# VGent: Visual Grounding via Modular Design for Disentangling Reasoning and Prediction

## Supplementary Material


Figure 1. Visualizations of VGent's output under single target and multiple targets challenges.

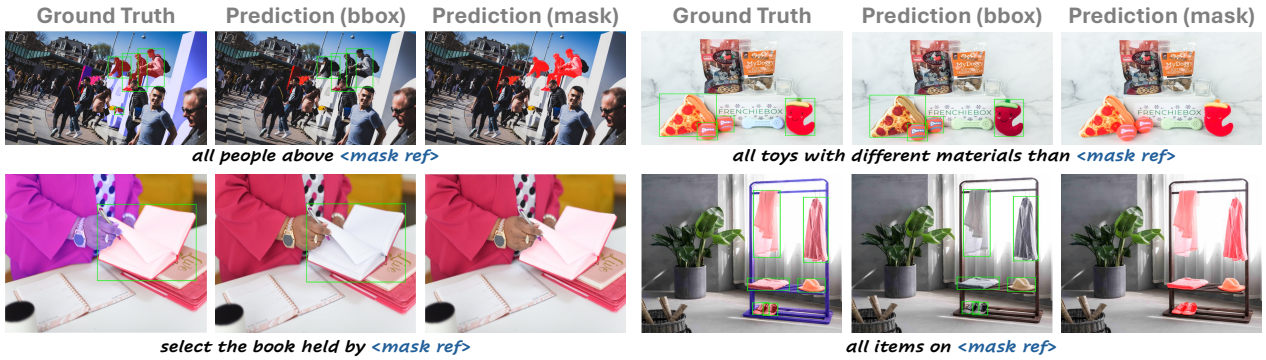


Figure 2. Visualizations of VGent's output under visual reference challenges. Blue masks indicate visual reference regions.

## 1. Additional Qualitative Results

In Fig. 1 and Fig. 2, we present additional qualitative examples to further illustrate the versatility and robustness of VGent across a wide range of grounding conditions, including single-target, multi-target, and visual reference–conditioned multi-target scenarios. These examples highlight VGent's ability not only to localize explicit referents but also to reason over subtle visual cues and contextual relationships in complex scenes.

As shown in Fig. 1 (top-left), VGent successfully identifies the person wearing glasses in a densely crowded environment. Despite the glasses covering only a few pixels and the presence of numerous distractor individuals without glasses, the model accurately grounds the intended target. This demonstrates VGent's strong sensitivity to fine-grained visual attributes and its capability to filter out semantically similar distractors.

Similarly, in Fig. 2 (top-left), VGent effectively resolves a visual reference-conditioned multi-target query, detecting all people above the provided visual reference. The model succeeds even under occlusion and when some targets appear at a small scale due to being farther from the camera. These results illustrate VGent's ability to integrate visual reference signals, reason over relational cues, and maintain stable grounding performance.

## 2. Additional Quantitative Results

In Tab. 1, we further report experimental results on generalized referring expression segmentation (GRES) evaluated on gRefCOCO val split and Reasoning Segmentation (Rea-

Table 1. **Results on generalized referring expression segmentation (GRES) and reasoning segmentation (ReasonSeg).** We highlight the best performance in bold and underline the second best.

| | GRES | | | | ReasonSeg | |
|---|---|---|---|---|---|---|
| Model | F1 | gIoU | cIoU | N-acc | gIoU | cIoU |
| MagNet [4] | - | - | - | - | - | - |
| Groundhog$_{7B}$ [23] | - | - | - | - | - | - |
| GLaMM$_{7B,\,FT}$ [14] | - | - | - | - | - | - |
| u-LLaVA$_{7B}$ [20] | - | - | - | - | - | - |
| UNINEXT-H [21] | - | - | - | - | - | - |
| PSALM$_{1.3B}$ [24] | - | - | - | - | - | - |
| LAVT [22] | - | 58.40 | 57.64 | 49.32 | - | - |
| HDC [13] | - | 68.28 | 65.42 | 63.38 | - | - |
| ReLA [10] | - | 63.60 | 62.42 | 56.37 | 21.3 | - |
| Seg-Zero [11] | - | - | - | - | 57.5 | 52.0 |
| GSVA$_{13B,\,FT}$ [19] | - | 70.04 | 66.38 | 66.02 | - | - |
| SAM4MLLM$_{7B}$ [3] | - | 71.86 | 67.83 | 66.08 | - | - |
| LISA$_{13B,\,FT}$ [7] | - | 65.24 | 63.96 | 57.49 | 61.3 | 62.2 |
| RAS$_{13B}$ [2] | <u>81.74</u> | <u>74.64</u> | **70.48** | <u>69.05</u> | - | - |
| VGent (Ours) | **82.91** | **77.14** | <u>69.33</u> | **83.33** | **62.2** | **64.0** |

sonSeg) evaluated on the ReasonSeg test split. GRES [9] involves an arbitrary number of targets, and ReasonSeg [8] evaluates grounding under complex and implicit language instructions. VGent achieves superior performance, demonstrating the robustness and generalization capability of our framework across diverse grounding scenarios. In particular, VGent achieves a substantial improvement in the GRES N-Acc metric—which evaluates whether the model hallucinates targets in non-target scenarios—surpassing the previous state-of-the-art RAS$_{13B}$ [2] by **+14.28%**. This result highlights the faithfulness of VGent and its significantly reduced tendency to hallucinate outputs.

## 3. Ablation on Upper Bounds

Table 2. Oracle selection for upper-bound performance on Omni-modal Referring Expression Segmentation (ORES).

| Model | Overall | | |
|---|---|---|---|
| | F1 | gIoU | cIoU |
| VGent (Ours) | 71.47 | 68.42 | 75.28 |
| UPN [5] (Oracle) | 91.27 | 79.97 | 81.40 |
| UPN [5] + GLEE [18] (Oracle) | 94.68 | 84.05 | 85.00 |
| UPN [5] + GLEE [18] + SAM [15] (Oracle) | **95.38** | **86.20** | **88.45** |

We evaluate how different detector combinations affect the upper-bound performance of VGent by applying oracle selection on ORES. For F1, we run Hungarian Matching between the grouth truth boxes and proposed boxes, and retain proposals whose IoU exceeds 0.5; for gIoU and cIoU, we keep proposals whose IoA exceeds 0.6. As shown in Table 2, different detectors provide complementary proposals that jointly increase coverage of the ground-truth boxes, thereby raising the achievable upper bound of VGent's performance.

## 4. Details of Implementation

**QuadThinker.** For the QuadThinker component used to initialize VGent's encoder, we perform GRPO training for one epoch based on Qwen2.5-VL-7B [1] using MaskGroups-HQ [2] and VisionReasoner-7K [12], with a batch size of 16 and a learning rate of 1e-6.

**Learnable Query.** Inspired by SegVG [6], we use multiple learnable queries to benefit proposal selection through self-attention within each decoder layer which propagates the global target information. Empirically, we find that using 10 learnable queries yields the best performance, where 5 queries are used to regress the number of targets and 5 are used to regress the number of positive proposals.

**Visual Reference.** MaskGroups-HQ [2] provides visual references in the form of segmentation masks. To integrate these visual references into the language query, we convert each mask into a bounding box. Specifically, we compute the minimum and maximum (x,y) coordinates that tightly enclose the mask, resize the resulting box to the resolution of the model's image input, and round all coordinates to integers. We then replace the placeholder token `<mask-ref>` in the textual query with this coordinate list. For example, the query *"the woman wearing a skirt behind the left side of* `<mask-ref>`*"* becomes *"the woman wearing a skirt behind the left side of [50, 490, 120, 637]"*.

**Training on ORES.** For experiments on ORES, which follows the evaluation split of MaskGroups-HQ [2], we combine proposals from UPN [5], SAM [15], and GLEE [18] during training. We first train on Objects365 [16] for 16K steps using 6 nodes (each with 8×A100-80G GPUs), with a per-GPU batch size of 1 and gradient accumulation of 2. We then train on the mixed dataset of Objects365 [16] and MaskGroups-2M [2], sampled with the 0.3 and 0.7 ratio of them under the same configuration. Finally, we train on the MaskGroups-HQ [2] training split for 48K steps using 1 node of 8×A100-80G GPUs. The BCE loss is weighted by 1 and the L1 loss by 10. We use a learning rate of 2e-5 with linear decay. For box-aware label, proposals with IoU $> 0.6$ are treated as positives and all others as negatives. For mask-aware label, we assign positives using IoA $> 0.6$. All images are resized to $840 \times 840$ resolution.

**Training on REC.** For REC experiments, we follow RAS [2] to further fine-tune on all training splits of RefCOCO, RefCOCO+, and RefCOCOg for 48K steps using 1 node of 8×A100-80G GPUs.

**Training on GRES and ReasonSeg.** For experiments on GRES and ReasonSeg, we fine-tune the checkpoint obtained after pre-training on Objects365 [16] and MaskGroups-2M [2]. During fine-tuning, we reweight the loss for mask-aware labels by a factor of $1 +$ IoA for each proposal on

GRES. All fine-tuning experiments are conducted on their respective training splits for 48K steps using a single node with 8×A100-80G GPUs. We report results based on the best-performing checkpoint and outputs.

**Inference.** We use UPN [5], SAM [15], and GLEE [18] for both training and inference, and for all inference-time speed measurements. The runtime consists of 0.696 seconds for VGent's encoder–decoder, 0.263 seconds for UPN, 0.213 seconds for GLEE, and 1.154 seconds for SAM.

**Ablation Studies.** For ablation experiments, QuadThinker is further trained for four additional epochs when being integrated into VGent. While this extended training does not improve QuadThinker's performance, it consistently yields better overall performance for VGent. All ablation studies are conducted on a single node with 8×A100-80G GPUs.

**Qwen3-VL Evaluation.** Following the official GitHub instructions of Qwen3-VL [17], we use the prompt: *"Locate {Question}, output the bbox coordinates using JSON format."*, where {*Question*} is replaced by the language query input. For consistency with our implementation, the input image is resized to a resolution of 840 × 840. Qwen3-VL outputs bounding boxes in a normalized format, where each coordinate is represented as a relative value multiplied by 1000. During post-processing, we divide the predicted values by 1000 and scale them by the image resolution to recover the absolute bounding box coordinates.

## References


[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Shengcao Cao, Zijun Wei, Jason Kuen, Kangning Liu, Lingzhi Zhang, Jiuxiang Gu, HyunJoon Jung, Liang-Yan Gui, and Yu-Xiong Wang. Refer to anything with vision-language prompts. *arXiv preprint arXiv:2506.05342*, 2025. 2

[3] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. SAM4MLLM: Enhance multi-modal large language model for referring expression segmentation. In *ECCV*, 2024. 2

[4] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *CVPR*, 2024. 2

[5] Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. Chatrex: Taming multimodal llm for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024. 2, 3

[6] Weitai Kang, Gaowen Liu, Mubarak Shah, and Yan Yan. Segvg: Transferring object bounding box to segmentation for visual grounding, 2024. 2

[7] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *CVPR*, 2024. 2

[8] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2

[9] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 2

[10] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023. 2

[11] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 2

[12] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 2

[13] Zhuoyan Luo, Yinghao Wu, Yong Liu, Yicheng Xiao, Xiao-Ping Zhang, and Yujiu Yang. HDC: Hierarchical semantic decoding with counting assistance for generalized referring expression segmentation. *arXiv preprint arXiv:2405.15658*, 2024. 2

[14] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *CVPR*, 2024. 2

[15] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[16] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[17] Qwen Team. Qwen3 technical report, 2025. 3

[18] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795, 2024. 2, 3

[19] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized segmentation via multimodal large language models. In *CVPR*, 2024. 2

[20] Jinjin Xu, Liwu Xu, Yuzhe Yang, Xiang Li, Fanyi Wang, Yanchun Xie, Yi-Jie Huang, and Yaqian Li. u-LLaVA: Unifying multi-modal tasks via large language model. In *ECAI*, 2024. 2

[21] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 2

[22] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H.S. Torr. LAVT: Language-aware

vision transformer for referring image segmentation. In *CVPR*, 2022. 2

[23] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, 2024. 2

[24] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. PSALM: Pixelwise segmentation with large multi-modal model. In *ECCV*, 2024. 2