GroundingSuite: Measuring Complex Multi-Granular Pixel Grounding

Rui Hu[1*] Lianghui Zhu[1*] Yuxuan Zhang[1] Tianheng Cheng[1†] Lei Liu[2] Heng Liu[2]
Longjin Ran[2] Xiaoxin Chen[2] Wenyu Liu[1] Xinggang Wang[1‡]
[1] School of EIC, Huazhong University of Science & Technology
[2] vivo AI Lab

Figure 1. **Examples of our GroundingSuite dataset.** Including four aspects: stuff-class segmentation for context-aware localization, part-level segmentation requiring fine-grained understanding, multi-object segmentation with complex referential relationships, and single-object segmentation across diverse appearance variations. Note that 'red' color means the referring is wrong.

Abstract

Pixel grounding, encompassing tasks such as Referring Expression Segmentation (RES), has garnered considerable attention due to its potential for bridging the gap between vision and language modalities. However, advancements in this domain are currently constrained by limitations inherent in existing datasets, including limited object categories, insufficient textual diversity, and a scarcity of high-quality annotations. To mitigate these limitations, we introduce ***GroundingSuite****, which comprises: (1) an automated data annotation framework leveraging multiple Vision-Language Model (VLM) agents; (2) a large-scale training dataset encompassing 9.56 million diverse referring expressions and their corresponding segmentations; and (3) a meticulously curated evaluation benchmark consisting of 3,800 images. The GroundingSuite dataset boosts model performance to state-of-the-art levels. Specifically, a cIoU of 68.9 on gRefCOCO and a gIoU of 55.3 on RefCOCOm. Moreover, the GroundingSuite annotation framework demonstrates superior efficiency compared to the current leading data annotation method, i.e.,* $4.5\times$ *faster than the GLaMM. Codes are available at:* https://github.com/hustvl/GroundingSuite.

1. Introduction

Recent advancements in pixel grounding have fostered significant interest due to their remarkable segmentation performance in following natural language descriptions. However, this task remains constrained by existing benchmark limitations. As shown in Tab. 1, widely-used datasets like RefCOCO series [8, 17, 19, 36], while valuable for early research, restrict their scope to closed-set categories and limited manually annotated data samples. Specifically, only 80 object categories from the COCO dataset [13] hinder the pixel grounding task from generalizing to open-vocabulary understanding, diverse granularity levels (*e.g.*, part-level segmentation), and complex scene compositions (*e.g.*, multi-object interactions and background separation). Besides, current datasets can not satisfy both the amount and annotation quality requirements, *i.e.*, manually annotated datasets can not scale up due to the heavy human effort, and automatically annotated datasets often show inferior label quality.

Some automatic annotation approaches, such as GLaMM [22] and MRES [28], are proposed to alleviate the heavy human annotation burden but also face low-quality

* Equal contribution; † Project lead; ‡ Corresponding author: xgwang@hust.edu.cn

annotation and high-cost utilization problems. From the perspective of annotation quality, GLaMM suffers from unresolved textual ambiguities, and MRES only works with restricted fixed vocabularies. As for the cost, GLaMM has 23 pipeline steps, requiring a large amount of GPU resources, and MRES relies on human-annotated boxes as initialization.

To address the above challenges, we propose GSSculpt, an annotation framework that involves vision-language models (VLM) as efficient annotation agents and effective quality checkers. Specifically, GSSculpt incorporates three critical components: entity spatial localization, grounding text generation, and noise filtering. Entity spatial localization is the first phase, in which we generate comprehensive image captions, employ precise phrase grounding techniques, and utilize SAM2 [23] to extract high-quality masks. Then, we use state-of-the-art multimodal models with carefully designed prompts to generate unambiguous descriptions with clear positional relationships in the grounding text generation phase. Last, we employ instruction-based segmentation models to filter the noisy references, ensuring high data quality throughout the collection.

Based on the GSSculpt, we further curated a large-scale training set, *i.e*., GSTrain-10M, and a comprehensive evaluation benchmark, *i.e*., GSEval. The above three components make up the proposed **GroundingSuite**. Specifically, GSTrain-10M comprises a 9.56-million-image training corpus automatically annotated on SA-1B dataset images [9] using our hybrid framework, featuring diverse textual descriptions averaging 16 words in length and unambiguous instruction-mask pairs. GSEval is a novel evaluation benchmark containing 3,800 images carefully selected from COCO's unlabeled datasets [13], ensuring zero overlap with existing annotated sets while maintaining natural scene diversity. As shown in Fig. 1, the proposed benchmark specifically covers four mainstream aspects of segmentation: stuff-class segmentation [3, 39, 40] for context-aware localization, part-level segmentation [4, 6, 12] of fine-grained understanding, multi-object segmentation [14] with complex referential relationships, and single-object segmentation [8, 17, 19, 36] across diverse appearance variations.

Our main contributions of GroundingSuite can be summarized as follows:

- To address the low-quality annotation and high-cost utilization problems of existing auto-labeling methods, we propose GSSculpt, a vision-language models (VLM) based automatic annotation framework, which produces accurate annotations and reduces 78% of pipeline steps when compared to GLaMM.
- We introduce a large-scale training dataset and a carefully curated evaluation benchmark, laying a solid foundation for future research in pixel grounding. This dataset addresses critical limitations in existing grounding datasets, *i.e*., limited object categories, insufficient textual diversity, and a scarcity of high-quality annotations, while supporting diverse segmentation scenarios.
- The proposed dataset presents substantial performance enhancements across baseline methods. Models trained on our data consistently demonstrate superior results, establishing new state-of-the-art benchmarks across multiple evaluation metrics. Specifically, our approach achieves a cIoU of 68.9 on gRefCOCO and a gIoU of 55.3 on RefCOCOm.

Benchmarks	Cat.	Len.	Stuff	Part	Multi	Single
RefCOCO [8, 36]	80	3.6				✓
RefCOCO+ [8, 36]	80	3.5				✓
RefCOCOg [17, 19]	80	8.4				✓
gRefCOCO [14]	80	3.7			✓	✓
RefCOCOm [28]	471	5.1		✓		✓
GSEval	∞	**16.1**	✓	✓	✓	✓

Table 1. **Comparisons with previous Referring Expression Segmentation benchmark.** 'Cat.' and 'Len.' denote the number of categories and average text length; Stuff: includes stuff classes; Part: includes part-level annotations; Multi: supports multi-object references; Single: supports single-object references.

2. Related Work

2.1. Automatic Annotation for Pixel Grounding

CLIP [20] and GLIP [11] pioneered using vision-language models for label generation, while SAM [9] enabled promptable segmentation at scale. Building on these, recent work explores automatic dataset creation through model collaboration. GranD [22] is Developed using an automatic annotation pipeline and verification criteria, it encompasses 7.5M unique concepts grounded in 810M regions. However, the GLaMM [22] process has many steps, leading to the accumulation of errors from multiple models, and as a result, it hasn't solved the problem of text ambiguity. MRES [28] build a large visual grounding dataset namely MRES32M, which comprises over 32.2M high-quality masks and captions on the provided 1M images. However, it relies on box annotations from existing datasets and only uses a fixed vocabulary, which limits its generality.

2.2. Pixel Grounding Benchmarks

Mainstream benchmarks such as RefCOCO [8, 36], RefCOCO+ [8, 36], and RefCOCOg [17, 19] have driven progress in language-guided segmentation. However, their reliance on COCO's categorical constraints (80 classes) and human-annotated referring expressions creates artificial domain limitations. Among recent works, RefCOCOm [28] contains 80 object categories and an associated 391 part

categories. Despite its part-level advancements, RefCOCOm's reliance on COCO's fixed 80 object categories limits its ability to benchmark open-vocabulary or cross-category referring segmentation, hindering evaluation of models' adaptability to unseen object types in real-world scenarios. GRES [14] is a new benchmark called Generalized Referring Expression Segmentation, which extends the classic RES to allow expressions to refer to an arbitrary number of target objects. The GRES benchmark has several limitations, including its restriction to the 80 object categories from the COCO dataset, the lack of part-level segmentation, and the absence of a background class for segmentation.

3. GSSculpt: Large-scale Grounding Labeling

3.1. Overview

We introduce our vision-language models (VLM) based automatic annotation framework **GSSculpt**, designed to automatically generate high-quality pixel grounding data at scale. Through a comprehensive analysis of existing approaches, we propose three critical components essential for the effective auto-labeling framework, as shown in Fig. 2, *i.e.*, (1) **entity spatial localization**: discovering regions/objects of interest and generating high-quality masks; (2) **grounding text generation**: generating precise and uniquely identifiable language descriptions for regions or objects; and (3) **noise filtering**: eliminating ambiguous or low-quality samples. The proposed framework GSSculpt provides elaborate designs for each component to ensure annotation quality and accuracy, collectively constructing an efficient streamlined annotation framework, which aims to advance the state of pixel-level grounding datasets.

Data source. For large-scale training data, we primarily utilize SA-1B [9] as our data source, which comprises 1.1 billion high-quality segmentation masks across 11 million diverse, high-resolution images. In this paper, we sample 2M images for annotation. While the segmentation annotations in SA-1B focus on diverse geometric prompts, we concentrate on semantically meaningful objects or regions. Therefore, we employed SAM-2 [23] for segmentation annotation instead of directly using the mask annotations provided by SA-1B.

3.2. Entity Spatial Localization

The proposed framework builds upon precise object/region recognition and localization within complex visual scenes. This critical first stage employs a sophisticated three-tiered approach that systematically addresses the challenges of visual parsing.

Global caption generation. We leverage a cutting-edge large visual-language model, *i.e.*, InternVL2.5 [5], to produce comprehensive scene descriptions, discovering all semantically significant objects or regions within each image with remarkable completeness.

Phrase grounding. The generated captions serve as input for Florence-2 [32], as the advanced phrase grounding model, which precisely grounds texts to corresponding spatial locations. This process provides preliminary bounding box regions for candidate objects.

Mask generation. Then we adopt SAM2 [23] to obtain pixel-level segmentation masks for the grounded texts using bounding boxes as spatial prompts.

3.3. Grounding Text Generation

The second stage primarily further optimizes the textual descriptions of grounded regions with Large Language Models, enriching the grounding texts with the context of the image.

We design specialized prompt templates to guide multimodal language models in generating distinct and unambiguous references. These templates strategically emphasize spatial relationships, distinctive visual features, and contextual cues. Using powerful models like InternVL2.5, we generate linguistically diverse and rich descriptions. Our method produces natural descriptions averaging 16 words, significantly more expressive than the shorter phrases in manually annotated datasets, while maintaining referential clarity and eliminating ambiguity.

3.4. Noise Filtering

The noise filtering stage ensures dataset quality by eliminating ambiguous or incorrect annotations:

We employ instruction-based segmentation models, *i.e.*, EVF-SAM [37], to identify potentially ambiguous referring expressions by measuring consistency between the generated expression and the corresponding mask. Specifically, we use the generated texts in previous steps to prompt a Referring Expression Segmentation (RES) model to produce masks, and then calculate the IoU between the generated masks and the annotated mask. By applying an IoU threshold, *i.e.*, 0.5, we can effectively filter out inaccurate text-mask pairs.

This comprehensive filtering approach achieves an optimal balance between dataset scale and annotation quality, resulting in 9.56 million high-quality training samples while ensuring the integrity and quality of the final dataset.

3.5. GSTrain-10M

Applying the proposed annotation framework to a diverse subset of images from the SA-1B dataset, we have created **GSTrain-10M**, a large-scale comprehensive training dataset, aiming for pixel grounding and comprising 9.56 million high-quality text-mask pairs across 2 million images. GSTrain-10M contains linguistically rich descriptions with an average length of 16 words, significantly more

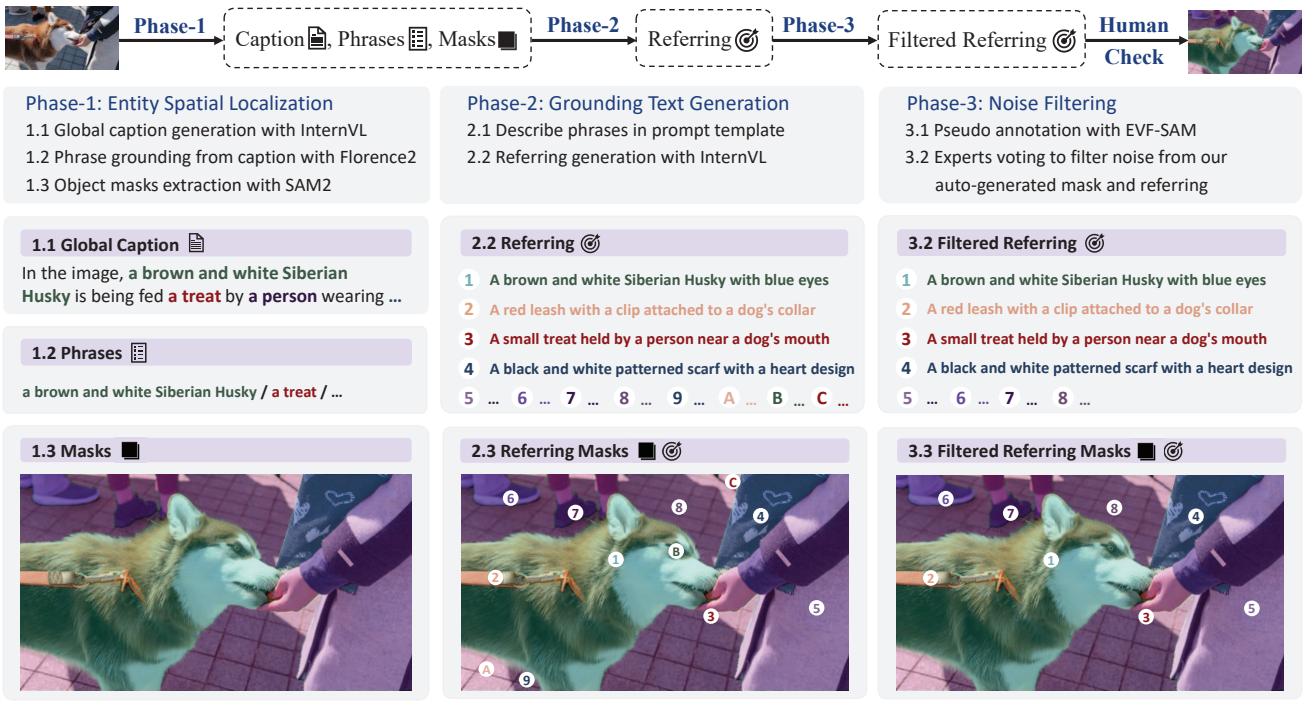

Figure 2. **GSScuplt Automatic Annotation Framework.** Our pipeline consists of three sequential phases: (1) Entity Spatial Localization, where we first identify potential objects of interest and generate high-quality segmentation masks; (2) Grounding Text Generation, where we then create unambiguous natural language descriptions that uniquely reference the segmented objects; and (3) Noise Filtering, where we finally eliminate ambiguous or low-quality samples to ensure dataset reliability.

detailed than existing manually annotated datasets. The GSTrain-10M dataset covers an open vocabulary of concepts including common objects, fine-grained parts, amorphous stuff categories, and diverse scenes. Each annotation undergoes a rigorous noise-filtering process, ensuring unambiguous references with clear spatial relationships. The presented GSTrain-10M represents a substantial advancement in both scale and quality for pixel grounding, enabling more robust and generalizable model training across a wider range of visual concepts and linguistic expressions.

4. GSEval: A Comprehensive Benchmark

Building upon the aforementioned annotation framework, we further develop **GSEval**, a comprehensive evaluation benchmark for pixel grounding tasks. As shown in Fig. 3, we propose a human-guided curation pipeline to ensure data quality and task diversity.

4.1. Automatic Data Curation

We first apply the proposed annotation framework (Sec. 3) to the unlabeled images of the COCO dataset [13]. The generated annotations and labeled images have no overlap with existing labeled datasets. Then, we use vision-language models (VLM) to assign categories for referring prompts. This pre-categorization helps to reduce the cost of subsequent manual selection and filters out noisy referring-mask pairs. Last, we adopt matting methods to refine the boundaries of object masks. Specifically, we translate the mask areas to trimaps and use the *off-the-shelf* matting model [34] to generate precise boundaries.

4.2. Human Data Curation

To ensure the quality of GSEval, every image was manually selected and verified by human reviewers. This process resulted in a dataset organized into four distinct categories:

- *Stuff Segmentation*: Comprises 1,000 images of background elements that require context-aware localization, such as `sky` and `sea`.
- *Part-level Segmentation*: Consists of 500 images that demand fine-grained understanding of object components, such as `the camera on a phone` or `a man's beard`.
- *Multi-object Segmentation*: Includes 800 images with complex referential relationships between entities, like `a flock of sheep` or `two dogs`.
- *Single-object Segmentation*: Contains 1,500 images that showcase diverse object appearances, such as `a brown cat` and `a colorful parrot`.

This enhanced human-in-the-loop approach significantly improves efficiency while maintaining strict quality standards. The VLM pre-classification reduces the human reviewers' annotation burden, allowing them to focus on

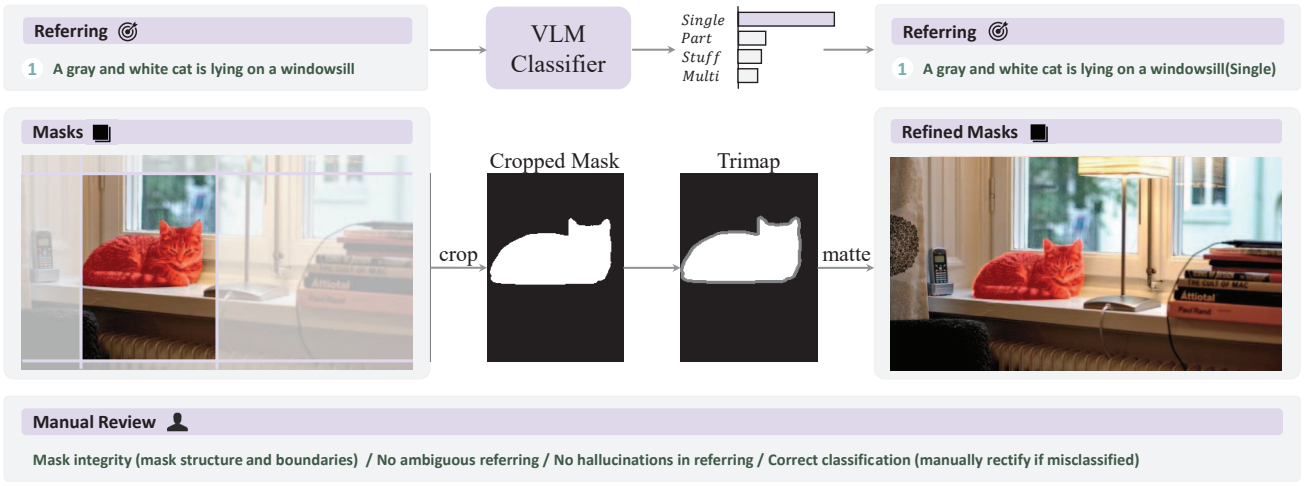

Figure 3. **Curation pipeline for GSEval Benchmark.** First, we apply our annotation pipeline to unlabeled COCO images. Next, we use a VLM classifier to ensure the categories of referring prompts. Then, we translate the coarse masks to trimaps and apply matting methods for precise object boundaries. Finally, we organize human reviewers for manual checks.

nuanced verification. By strategically distributing images across the four categories, we ensure comprehensive evaluation across varying levels of granularity and complexity. The curated benchmark covers multiple domains and object types, providing a robust assessment of models' generalized segmentation capabilities in real-world scenarios.

4.3. GSEval-BBox

In addition to the pixel-based segmentation benchmark, we construct a bounding box version, namely **GSEval-BBox**. Specifically, GSEval-BBox is designed to evaluate the visual grounding capabilities of multimodal large language models. GSEval-BBox converts the high-quality segmentation masks into corresponding bounding boxes, enabling direct assessment of referential object localization.

4.4. Evaluation Metrics

Pixel-level grounding. We adopt the standard metric for pixel level evaluation, *i.e.*, gIoU. The gIoU metric calculates the average of per-image Intersection-over-Union (IoU) scores:

$$\text{gIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i|}{|P_i \cup G_i|},$$

where P_i and G_i denote the predicted segmentation mask and ground truth mask for the i-th sample, respectively. N is the total number of samples. In contrast to previous methods using cIoU, the proposed GSEval includes part-level segmentation tasks and we prioritize gIoU as the primary evaluation metric. Specifically, gIoU provides a more balanced assessment across objects of varying sizes, while cIoU exhibits bias toward larger objects and demonstrates greater volatility in measurements.

Box-level grounding. For box-level grounding, we adopt the same metrics as in referring expression comprehension (REC), which typically includes accuracy at different IoU thresholds (e.g., Acc@0.5).

5. Evaluation on GSEval

In this section, we evaluate several representative methods on our proposed GSEval under zero-shot settings.

5.1. Benchmark Details

As shown in Tab. 1, GSEval significantly surpasses existing referring expression segmentation benchmarks in multiple dimensions. Compared to the RefCOCO dataset, which primarily relies on COCO category annotations, our proposed GSEval adopts an open-vocabulary setting, encompassing not only foreground objects but also stuff categories and various part-level categories. In addition, the average text length in GSEval reaches 16.1 words, making its descriptions substantially more detailed and nuanced compared to others. Furthermore, GSEval is uniquely comprehensive in supporting all key features: stuff classes, part-level annotations, and both multi-object and single-object references. This comprehensive design enables more challenging and realistic evaluation scenarios that better reflect real-world language grounding applications.

5.2. Benchmark Results on GSEval

Tab. 2 presents the zero-shot performance of previous state-of-the-art RES methods on our GSEval across four challenging subsets, *i.e.*, stuff, part, multi-object and single-object, which provides a comprehensive assessment of model capabilities. For comparison, we also report the performance on RefCOCO/+/g benchmarks (averaged across val/testA/testB splits) using compound IoU (cIoU).

Limited generalization ability of specialists. As shown in Tab. 2, although the specialists (*e.g.*, LAVT [33] and ReLA [14]) achieved good results on RefCOCO-series benchmarks, the performance on GSEval is poor, especially for stuff and part cases, which were not a major focus in the previous training and evaluation data. Compared to MLLM-based methods (multimodal large language models), it is evident that those specialists have weaker generalization capabilities. In addition, the dramatic performance drop highlights the limitations of existing benchmarks and validates the need for more comprehensive evaluation benchmarks like GSEval.

Limited fine-grained localization ability of MLLMs. As shown in Tab. 2, although some MLLM-based methods have achieved good performance on GSEval, especially for the stuff category, their performance on fine-grained part-level pixel grounding is clearly inferior to other subsets. While LLMs possess strong text understanding and reasoning capabilities, they still have significant limitations in fine-grained image localization.

Notably, we observe a significant discrepancy between the GSEval and RefCOCOs scores, such as InstructSeg [29] and LISA [10]. InstructSeg achieves state-of-the-art performance on RefCOCOs (81.9 cIoU) but performs poorly on GSEval (52.5 gIoU). In contrast, LISA has relatively low scores on RefCOCOs (67.9 cIoU) but achieves decent accuracy on GSEval (57.6 gIoU). We further analyze the training data of both methods and find that LISA incorporates a wide variety of datasets from different tasks including ADE20K [39, 40] and COCO-Stuff [3], whereas InstructSeg was more focused on referring expression segmentation. This makes LISA more generalizable compared to InstructSeg and further validates that our proposed GSEval is better suited than RefCOCOs for evaluating the performance of general-purpose multi-task pixel grounding methods.

Moreover, we provide the visualization results on GSEval. As shown in the Fig. 4, InstructSeg performs poorly on stuff and part categories, tending to segment entire objects instead. In contrast, LISA and EVF-SAM are better at following instructions to locate the target regions, but their segmentation accuracy still has some limitations.

5.3. Benchmark Results on GSEval-BBox

Additionally, we further evaluated the bbox-level grounding performance of multimodal large language models, as shown in Tab. 3. Gemini-1.5-Pro [27] demonstrates exceptional performance on GSEval, particularly in part-level grounding. However, for open-source models, part-level grounding remains a challenging subset. As mentioned earlier, fine-grained object localization is relatively difficult for multimodal large language models.

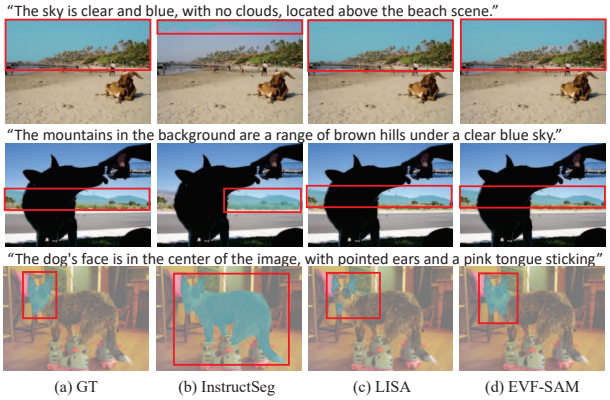

Figure 4. **The visualization comparisons of different methods on GSEval.** All methods are evaluated under the zero-shot setting with the public code and weights.

In Fig. 5, we further visualize the results of different multimodal large language models. It is clear that Qwen2.5VL [2] and InternVL2.5 [5] struggle to locate part-level targets. Even when the targets are located, the bounding box precision is relatively low, making it difficult to achieve accurate object localization. In contrast, Gemini-1.5-Pro currently performs better, but there is still a significant gap in overall performance.

Since RefCOCO is more focused on foreground object-level grounding, with the development of multimodal large language models, it has become increasingly difficult to use RefCOCO as the primary benchmark for evaluating the performance of multimodal models. A comprehensive, multi-granularity grounding benchmark is therefore greatly needed. As a result, we believe that GSEval plays a crucial role in the grounding evaluation of multimodal models.

6. Experiments

In this section, we conduct a series of ablation studies to further validate the effectiveness of the automatically annotated training dataset, GSTrain-10M. We describe the details of each experiment in the corresponding subsection.

6.1. Effectiveness on Pixel Grounding Methods

To evaluate the impact of our GSTrain-10M, we select two representative methods, *i.e.*, EVF-SAM [37] and LISA-7B [10], and test them with and without our GSTrain-10M.

Experimental details. For the baseline, we followed the original settings of EVF-SAM and LISA-7B. Specifically, EVF-SAM utilizes multiple datasets, including the RefCOCO series [8, 17, 19, 36], Objects365 [24], ADE20K [39, 40], Pascal-Part [4], HumanParsing [12], and PartImageNet [6].

Meanwhile, LISA-7B employs the RefCOCO series [8, 17, 19, 36], ADE20K [39, 40], COCO-Stuff [3], PACO-

Methods	Type	RefCOCO/+/g (AVG)	GSEval				
			Stuff	Part	Multi	Single	All
LAVT [33]	*Specialist*	65.8	6.0	10.7	45.0	25.8	22.5
ReLA [14]		68.3	3.8	8.8	46.7	19.7	19.7
LISA-7B [10]	*MLLM*	67.9	85.2	21.2	71.5	42.8	57.6
GSVA-7B [31]		70.1	76.0	20.0	57.8	34.2	48.6
GLaMM [22]		75.6	**86.9**	16.5	70.4	42.1	57.2
PSALM [38]		77.1	39.0	10.0	53.7	36.9	37.7
EVF-SAM [37]		79.3	85.1	23.1	**72.1**	**54.5**	**62.6**
InstructSeg [29]		**81.9**	56.2	**24.2**	66.8	51.3	52.5

Table 2. Comparison among previous SOTA RES methods on our GSEval in terms of gIoU, while we report average cIoU for RefCOCO/+/g.

Methods	Type	RefCOCO/+/g (AVG)	GSEval-BBox				
			Stuff	Part	Multi	Single	All
Gemini-1.5-Pro [27]	*Proprietary Model*	-	86.7	58.5	61.7	**77.3**	74.3
Doubao-1.5-thinking-vision-pro[25]		91.3	86.8	**69.7**	83.1	74.3	**79.0**
Doubao-1.5-vision-pro [26]		91.6	80.0	28.8	**85.2**	53.3	64.2
Claude-3.7-sonnet [1]		-	56.7	2.6	20.7	9.4	23.8
GPT-4o [7]		-	42.2	2.0	15.3	6.1	17.3
InternVL3-78B[41]	*Open-source Model*	91.4	69.3	13.8	61.1	49.5	52.9
InternVL3-8B[41]		89.6	77.3	8.4	56.6	46.5	52.3
InternVL2.5-78B [5]		92.3	85.3	16.8	63.2	55.7	62.2
InternVL2.5-8B [5]		87.6	91.3	7.3	65.7	47.3	58.2
Qwen2.5-VL-72B [2]		90.3	88.4	31.2	42.8	64.6	62.5
Qwen2.5-VL-7B [2]		86.6	**93.0**	17.6	75.9	59.1	66.7
DeepSeek-VL2 [30]		**93.0**	86.5	12.7	64.8	51.2	60.8
Ferret-13B [35]		85.6	80.6	21.3	58.0	46.6	55.1
Ferret-7B [35]		83.9	82.1	17.6	56.0	43.2	53.3
Mistral-Small-3.1-24B [18]		-	16.0	3.5	20.3	10.7	13.2

Table 3. Comparison among previous grounding methods on our GSEval-BBox. All metrics measure referring expression comprehension accuracy (%).

Methods	GSTrain-10M	GSEval	gRefCOCO	RefCOCOm
EVF-SAM		62.6	63.5	51.3
EVF-SAM	✓	77.3 +14.7	66.4 +2.9	55.3 +4.0
LISA-7B		57.6	32.2	34.2
LISA-7B	✓	73.6 +16.0	36.3 +4.1	39.3 +5.1

Table 4. Ablation study on the effectiveness of our dataset on different methods.

LVIS [21], PartImageNet [6], Pascal-Part [4], as well as LLaVAInstruct-150k for LLaVA-v1 [15], LLaVA-v1.5-mix665k for LLaVA v1.5 [16], and ReasonSeg [10].

Results. In Tab. 4, we evaluate the performance using the proposed GSTrain-10M of EVF-SAM and LISA on several benchmarks. The results demonstrate that our dataset brings significant performance improvements to both methods across all benchmarks.

For EVF-SAM, we observe 14.7, 2.9, and 4.0 points improvements on GSEval, gRefCOCO, and RefCOCOm benchmarks, respectively. The consistent gains across different evaluation benchmarks highlight the transferability of knowledge acquired from our diverse training corpus.

For LISA-7B, we observe 16.0, 4.1, and 5.1 points improvements on GSEval, gRefCOCO, and RefCOCOm benchmarks, respectively. The greater relative improvement observed in LISA-7B suggests that language-centric models particularly benefit from our dataset's text diversity and the richness of our referring expressions (averaging 16 words in length).

Furthermore, the consistent improvements across both external benchmarks (gRefCOCO and RefCOCOm) demonstrate that the knowledge gained from our dataset generalizes well beyond the specific distribution of GSEval. Considering the gRefCOCO focuses on multi-object relationships, while RefCOCOm evaluates fine-grained semantic understanding, the performance gains across these diverse evaluation scenarios demonstrate the robustness and generality of the proposed GSTrain-10M.

(a) GT (b) Qwen2.5VL (c) InternVL2.5 (d) Gemini1.5 Pro

Figure 5. **The visualization comparisons of differnet methods on our GSEval-BBox.** All open-source methods are evaluated under the zero-shot setting with the public code and weights.

These results collectively validate that the proposed GSSculpt produces high-quality training data that enhances model performance across architectural paradigms, supporting more accurate and nuanced referring expression segmentation capabilities.

6.2. Comparisons with Other Automated Datasets

To further validate our approach, we conduct a comparative analysis between the proposed GSTrain-10M and GranD, another state-of-the-art automatically generated dataset introduced with GLaMM [22].

Experimental details. For fair comparisons, we randomly selected 100k samples from both GranD and the proposed GSTrain-10M to train models under the LAVT, EVF-SAM and LISA framework.

Results. As evidenced in Tab. 5, our proposed GSTrain dataset consistently outperforms the GranD dataset across all evaluated models and benchmarks. This benefit is clear for the specialist model, LAVT, which sees performance boosts of 2.2, 1.1, and 2.0 percentage points on RefCOCO, gRefCOCO, and RefCOCOm, respectively. The advantages of GSTrain are even more pronounced for the MLLM-based models. EVF-SAM, for instance, achieves gains of 2.8, 0.9, and 1.2 points across the same benchmarks. The most significant improvement is observed with LISA-7B, which not only improves by 2.5 points on RefCOCO but also exhibits a substantial 8.0-point increase on RefCOCOm (37.6 vs. 29.6). This considerable performance gain on a benchmark known for its linguistic complexity underscores that GSTrain provides the rich, high-quality data crucial for enabling advanced models to better interpret fine-grained textual descriptions.

Methods	Dataset	RefCOCO	gRefCOCO	RefCOCOm
LAVT	GranD	39.5	26.3	27.1
LAVT	GSTrain	41.7 +1.2	27.4 +1.1	29.1 +2.0
EVF-SAM	GranD	60.7	28.0	31.5
EVF-SAM	GSTrain	63.5 +2.8	28.9 +0.9	32.7 +1.2
LISA-7B	GranD	63.1	29.5	29.6
LISA-7B	GSTrain	65.6 +2.5	31.3 +1.8	37.6 +8.0

Table 5. Ablation study on the effectiveness of different automated datasets.

Ratio	GSEval				
	Stuff	Part	Multi	Single	All
0%	85.1	23.1	72.1	54.5	62.6
20%	93.2	43.2	85.0	68.2	75.4
50%	94.6	54.0	88.3	72.7	79.6
100%	94.7	59.2	89.0	72.4	80.3

Table 6. Ablations on the scalability.

6.3. Scalability

To investigate the effect of data scalability on model performance, we train EVF-SAM with different proportions of our GSTrain-10M without other datasets.

Experimental details. We follow the training settings from [37] and train EVF-SAM with different proportions of our GSTrain-10M (0%, 20%, 50%, 100%) for 10 epochs and then evaluate it on the proposed GSEval.

Results. As shown in Tab. 6, the results show that model performance increases significantly across all aspects as the data proportion increases.Notably, we can observe a clear rising curve with the increment of training data ratio.

7. Conclusion

In this paper, we present GroundingSuite, which contains a vision-language model (VLM) based automatic annotation framework, a large-scale, textually diverse training corpus (9.56M masks), and a comprehensive evaluation framework. Extensive experiments demonstrate that models trained on GSTrain-10M consistently establish new state-of-the-art results across multiple benchmarks. The proposed GroundingSuite lays a solid foundation for the visual language understanding domain, providing valuable support for future research endeavors.

Acknowledgement: This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62276108.

References

[1] Anthropic. Claude 3.7 sonnet system card. https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf, 2024. 7

[2] 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. 6, 7

[3] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *CoRR*, abs/1612.03716, 2016. 2, 6

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014. 2, 6, 7

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 6, 7

[6] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 2, 6, 7

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[8] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1, 2, 6

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[10] 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. 6, 7

[11] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[12] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, pages 1386–1394, 2015. 2, 6

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2, 4

[14] 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, 3, 6, 7

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 7

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 7

[17] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 2, 6

[18] mistralai. Mistral-small-3.1-24b-instruct-2503. https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503, 2024. 7

[19] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 1, 2, 6

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[21] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 7

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 1, 2, 7, 8

[23] 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

[24] 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. 6

[25] ByteDance Seed Team. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 7

[26] Doubao Team. Doubao-1.5-pro. https://team.doubao.com/en/special/doubao_1_5_pro, 2024. 7

[27] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6, 7

[28] Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12998–13008, 2024. 1, 2

[29] Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Zheng Zhao, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. *arXiv preprint arXiv:2412.14006*, 2024. 6, 7

[30] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 7

[31] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 7

[32] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 4818–4829. IEEE, 2024. 3

[33] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022. 6, 7

[34] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 4

[35] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 7

[36] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 1, 2, 6

[37] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 3, 6, 7, 8

[38] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 7

[39] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5122–5130. 2, 6

[40] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 2, 6

[41] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 7