructions through a multi-step iterative process. VisualCot (Shao et al., 2024) is trained end-to-end with a customized dataset to achieve target localization capabilities. These two methods allow LMMs to dynamically focus on visual inputs until the correct answer is derived. However, these complex inference pipelines lead to substantial computational overhead, and their heuristic designs further hinder the extension beyond VQA to other multimodal instruction following tasks such as robotic control.

Besides these explicit strategies incorporating additional visual grounding modules, other studies pursue refining data or introducing extra training targets to enhance the grounding capabilities of LMMs implicitly. ViGor (Yan et al., 2024) proposes a fine-grained reward modeling to enhance visual grounding of LMMs, and SynGround (He et al., 2024) introduces a pragmatic framework for image-text-box synthesis tailored for visual grounding. These methods, however, are primarily focused on the visual grounding task itself, overlooking its influence on downstream multimodal instruction following tasks. Distinct from previous efforts, this paper introduces a generic visual grounding model that is adaptable to any multimodal instruction following tasks, and provides a systematic investigation into the advantages of integrating an additional visual grounding model into downstream applications.

B. Limitation and Future Work

Here, we discuss our limitations, potential solutions and interesting future works.

1. **Computational Overhead**. Note that IVM introduces additional parameters and computational overhead to directly enhance visual grounding ability of LMMs, which in turn indirectly improve the VQA performances. However, more VQA performance gains can be obtained if the same amount of additional parameters are end-to-end trained directly on VQA data (LLaVA-13B v.s IVM-Enhanced LLaVa-7B in Table 2).

   Solution and future work: Nevertheless, this is quite reasonable because IVM primarily focuses on improving the visual grounding ability, but accurate VQA also requires other abilities which can be learned through end-to-end training. End-to-end training, however, requires tremendous VQA data to implicitly and slowly improve the visual grounding ability, which is quite data-intensive. Both Table 1 and Figure 1 can show that even trained on billions of data, GPT4-V still performs subpar on tasks that require strong visual grounding ability. IVM, instead, can significantly boost the visual grounding ability of GPT4-V using just 7B parameters and less computations. One promising and interesting future direction is to include some auxiliary tasks to directly absorb the strong visual grounding ability in the IVM-Mix-1M dataset through end-to-end training like (Wu et al., 2023).

2. **Data Quality**. Due to task complexity, the machine-annotated data in IVM-Mix-1M inevitably includes wrong labels that mistakenly exclude instruction-sensitive image regions or suboptimal labels that not

fully mask out all instruction-irrelevant areas. These inaccuracies may lead to suboptimal IVM model. We propose a DWSL framework to tackle this. However, the DWSL framework relies on a learned discriminator and a human-designed $f(x)$ function, which may not exclude all inaccuracies.

Solution and future work: We have clearly demonstrated in Figure 11 that with a simple $f(x)$ and a lightweight discriminator, DWSL consistently outperforms the naive Supervised Learning (SL), doing pretty well on prioritizing good samples and meanwhile identifying inaccurate labels. To further enhance this, one can use other advanced techniques such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022; Hu et al., 2024) to provide more fine-grained judgement on annotation qualities, or resort a theoretical-soundness $f(x)$ (Xu et al., 2022) to achieve better results. In addition, one can also use our pretrained IVM model to directly generate high-quality heatmaps to enhance the machine annotations.

3. **Mask Deployment Methods**. In this paper, we directly use the simple post-processing method to apply the IVM generated heatmaps on images, which then are fed into LMMs to perform downstream tasks. However, these post-processing methods introduce some heuristics, which may be suboptimal for downstream LMMs. In addition, LMMs may not see many masked images during pretraining, thus some distributional shift may occur.

   Solution and future work: Although these limitations exist, IVM still obtain consistent improvements using diverse mask deployment methods, as shown in Table 6, which showcases the great versatility of IVM to inject visual grounding abilities. To further improve this, one strategy is employ some task-specialized visual search method (Wu & Xie, 2023), but will bring many computational load during inference and limit the versatility on embodied agents. Another promising direction is directly using the IVM generated heatmaps as an additional input channel to finetune the LMMs like (Sun et al., 2023), which can fully eliminate the heuristics of post-process methods, may bring larger performance gains. Due to resources limits, we leave this for a future work.

4. **Fine-grained Heatmaps**. Note that the IVM generated heatmaps cannot provide exact semantic object segmentation with clear contours like reasoning segmentation (Lai et al., 2024) offers.

   Discussions: We want to clarify that this is an advantage of the IVM model rather than a limitation. This is because of the ambiguous nature of the visual masking task. For this task, the ground truth heatmaps are mostly less semantic-meaningful for annotations as discussed in 2. So, we ensemble the annotation proposals from different visual grounding methods for data annotation, which will make the trained IVM model robust to include instruction-relevant image areas, rather than being aggressive to exclude some instruction-sensitive pixels like reasoning segmentation (Lai et al., 2024) does illustrated in Figure 2.

Overall, although some limitation exist, we have thoroughly discussed potential solutions to these limitations. Moreover, in this paper, we have demonstrated the superior effectiveness and versatility of IVM to directly inject strong visual grounding ability to downstream LMMs or embodied agents, representing a pioneer effort to extend traditional visual grounding methods towards a more complex and generic setting that covers diverse multimodal instruction following tasks.

C. Broader Impact

This paper aims to advance the field of artificial intelligence, where no significant negative social impact is observed in this paper. The IVM-Mix-1M may contain some potential privacy issues and biases. However, in this paper, nearly all data are collected from open-sourced data, which have been well peer-reviewed, thus resolved this ethical concern.

D. Data Preparation

To train an IVM model, the first main challenge is the scarcity of training data. Most existing Visual Grounding (VG) datasets (Yu et al., 2016; Lai et al., 2024) typically feature simple instructions focused primarily on prominent objects within images, lacking both diversity and complexity required for IVM. To tackle this, we compiled one million data from various sources, including labeled visual grounding, unlabeled multimodal instruction following, and robotics data. As outlined in Section 2, scaling human annotations is challenging due to the high complexity of such data. Therefore, we introduce an *LLM-empowered Mixture of Expert pipeline* that integrates SOTA visual grounding models to efficiently generate reliable annotations. We further manually annotate a smaller dataset to compensate inaccuracies in auto-generated labels. The resulted combined dataset, IVM-Mix-1M, comprises one million data samples ready for IVM training, which will be released for future study upon acceptance.

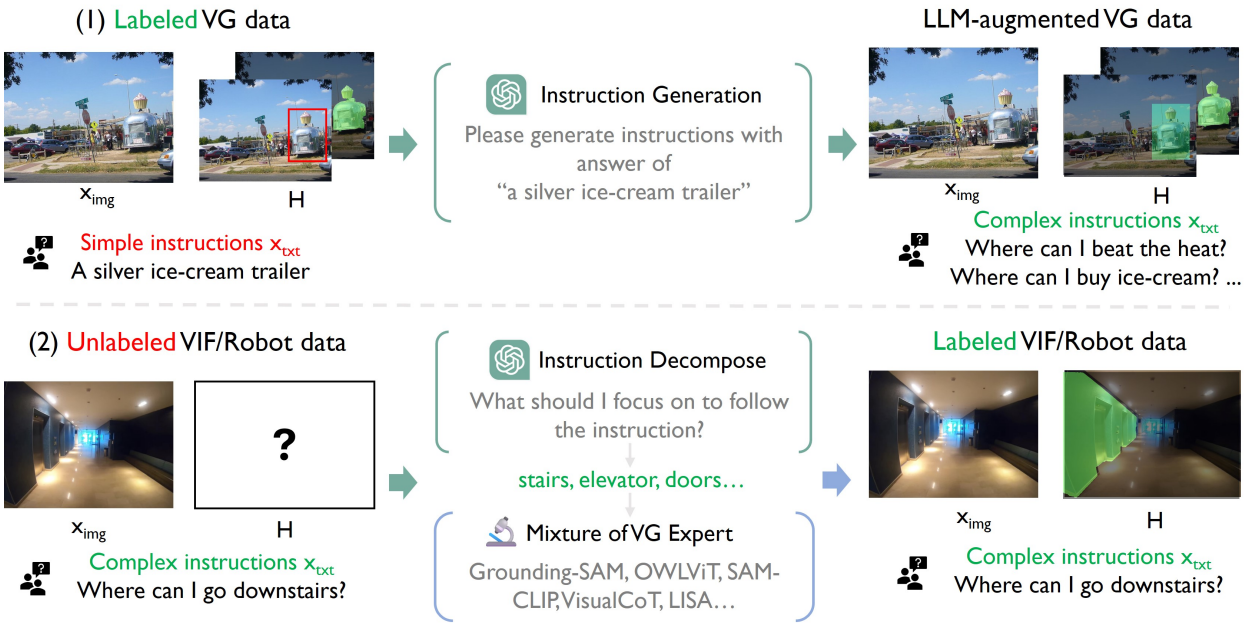


Figure 6. LLM-empowered Mixture-of-Expert pipeline for auto-annotation. (1) For labeled VG data, we utilize an LLM to generate complex instruction annotations. (2) For unlabeled VIF or robot data, we first use an LLM to simplify the instruction and then leverage a mixture of VG models to generate candidate labels.

D.1. LLM-empowered Mixture of Expert Annotation Pipeline.

Leveraging the power of LLM, this pipeline can efficiently generate high-quality annotation, which consists of two components (Figure 6): 1) *Labeled visual grounding data.* We collect 250K labeled *VG* data from multiple sources including VG caption (Liang et al., 2019), Flickr30K (Young et al., 2014), VSR (Alayrac et al., 2019), OpenImage (Kuznetsova et al., 2020), and RefCoCo (Yu et al., 2016; Lin et al., 2014), which provide bounding boxes with simple instructions for each image. To increase the diversity and complexity of instructions, we utilize GPT-4 (Achiam et al., 2023), known for its robust language understanding and generation capabilities, to create diverse instruction-answer pairs based on existing language instructions. 2) *Unlabeled Visual-Instruction-Following (VIF) and robotics data.* We sample a 700K subset from LLaVA-Instruction-tuning (Liu et al., 2024c) for VQA-type data, and a 50K subset from OpenX (Vuong et al., 2023) for robotics data. Given that these data lack grounded labels but contain complex instructions, we use GPT-4 to simplify the language instructions by prompting it to infer the names of targeted objects necessary for following the instructions. These simplified instructions then guide existing VG models to generate candidate labels. To ensure the quality of these labels and compensate for the ambiguous nature of the IVM task, we integrate proposals from several *VG* experts, such as Grounding-Sam (Ren et al., 2024), LISA (Lai et al., 2024), AlphaClip (Sun et al., 2023), and OwlViT (Gupta et al., 2022), via an ensemble approach.

D.1.1. Labeled Visual Grounding Data

For labeled visual grounding data, We provide the following prompt to drive GPT-4 (Achiam et al., 2023) to generate more complex instructions based on given language annotations.

> [Image Description] %s
>
> [System] You are an AI visual assistant, and you are seeing a single image. What you see are part of the image and are provided with a simple phrase. Please generate any instructions that can be executed based on the content of the picture described, including simple queries about the content of the picture, such as the object types, counting the objects, object actions, relative positions between objects, etc. Also consider more complex questions that require reasoning. For example, you can ask what time it is now for a clock and what can I use to clean the room for a broom. Ensure that the questions you ask can be clearly answered only based on what you see. Please generate as many five questions as possible and return them in a single line separated by

';' and avoid any other output.

D.1.2. Unlabeled Visual Instruction Following Data

For unlabeled visual instruction following data, we first try to simplify complex instructions. Specifically, we employ GPT-4 to infer the necessary object for executing the given instructions based on these instructions and a simple image caption. If the dataset lacks captions, they can be generated using an existing caption model like BLIP-Caption (Li et al., 2023b). Below, we outline the prompts specifically designed for GPT-4.

> [Image Caption] %s
>
> [Instruction] %s
>
> [System] You are an helpful AI assistant. I need to reply to the previous instruction based on an image, and I have a simple caption for the image. Please note that there may be objects in the image that I did not detect. Since you cannot view the image, please list any potential objects that might influence my responses, separated by semicolons, in a single line without any additional output. If you believe that the number of objects could be too extensive and might hinder my judgment, print 'None'.

With the simplified instruction, we can adopt existing visual grounding models to generate the candidate label. Specifically, we utilize four models: AlphaCLIP (Sun et al., 2023), LISA (Lai et al., 2024), OwVIT (Gupta et al., 2022) and Grounding-SAM (Ren et al., 2024) and the inference pipelines are provided in the official implementation of these models.

D.2. Manual Annotation

. Despite integrating the most advanced models, the auto-generation design still faces challenges that can lead to data inaccuracies. First, employing LLM to simplify or complicate language annotations without considering image content can introduce uncontrollable biases. Second, as the task exceeds the capabilities of existing models, it becomes impossible to totally exclude low-quality annotations that contain irrelevant visuals or mistakenly filter out critical contents. Thus, to enhance the overall quality of the dataset, we further manually annotate a 10K subset of the constructed dataset to inject human expert knowledge.

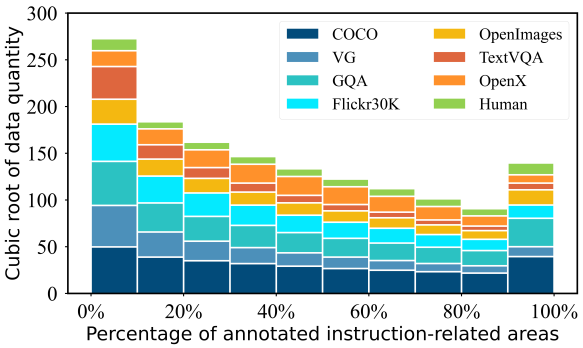


Figure 7. Data analysis on the IVM-Mix-1M dataset: data quantity v.s percentage of instruction-related areas.

D.3. Data Analysis

. Here, we provide quantitative analysis on the IVM-Mix-1M dataset. Figure 7 depicts the data quantities w.r.t the percentage of annotated instruction-related image area. Here, each ratio range is further categorized by different data sources, where manually annotations are treated as a separate category (Human), while all others are machine-generated. Our analysis reveals that the instruction-related image regions only occupy a small fraction of the total image area (*e.g.* most data have less than 40% instruction-relevant image regions), indicating that most visual contents may cause distraction and corroborating the necessity of visual masking for instruction following tasks.

E. Training and Evaluation Details

E.1. Discriminator-Weighted Supervised Learning Framework

The challenge now is to train the IVM model with a small high-quality human-annotated dataset ($\mathcal{D}_e$) as well as a large but mixed-quality auto-generated dataset ($\mathcal{D}_o$). Training naively on the combined dataset may yield suboptimal results due to inaccuracies in auto-generated labels, while solely using limited human-annotated data is insufficient. Inspired by recent advances in imitation learning using mixed-quality data (Xu et al., 2022; Zhang et al., 2023), we employ a Discriminator-Weighted Supervised Learning (DWSL) framework to effectively leverage the strengths of both auto- and human-annotated data.

Discriminator Training. Specifically, we introduce a discriminator d optimized by Eq. (1) to assign high weights to high-quality annotations and vice versa:

$$\min_d \mathbb{E}_{(\mathbf{x}_{\text{img}},\mathbf{x}_{\text{txt}},\mathbf{H})\sim\mathcal{D}_e}\left[-\log d(\mathbf{x}_{\text{img}},\mathbf{x}_{\text{txt}},\mathbf{H})\right] + \mathbb{E}_{(\mathbf{x}_{\text{img}},\mathbf{x}_{\text{txt}},\mathbf{H})\sim\mathcal{D}_o}\left[-\log(1-d(\mathbf{x}_{\text{img}},\mathbf{x}_{\text{txt}},\mathbf{H}))\right], \tag{1}$$

where $(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}}, \mathbf{H})$ are image-instruction-heatmap pairs

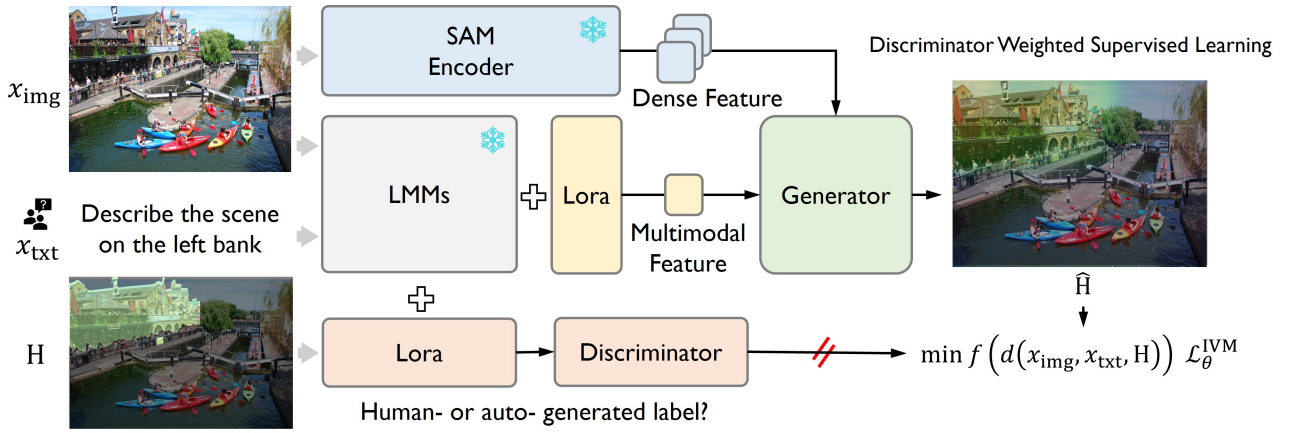


Figure 8. IVM model architecture. A frozen vision backbone and a LORA-tuned LMM are utilized to extract dense image features and multimodal representations, respectively. These features are then fed into a generator for dense prediction and a discriminator for generating training weights. See Appendix E.3 for more details.

sampled from $\mathcal{D}_o$ and $\mathcal{D}_e$ datasets. Eq. (1) is similar to the one in GAN (Goodfellow et al., 2014), but the "fake" data in (Goodfellow et al., 2014) is replaced by machine-generated data from $\mathcal{D}_o$. After training with Eq. (1), the discriminator d assigns high weights to high-quality human-annotated data from $\mathcal{D}_e$ and relatively high values to similarly high-quality data from $\mathcal{D}_o$ that aligns with human preferences, acting as a judge for annotation quality.

Discriminator-weighted IVM Training. Then, we apply the trained discriminator as a weighting function for the IVM training objective:

$$\min_\theta \mathbb{E}_{(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}}, \mathbf{H}) \sim \mathcal{D}_o \cup \mathcal{D}_e} \left[f\left(d(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}}, \mathbf{H})\right) \mathcal{L}_\theta^{\text{IVM}}(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}}, \mathbf{H}) \right], \tag{2}$$

$$\mathcal{L}_\theta^{\text{IVM}}(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}}, \mathbf{H}) = \lambda_{\text{bce}} \mathbf{BCE}(\hat{\mathbf{H}}_\theta, \mathbf{H}) + \lambda_{\text{dice}} \mathbf{DICE}(\hat{\mathbf{H}}_\theta, \mathbf{H}), \tag{3}$$

where λ_{bce} and λ_{dice} are set to 1.0 and 1.0 to balance the binary cross-entropy loss (**BCE**) and the DICE loss for segmentation (**DICE**) (Jadon, 2020), respectively. $f(x) \geq 0$ can be any non-negative, non-decreasing function. For simplicity, we set $f(x) := \min(\max(0.1, x), 1)$. This allows the weighting function $f(d(\cdot))$ in Eq. (2) to dynamically prioritize training with high-quality data determined by the discriminator d. This approach maximizes the usage of reliable annotations in $\mathcal{D}_o$ to compensate for the small $\mathcal{D}_e$, while minimizing the impact of low-quality data in $\mathcal{D}_o$, thus optimizing performance.

E.2. Model Architecture

The overall model framework is illustrated in Figure 8. Due to its complexity, IVM requires both reasoning and precise localization of the target object, closely paralleling reasoning segmentation (Lai et al., 2024). Consequently, for the heatmap generator, we adopt a model design similar to that of LISA (Lai et al., 2024). Specifically, we first extract dense image features using an isolated vision backbone and multimodal representation from an LMM, which processes image-instruction pairs. These two types of features are then fed into a lightweight generator that integrates them to produce a dense prediction.

For the discriminator, we deploy a lightweight discriminator that encodes the segmentation map using a two-layer convolution network. This discriminator interacts with the outputs of the LMMs through multiple cross-attention operators and finally outputs a quality score for each sample.

Trainable Parameters. To enhance training efficiency, we freeze the pre-trained large foundation models and perform LORA finetuning (Hu et al., 2022). The vision backbone, inherited from *Segment Anything Model* (Kirillov et al., 2023), is completely frozen, while the lightweight generator and discriminator are fully finetuned. Notably, we utilize a shared LMM for both the generator and discriminator branches but employing separate LORA parameters to avoid interference between the two tasks.

E.3. Architecture Details

In this section, we primarily focus on the architectural design of the lightweight generator and discriminator, as both the Language Model Multitask (LMM) and the vision backbone are derived from the powerful foundation models (LLaVA & SAM). Both the generator and discriminator utilize the same transformer-based decoder block, as depicted in Figure 9. We employ two such blocks for both the generator and discriminator. Specifically, the generator produces dense predictions by upscaling the output features of the decoder block through a straightforward upsampling operation. In contrast, the discriminator first employs a

two-layer convolutional downsampling network to encode segmentation labels. This network, in conjunction with the decoder block and a simple MLP (multi-layer perceptron) head, outputs the weights.

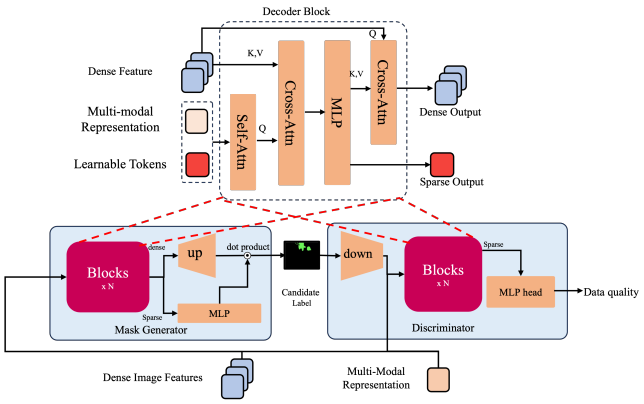


Figure 9. Generator/Discriminator Architecture Details

E.4. Training Details

We adopt 8 NVIDIA 80G A100 GPUs and take 4 days to train our IVM model. The training scripts are based on deepspeed (Aminabadi et al., 2022) engine and the training hyperparameters can be found in Table 3.

Table 3. Hyper-parameters for pretraining.

config	value
training iteration	200K
optimizer	AdamW (Loshchilov & Hutter, 2017)
learning rate	1×10^{-5}
batch size	32
weight decay	0
optimizer momentum	β_1, β_2=0.9, 0.95
data augmentation	*RandomCropResize*

E.5. Multimodal Benchmarks Evaluations

We evaluate our IVM on diverse multimodal benchmarks, including general VQA (VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), MME (Fu et al., 2023)), first-person perspective QA (EgoThink (Cheng et al., 2024)), scientific QA (SQA (Lu et al., 2022)), hallucination adversarial QA (POPE (Li et al., 2023c)) and V* (Wu & Xie, 2023), a recently proposed challenging benchmark with high-resolution and complex visual input.

Our evaluation employs a two-stage inference pipeline: the image is firstly simplified by IVM-generated heatmap and mask deployment methods; Subsequently, the simplified image is fed into downstream LMMs(GPT4-V (Achiam et al., 2023), LLaVA (Liu et al., 2023a)) without finetuning. We adhere to the official procedures of each benchmark to evaluate the output of LMMs and report the results.

E.6. Real Robot Evaluations

Task descriptions. The real robot experiments evaluate several `pick and place` manipulation tasks that require strong visual grounding abilities. Specifically, we evaluate on 4 tasks as shown in Table 4, following the task definitions in DecisionNCE (Li et al., 2024). For each task, we collect around 100 demonstrations using the demonstration collection system in BridgedataV2 (Walke et al., 2023). We take both a side camera view and a wrist camera view as the vision inputs, as shown in Figure 10. For each demonstration, the environmental steps are around 50 steps. During data collection, the object and robot locations are randomly initialized, and the scene also has lots of randomly located distractors with varied shape and color.

(a) WidowX RealRobot Side Camera View (b) WidowX RealRobot Wrist Camera View

Figure 10. Visual input view for LCBC policy.

Training details. Here, we train Language-Conditioned Behavior Cloning (LCBC) policies using DDPM (Ho et al., 2020) loss since diffusion policies are good at fitting complex data distributions (Zheng et al., 2024; Walke et al., 2023; Ajay et al., 2022), especially human demonstrations (Chi et al., 2023). For model architecture, the side and wrist images are augmented and then passed through a shared ResNet50 (He et al., 2016) image encoder and get an image embedding for each camera view, following (Walke et al., 2023). As the downstream data is quite limited, we load the ImageNet (Deng et al., 2009)-pretrained ResNet50 image encoder and further train it on the small robot data. Meanwhile, the language instruction is passed through a frozen T5 text-encoder (Raffel et al., 2020), which is fused into the image encoder via Film conditioning layers (Perez et al., 2018). Then, this language-conditioned image embedding is passed through a MLPs with residual connections similar to IDQL (Hansen-Estruch et al., 2023), which then outputs the predicted noise in DDPM (Ho et al., 2020). To obtain smoothed policy rollouts, we adopt Action Chunking and Temporal Ensemble from ACT (Zhao et al., 2023) with a chunking size 4 rather than 100 in (Zhao et al., 2023) because the episode horizons in this paper are only around 50. The LCBC policies are trained either on the original

Table 4. Real Robot Tasks

Environment ID	Language Instruction
`Red cup on silver pan`	`Pick up the red cup and place it on the silver pan`
`Red cup on red plate`	`Pick up the red cup and place it on the red plate`
`Duck on green plate`	`Pick up the duck and place it on the green plate`
`Duck in pot`	`Pick up the duck and place it in the pot`

Table 5. Real robot LCBC training details

Backbones	
Visual encoder	Resnet50 (He et al., 2016) (ImageNet (Deng et al., 2009) pretrained)
text encoder	T5 (Raffel et al., 2020) (frozen)
DDPM hyperparameters	
noise schedule	VP (Song et al., 2021)
denoising time steps	25
Other hyperparameters	
Chunking size	4
Optimizer	AdamW (Loshchilov & Hutter, 2017)
Learning rate	1e-4
Lr schedule	cosine annealing
Warm up steps	2000
Batch size	64
Gradient Steps	200K
Augmentation	Yes (Walke et al., 2023)

side camera view (without IVM assistance) or on the IVM-masked side camera view (with IVM assistance) for 200K steps with a batch size of 64. The training can be completed on 2 NVIDIA RTX4090 GPU in 17h. All hyperparameters are summarized in Table 5.

Evaluation details. We first evaluate the trained LCBC policies without strong distractions, where no or only small distractors appear in the image. Then, we add lots of distracting objects with varied shapes and colors, and even introduce strong human disturbance to attack the LCBC policies. For each score reported in Figure 5, we evaluate 10 episodes and report the success rates.

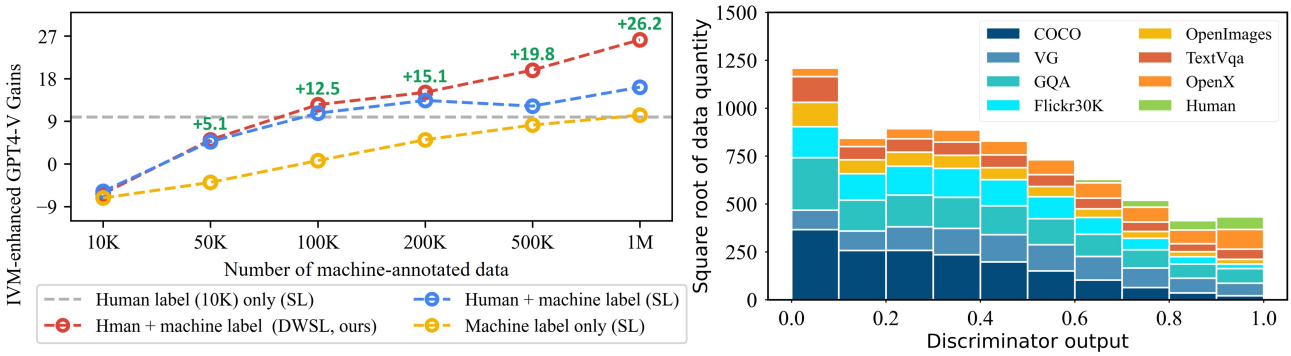
F. More result

F.1. Ablation

We ablate the key components of IVM and report overall accuracy improvement(%) of IVM-Enhanced GPT4-V, evaluated on V* bench (Wu & Xie, 2023) due to its high demand on precise visual grounding abilities.

Training Data. We investigate the impact of IVM-Mix data characteristics on IVM performance from two key perspectives: 1) Large machine-annotated data volume clearly enhances IVM model performance, as illustrated by the progressive improvement in Figure 11 (a) with increased machine-annotated data volume (red, blue and yellow line). This demonstrates the effectiveness of our proposed LLM-empowered Mixture-of-Expert pipeline in generating reliable data for IVM training. 2) Figure 11 (a) also reveals that incorporating human annotations significantly boosts training efficiency (red and blue *v.s* yellow line), highlighting the critical role of introducing human preferences in IVM-Mix-1M dataset, despite its relatively small volume compared to machine-annotated data (only 1:100).

DWSL Framework. We also explore the efficacy of the DWSL framework in Figure 11 (a) by comparing IVM training using: 1) DWSL (red line), 2) traditional Supervised Learning (SL) without DWSL (blue line), and 3) SL on limited human data (gray line). The results demonstrate that DWSL effectively leverages both human- and auto-annotated data, particularly as the volume of machine-annotated data increases, enjoying higher asymptotic performances. This is expected as machine-annotated data often contain inaccuracies and training naively using all these data can lead to suboptimal results. Meanwhile, the limited human data alone cannot provide satisfactory outcomes. DWSL, however, addresses these challenges by dynamically



(a) IVM-enhanced GPT-4V gains trained with different data and methods

(b) Discriminator output values for different data sources

Figure 11. Ablations on training data and the proposed DWSL framework.

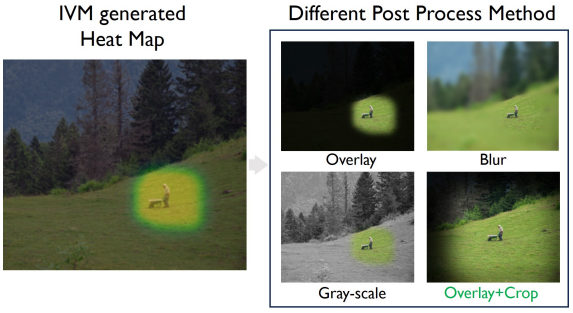


Figure 12. Different mask deployment methods.

Table 6. Ablations on different mask deployment methods on the V*bench.

	Overlay	Blur	Gray-scale
w/ crop	**+26.2**	+24.4	+22.1
w/o	+19.1	+17.2	+10.2

prioritizing good samples and discarding misleading ones, resulting in stable and improved results. This is further illustrated in Figure 11 (b) which visualizes the outputs of the discriminator for each sample, where the discriminator can correctly retain good samples (e.g. Human) and filter out low-quality data with lower weights.

Mask Deployment Strategy. We investigate the impact of mask deployment strategy on downstream applications. While more complex solutions such as visual search algorithms (Wu & Xie, 2023) can be employed, our investigation focuses solely on simpler approaches to understand the intrinsic capabilities of IVM model. Specifically, we examine four basic masking methods: overlay, blur, grayscale, and cropping, as illustrated in Figure 12. In particular, for the crop method, we find the smallest area that retains all the activated (¿0) values in the heatmap and crop it. Table 6 demonstrates that IVM maintains robustness across all simple post-processing methods, where overlay+crop enjoys the most performance enhancement and thus is used as our default mask deployment method.

F.2. Referring Expression Comprehension

As IVM is an extension of traditional visual grounding task, we also evaluate our IVM on RefCoCo, RefCoCo+ and RefCoCog (Yu et al., 2016). We reported the accuracy (IOU-50%) on the validation split in Table 13. As a generalist model capable of handling complex instructions, our IVM achieves performance comparable to that of state-of-the-art (SOTA) specialist models.

F.3. Visualization Result

In this section, we provide more visualization result in VQA-type data as shown in Figure 14.

Failure Case. Although we observe numerous successful instances, our IVM still faces significant challenges, as illustrated in Figure 15. We summarize these challenges into three categories: missing target, misguided target, and insufficient reasoning.

(a) **Missing Target**: Challenges arise when target objects are relatively small and scattered around many separate image corners. In this case, accurately detecting all of the targeted objects is quite difficult. Even specialized open vocabulary detection models struggle with this task. For example, the cup on the right in the image is masked by the IVM mistakenly. However, we still observe that the IVM-generated heatmap for the right cup is partially activated, meaning that IVM have partially focus this regions. We believe by providing more training data, IVM can handle this better.

Figure 13. result in REC

Methods	RefCoCo	RefCoCo+	RefCoCog
Specialist models			
G-DINO-L (Liu et al., 2023b)	90.56	82.75	86.13
Generalist models			
LLaVA-7B (Liu et al., 2024c)	76.29	66.76	70.4
IVM(Ours)	**90.1**	**83.3**	**82.9**

Figure 14. Visualization results of IVM generated masks.

(b) **Misguided Target**: Accurately Localizing tiny target objects is a recognized challenge (Wu & Xie, 2023), especially when similar but more obvious objects are present. For instance, IVM incorrectly focuses on the more centrally located shoes of another man, instead of the shoes of the man wearing the **red hat** at the edge of the picture. However, this instruction is pretty challenging that at first glance, even a human might struggle to spot the man with the red hat in the left corner. We will leave challenge scenarios like this for future research.

(c) **Insufficient reasoning**: The objective of the IVM task is to assist LMMs in extracting visual features more effectively to better follow instructions. Thus, the demands on the model's reasoning capabilities extend far beyond mere object localization. Although IVM demonstrates strong performance, it sometimes overlooks additional image content necessary for accurately following instructions after correctly locating the target object. For instance, while IVM successfully identified the braking motorcycle, it failed to recognize that answering the question requires knowledge of the positions of both motorcycles simultaneously. We attribute this issue to biases in the training data. By incorporating more complex instructions and diversified labels, we anticipate that our model will achieve improved performance

F.4. Robotics Result

Here, we provide more evaluation rollouts of the IVM-assisted LCBC agents under strong distractions. Figure 16 clearly demonstrates that even under strong distractions like the background are full of distracting objects with similar colors or shapes to the targeted objects, and strong human disturbances that adversarially attack the robots, the IVM-assisted LCBC agents can still complete the tasks pretty well, enjoying high-level of generalization and robustness thanks to the superior visual grounding ability injected by IVM. More videos can be found in the supplementary materials.

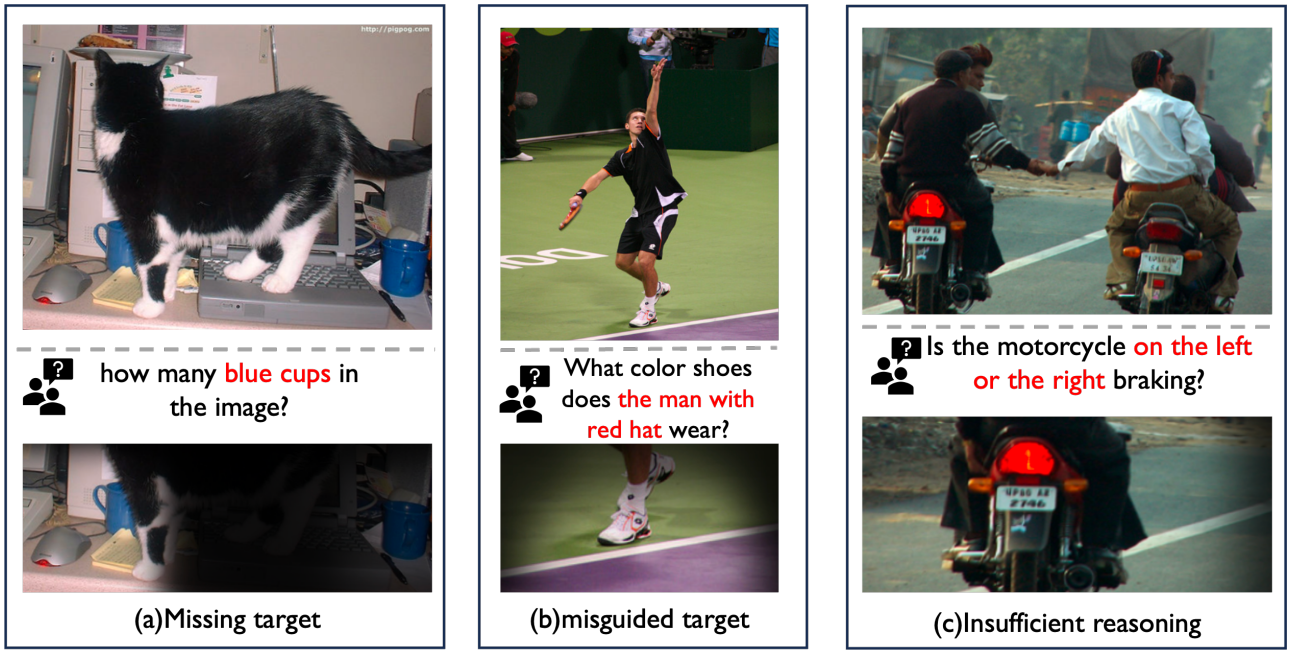


Figure 15. Some failure cases.

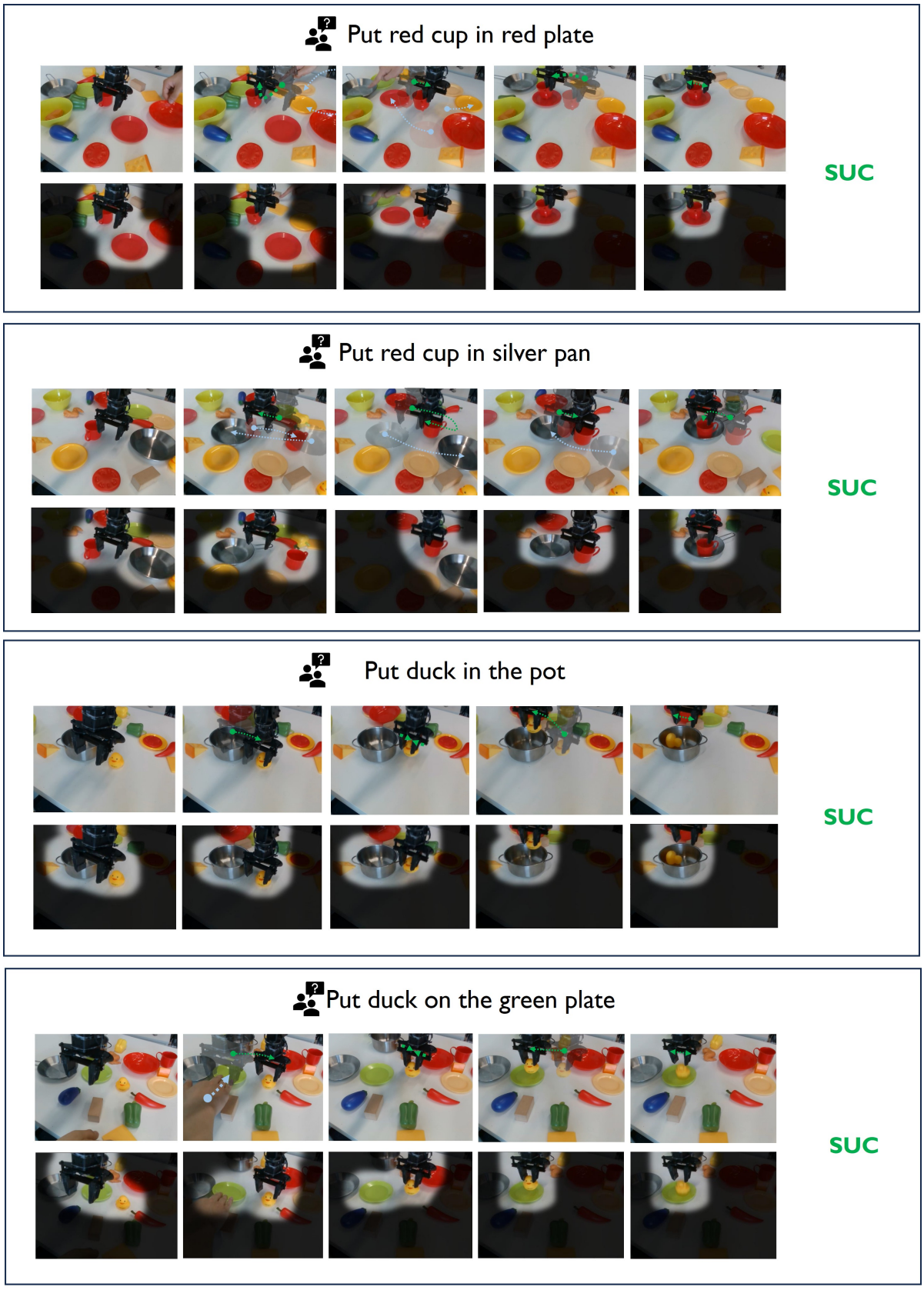


Figure 16. Real robot LCBC results with IVM assistance.