SimVG: A Simple Framework for Visual Grounding with Decoupled Multi-modal Fusion

Ming Dai[1], Lingfeng Yang[2], Yihao Xu[1], Zhenhua Feng[3], Wankou Yang[1,4]*

[1]Southeast University   [2]Nanjing University of Science and Technology
[3]Jiangnan University   [4]Advanced Ocean Institute of Southeast Univerisity, Nantong
{mingdai, 220211848, wkyang}@seu.edu.cn,
yanglfnjust@njust.edu.cn, fengzhenhua@jiangnan.edu.cn

Abstract

Visual grounding is a common vision task that involves grounding descriptive sentences to the corresponding regions of an image. Most existing methods use independent image-text encoding and apply complex hand-crafted modules or encoder-decoder architectures for modal interaction and query reasoning. However, their performance significantly drops when dealing with complex textual expressions. This is because the former paradigm only utilizes limited downstream data to fit the multi-modal feature fusion. Therefore, it is only effective when the textual expressions are relatively simple. In contrast, given the wide diversity of textual expressions and the uniqueness of downstream training data, the existing fusion module, which extracts multimodal content from a visual-linguistic context, has not been fully investigated. In this paper, we present a simple yet robust transformer-based framework, SimVG, for visual grounding. Specifically, we decouple visual-linguistic feature fusion from downstream tasks by leveraging existing multimodal pre-trained models and incorporating additional object tokens to facilitate deep integration of downstream and pre-training tasks. Furthermore, we design a dynamic weight-balance distillation method in the multi-branch synchronous learning process to enhance the representation capability of the simpler branch. This branch only consists of a lightweight MLP, which simplifies the structure and improves reasoning speed. Experiments on six widely used VG datasets, *i.e.*, RefCOCO/+/g, ReferIt, Flickr30K, and GRefCOCO, demonstrate the superiority of SimVG. Finally, the proposed method not only achieves improvements in efficiency and convergence speed but also attains new state-of-the-art performance on these benchmarks. Codes and models are available at https://github.com/Dmmm1997/SimVG.

1 Introduction

Visual grounding (VG) aims to predict the corresponding regions of an image through linguistic expressions. The task necessitates a comprehensive understanding of each modality, as well as the modeling of consistency between image context and text. Some benchmarks focus on addressing *phrase localization* [26, 50], which entails locating all objects mentioned in a sentence within an image. Another aspect emphasizes resolving *referring expression comprehension* (REC) [80, 48, 47], characterized by only one target corresponding to a sentence. Recently, a new type of *general referring expression comprehension* (GREC) [37, 17] task has emerged. GREC is similar to REC, but in which a sentence can have multiple targets or no target at all.
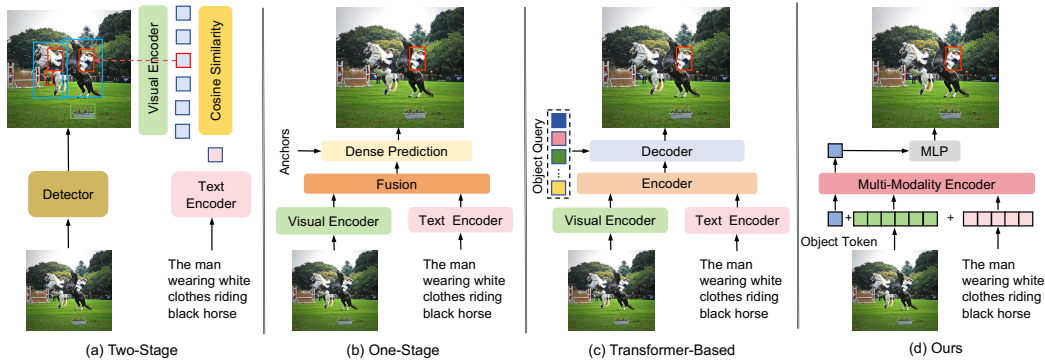
---

*Corresponding authors.

Figure 1: An overview of visual grounding structures: (a) Two-Stage: Applying a detector for proposals, followed by image-text encoding and feature similarity calculation for region matching. (b) One-Stage: Grounding in the fused features through dense prediction. (c) Transformer-based: Employing an encoder-decoder structure in the head. (d) Proposed SimVG: Utilizing Multi-Modality Encoder for multimodal interaction among object, image, and text tokens, directly applies a lightweight MLP for grounding.

Existing visual grounding models can be roughly divided into three categories: two-stage, one-stage, and transformer-based. Among them, as shown in Fig. 1(a), two-stage methods [19, 83, 79, 44, 38] require a pre-trained detector to generate proposals and perform localization through region-text retrieval. These methods rely on a complex module with manually designed mechanisms to achieve query reasoning and multi-modal fusion. One-stage methods [75, 36, 46, 73], on the other hand, employ an end-to-end architecture, as shown in Fig. 1(b). Most of them primarily perform dense prediction on multimodal fusion features defined in the form of anchors. Some recent algorithms [8, 25, 87, 57], depicted in Fig. 1(c), adopt an encoder-decoder architecture to perform multimodal fusion in the encoder and then decode the response target position using an object query similar to DETR[2]. The existing methods share a commonality: they adopt architectures that independently encode each modality before merging them, with multimodal fusion intricately linked to each visual grounding task. The feature extraction part of these methods generally employs specific classification [7, 55] or autoregressive [9, 30] tasks in each modality for pre-training. However, the alignment and mutual understanding between modalities only utilize a limited amount of downstream data, which undoubtedly underestimates the difficulty of achieving mutual understanding between modalities. Another observed trend, as noted in [25, 87], is the notable enhancement in the performance of visual grounding with a significant augmentation of pretraining data on large corpora. This implicitly suggests that leveraging a small amount of downstream data does not fully capitalize on the potential for mutual understanding between images and text. Nevertheless, this type of pretraining increases the burden of training resources.

Furthermore, the mutual understanding between multiple modalities is crucial for downstream tasks. As shown in Fig. 2, for a dataset with long sentence characteristics (like RefCOCOg [48]), adopting the decoupled multimodal understanding method can significantly improve model performance, while the improvement is relatively modest on datasets with short sentences. This observation aligns with our expectation that shorter captions pose less challenge for inter-modal understanding, whereas SimVG's decoupling approach proves advantageous for challenging longer descriptions that require intricate multi-modal comprehension.
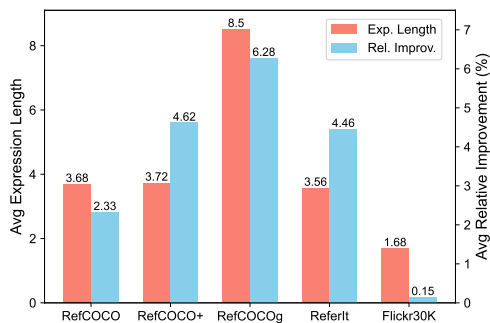

Figure 2: The expression length and relative improvement between Dynamic MDETR [57] and SimVG.

To be specific, we delve into existing multimodal understanding methods in the form of multimodal pre-training. Existing approaches can be broadly categorized into three categories. Dual-stream structures [51, 23, 33] encode image-text modalities independently and supervise them with contrastive learning. One-stream models [6, 27, 30, 22] concatenate multimodal features for feature extraction. Other works [31, 1, 58, 64] use a dual-stream design with a fusion encoder to balance complexity and computational cost, fusing multimodal features in intermediate layers. Some methods

also have applied multimodal pre-training to VG tasks, such as Dynamic MDETR [57], which uses an image-text encoder pretrained on CLIP [51] to enhance the performance. Recently, works like CPT [76], ReCLIP [59], and FGVP [71] have improved the performance of zero-shot visual grounding using dual-stream pre-trained models that employ a two-stage-like architecture and prompt engineering techniques. However, these approaches focus on multimodal alignment rather than mutual understanding. Thus, the dual-stream with fusion encoder architecture [31, 1, 58, 64] has been thrown into our sight. Specifically, based on BEiT-3 [64], we propose a simple framework called SimVG that decouples multimodal fusion from downstream tasks and simultaneously encodes image, text, and object tokens, as illustrated in Fig. 1 (d). The decoder structure used for query reasoning, similar to DETR [2], is effective but inevitably increases the model's complexity and computational overhead. We aim to develop a more efficient and simpler architecture for visual grounding. To achieve this, in addition to the decoder branch, we introduce a lightweight token branch. This branch leverages object tokens that are deeply integrated with image-text features to enhance grounding performance. To ensure that the token branch maintains high performance while enabling efficient inference, we adhere to the principles of the existing knowledge distillation methods [21, 5, 3] and introduce an innovative dynamic weight-balance distillation (DWBD). This approach enhances the token branch's capability by dynamically assigning weights to the decoder's predictions and ground truth at various learning stages, facilitating more effective learning in the token branch. Furthermore, we introduce a text-guided query generation (TQG) module to integrate textual information into queries, enabling the adaptive incorporation of textual prior knowledge into object queries. Notably, the design of the TQG module allows for the expansion of object tokens, increasing the number of objects it can handle and thereby adapting effectively to the GREC [17] task. The experiments demonstrate that, by decoupling multimodal fusion from downstream tasks, SimVG achieves rapid convergence and superior performance even with a limited amount of data. Coupled with our proposed DWBD and TQG modules, SimVG sets new state-of-the-art performance across various benchmarks.

Our main contributions are summarized as follows:

• We propose a simple yet strong visual grounding architecture called SimVG, which decouples multimodal understanding from downstream tasks and fully utilizes the inter-modal fusion ability of existing multimodal pre-trained models. To the best of our knowledge, SimVG is the first to employ a unified structure for encoding image, text, and object tokens in visual grounding.

• We propose a novel dynamic weight-balance distillation (DWBD) to dynamically allocate weights to decoder predictions and ground truth at different stages of multi-branch synchronous learning, aiming to minimize the discrepancy between the token and decoder branches. Furthermore, we introduce a text-guided query generation (TQG) module to incorporate textual prior information into object queries, thereby extending its applicability to the GREC task.

• The proposed SimVG architecture has demonstrated state-of-the-art performance across six prominent datasets, while also exhibiting notable gains in efficiency and convergence speed. Particularly noteworthy is that SimVG (ViT-B/32) achieves these results with just 12 hours of training on a single RTX 3090 GPU when applied to the RefCOCO/+/g datasets.

2 Related Work

Vision-Language Pre-Training.   Existing vision-language pretraining (VLP) models can be broadly categorized into three main types: one-stream, dual-stream, and dual-stream with fusion encoder architectures. One-stream models [6, 27, 30, 22] process both image and text inputs in a single stream, concatenate image and text embeddings, and interact cross-modals information throughout the whole feature extraction process. In contrast, dual-stream models [51, 23, 33] employ separate encoders for each modality. These models do not concatenate modalities at the input level. Instead, the interaction between pooled image and text vectors occurs at a shallow layer. Dual-stream models with fusion encoder [31, 1, 58, 64] combine aspects of both one-stream and dual-stream models. They allow for intermediate interaction between modalities, potentially offering a balance between complexity and performance. In this paper, we improve the performance of visual grounding by decoupling multi-modal fusion from downstream tasks into upstream VLP models [64].
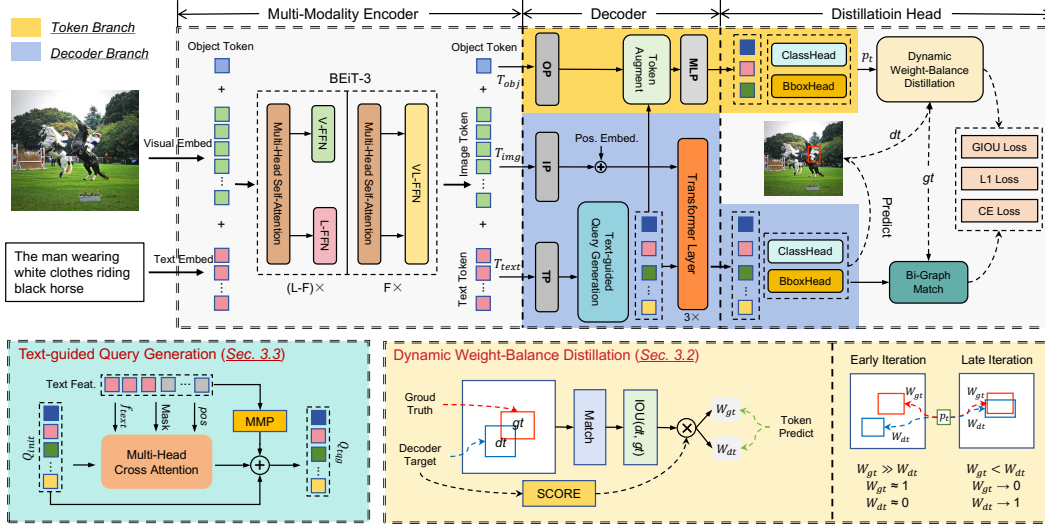
Figure 3: Overview of the proposed SimVG. The token branch refers to the upper light yellow region, while the decoder branch refers to the lower light blue region. During model inference, we can independently apply the more lightweight token branch to improve inference speed and simplify the model architecture.

Referring Expression Comprehension. Early approaches in REC typically followed a two-stage pipeline [19, 83, 88, 79, 44, 72, 18, 38]. This pipeline involves first extracting region proposals [53], which are then ranked based on their similarity scores with the language query. In contrast, a more recent line of research [85, 75, 36, 46, 73, 20] advocates for a simpler and faster one-stage pipeline based on dense anchors. Several recent approaches [8, 86, 25, 10, 32, 87] have employed a transformer-based structure [61] for multi-modal fusion. Furthermore, with the vigorous development of multimodal large language models (MLLM) [77], some of the latest methods have further enhanced the generalization performance of REC through zero-shot [71, 59] or fine-tuning [69, 49] methods, leveraging the powerful capabilities of large models [39] and general models [28]. In contrast to existing methods, our proposed SimVG method directly feeds object, image, and text tokens into the multi-modality encoder for multimodal feature interaction. We eschew the complex encoder-decoder structure and perform visual grounding directly using a simple MLP.

Knowledge Distillation in Object Detection. The majority of research in knowledge distillation has primarily focused on classification tasks [15, 66, 81]. Several studies [45, 65, 84, 63] have extended knowledge distillation techniques to dense prediction tasks, such as semantic segmentation and object detection. These works commonly exploit pixel-wise correlations or channel-wise interactions between dense features of teacher and student models. Recently, there has been a growing interest in developing tailored knowledge distillation losses for query-based detectors like DETR [2], as demonstrated by works such as [21, 5, 3]. Unlike previous methods that guide a lightweight student model using a pre-trained teacher model, this paper introduces knowledge distillation during synchronous learning to enhance the performance of the lightweight branch.

3 The Proposed Method

3.1 The Overview of SimVG

As shown in Fig. 3, the SimVG structure can be roughly divided into three parts: multi-modality encoder, decoder, and head. The multi-modality encoder adopts a structure similar to BEiT-3 [64], and additionally sets a learnable object token. The decoder is divided into two branches: one is similar to the transformer decoder in DETR [2] (*decoder branch*), and the other utilizes a lightweight MLP (*token branch*). The head is referred to as the "Distillation Head". Unlike conventional prediction heads, to reduce the performance gap between the token and decoder branches, we employ a dynamic weight-balance distillation (DWBD) to minimize the performance difference between the two branches during synchronous learning.

4

Multi-Modalilty Encoder. The input of SimVG include an image $I \in \mathbb{R}^{3 \times H \times W}$ and a caption text $T \in \Omega^M = \{t^1, t^2, ..., t^M\}$, where Ω denotes the set of words. First, the image is compressed to a scale of 1/32 of the original size using visual embedding to obtain $P_{img} = \{p^1, p^2, ..., p^N\}$. The text is then mapped to the vector space $L_{text} = \{l^1, l^2, ..., l^K\}$ and an text padding mask $L_{mask} = \{l_m^1, l_m^2, ..., l_m^K\}$. Additionally, we define a learnable object token T_{obj} as the target feature for the token branch. The query and attention padding mask of the transformer can be generated as:

$$\text{Query} = \{T_{obj}, p^1, ..., p^N, l^1, ..., l^K\}, \quad \text{Mask} = \{T_{obj_m}, p_m^1, ..., p_m^N, l_m^1, ..., l_m^K\}. \tag{1}$$

In the case where the image has no padding, T_{obj_m} and $\{p_m^1, p_m^2, ..., p^N\}$ are set to 0. In the independent encoding part, FFN adopts a setting of non-shared weights between the image and text modalities, while the rest remains largely the same with the original ViT [11] model.

The Decoder Branch. We first map the channel dimensions of the image tokens $T_{img} \in \mathbb{R}^{\frac{H}{32} \times \frac{H}{32} \times C}$ and text tokens $T_{text} \in \mathbb{R}^{N_{text} \times C}$ using a linear layer without sharing weights. Then, on the text side, we apply the text-guided query generation (TQG) module to interact the predefined object queries $Q_{obj} \in \mathbb{R}^{N_{oq} \times C}$ with the text tokens. On the image side, additional positional encoding is applied. Lastly, cross-attention interaction is performed through transformer modules. The entire process can be represented as follows:

$$Q_{decoder} = \mathbf{MCA}(\mathbf{IP}(T_{img}) + pos, \mathbf{TQG}(\mathbf{TP}(T_{text}), Q_{obj})), \tag{2}$$

where $\mathbf{TP}$ and $\mathbf{IP}$ refer to the text projection and image projection. $\mathbf{MCA}$ refers to multi-head cross attention. pos refers to position encoding, which applies 2D absolute sine encoding.

The Token Branch. We employ a linear layer $\mathbf{OP}$ to project object token $T_{obj} \in \mathbb{R}^{1 \times C}$ and use the results of TQG to augment the object token. Lastly, we use an MLP layer to further interact with and enrich the representation of the object token. The process of this branch can be defined as:

$$Q_{token} = \mathbf{MLP}(\mathbf{OP}(T_{obj}) + \mathbf{TQG}(\mathbf{TP}(T_{text}), Q_{obj})). \tag{3}$$

The Distillation Head. We adopt the same Hungarian matching as DETR [2] for the decoder branch. The matching cost consists of three parts: binary cross-entropy loss, l1 loss, and giou loss [54]. However, to further simplify the inference pipeline, we use the decoder branch as a teacher to guide the learning of the token branch during the whole training process. Therefore, the complete loss can be represented as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{det}(p_d, gt) + \mathcal{L}_{dwbd}, \quad \mathcal{L}_{det} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{l1} + \lambda_3 \mathcal{L}_{giou}, \tag{4}$$

where λ_1, λ_2, and λ_3 are set to 1, 5, and 2. p_d refers to the decoder branch prediction. $\mathcal{L}_{dwbd}$ refers to the dynamic weight-balance distillation loss, which will be explained in the next part.

3.2 Dynamic Weight-Balance Distillation

To make SimVG both efficient and effective, knowledge distillation is introduced, which leverages the predictions from the decoder branch as a teacher to guide the predictions of the token branch. Since the two branches share the features of the multi-modality encoder, training the teacher model independently using traditional knowledge distillation methods is not feasible. Instead, we employ a synchronous learning approach for both branches. This approach requires a delicate balance, ensuring that the performance of the teacher model is not compromised while maximizing the transfer of knowledge from the teacher branch to the student branch.

Therefore, we design a dynamic weight-balance distillation (DWBD), whose architecture is shown in Fig. 3. Let us denote the ground truth by y, and the set of N_q decoder prediction by $\hat{y}^d = \{\hat{y}_i^d\}_{i=1}^{N_q}$. To find a bipartite matching between these two sets, we search for a permutation of N_q elements $\sigma \in \Sigma_{N_q}$ with the lowest cost:

$$\hat{\sigma} = \mathbf{ARGMIN}_{\sigma \in \Sigma_{N_q}} \sum_{i}^{N_q} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}), \tag{5}$$

where $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is a pair-wise matching cost between the ground truth y_i and a prediction with index $\sigma(i)$. After pairing, the next step is to assess the decoder branch's understanding capability at the current stage. This is done by measuring the confidence of the joint target box of the decoder target dt and ground truth gt based on the IOU of their pairing:

$$W_{dt} = \frac{1}{N_{gt}} \sum_{i}^{N_{gt}} \left[\mathbf{IOU}(b_i, \hat{b}_{\hat{\sigma}}(i)) \times \mathbf{SCORE}(\hat{p}_{\hat{\sigma}(i)}) \right], \tag{6}$$

where N_{gt} is the number of ground truth boxes, SCORE represents the foreground score extracted from the predictions. W_{dt} can be seen as a reflection of the current stage's decoder branch capability, where a higher value indicates a stronger confidence. Lastly, $\mathcal{L}_{dwbd}$ can be expressed as follows:

$$\mathcal{L}_{dwbd} = \gamma_1(W_{dt} \times \mathcal{L}_{det}(p_t, dt)) + \gamma_2(W_{gt} \times \mathcal{L}_{det}(p_t, gt)), \quad W_{gt} = 1 - W_{dt}, \qquad (7)$$

where $\mathcal{L}_{det}$ is computed exactly the same as in Eq. 4. p_t refers to the token branch prediction. γ_1 and γ_2 are set to 2 and 1 in this paper. By design, in the early stage of network training, $W_{gt} \gg W_{dt}$, the training process of the entire token branch is guided by the ground truth. However, in the later training stage, $W_{gt} < W_{dt}$, the guidance from the decoder target becomes more significant. This dynamic adjustment of weights during training is the core idea of the proposed DWBD. We will further analyze the changes in W_{dt} and $\mathcal{L}_{dwbd}$ in Sec. 4.4.3. Additionally, to further enhance the performance of the token branch, we use a two-stage distillation approach. In the first stage, we train the decoder branch separately. In the second stage, we apply DWBD under the premise of synchronous learning for the two branches.

3.3 Text-guided Query Generation

The initial object queries, Q_{init}, are defined using learnable embeddings, without any prior information to guide them. From a macro perspective, visual grounding involves using text as a query to locate optimal regions in an image. Embedding text into queries to adaptively provide priors offers a viable solution. Therefore, we propose a text-guided query generation (TQG) module to generate object queries with text priors. As illustrated in Fig. 3, the process of generating queries through TQG can be expressed as follows:

$$Q_{tqg} = \mathbf{MCA}(Q_{init}, f_{text} + pos, \mathrm{Mask}) + \mathbf{MMP}(f_{text}, \mathrm{Mask}) + Q_{init}, \qquad (8)$$

where $f_{text} \in \mathbb{R}^{K \times C}$ is the feature after text projection, Mask is consistent with Eq. 1, and pos here is 1D absolute sine positional encoding. $\mathbf{MMP}$ is the process of filtering out valid tokens from f_{text} using the Mask and applying a max operation: $\mathbf{MMP}(f, m) = \max_i [f_i \times (\sim m_i)]$.

4 Experimental Results

4.1 Datasets and Evaluation Metric

The experiments in this paper are conducted on six widely used datasets: RefCOCO/+/g [80, 47, 48], Flickr30K Entities [50], ReferItGame [26], and GRefCOCO [17]. For the referring expression comprehension and phrase localization tasks, Precision@0.5 is used as the evaluation metric. The GRefCOCO dataset is evaluated using the Precision@(F_1=1, IoU$\geq$0.5) and N-acc metrics. More descriptions about datasets and evaluation metrics are provided in Appendix A and B.

4.2 Implementation Details

We train SimVG 30 epochs for REC and phrase localization, and 200 epochs for GREC, using a batch size of 32. Following standard practices, images are resized to 640×640, and the length of language expressions is trimmed to 20 for all the datasets. For pre-training, SimVG is trained for 30 epochs and then fine-tuned for another 10 epochs. The pre-training experiments are run on 8 NVIDIA RTX 3090 GPUs. All the other experiments are conducted on 2 NVIDIA RTX 4090 GPUs. More implementation details are reported in Appendix C.

4.3 Comparison with The State-of-the-art

In this part, we compare our SimVG with the SOTA methods on six mainstream datasets. We combine the results of RefCOCO/+/g, ReferItGame and Flickr30K datasets in Table 1, and the results of GREC are reported in Table 2. Table 3 reports the results pre-trained on the large corpus of data.

According to Table 1, our model performs better than two-stage models, especially MAttNet [79] while being 7 times faster. We also surpass one-stage models that exploit prior and expert knowledge, with +14% absolute average improvement over ReSC [73]. Additionally, the use of a patch stride of 32 and a lightweight head design has enabled SimVG to achieve an inference speed of only 44 ms on a GTX 1080Ti GPU. For transformer-based models, SimVG surpasses the recent SOTA method Dynamic MDETR [57] with an average of up to 4.4% absolute performance improvement.

Table 1: Comparison with some SOTA methods. RN101 refers to ResNet101 [16], DN53 denotes DarkNet53 [52], ViT-B/32 means ViT-Base [11] with stride of 32 in visual embedding. ∗ denotes testing with a NVIDIA RTX 3090 GPU, while other entries are tested with a GTX 1080Ti GPU. SimVG-TB and SimVG-DB refer to the SimVG model using only the token and decoder branch for inference, respectively.

Models	Visual Encoder	RefCOCO			RefCOCO+			RefCOCOg			ReferIt	Flickr30k	Time
		val	testA	testB	val	testA	testB	val-g	val-u	test-u	test	test	(ms)
Two-Stage													
MAttNet [79]	RN101	76.40	80.43	69.28	64.93	70.26	56.00	-	66.58	67.27	29.04	-	320
CM-Att-Erase [44]	RN101	78.35	83.14	71.32	68.09	73.65	58.03	-	67.99	68.67	-	-	-
DGA [72]	VGG16	-	78.42	65.53	-	69.07	51.99	-	-	63.28	-	-	341
RvG-Tree [18]	RN101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51	-	-	-
NMTree [38]	RN101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44	-	-	-
One-Stage													
RealGIN [85]	DN53	77.25	78.70	72.10	62.78	67.17	54.21	-	62.75	62.33	-	-	35
FAOA [75]	DN53	71.15	74.88	66.32	56.86	61.89	49.46	-	59.44	58.90	60.67	68.71	39
RCCF [36]	DLA34	-	81.06	71.85	-	70.35	56.32	-	-	65.73	63.79	-	25
MCN [46]	DN53	80.08	82.29	74.98	67.16	72.86	57.31	-	66.46	66.01	-	-	56
ReSC$_L$ [73]	DN53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20	64.60	69.28	36
LBYL [20]	DN53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-	67.47	-	30
Transformer-Based													
TransVG [8]	RN101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73	70.73	79.10	62
TRAR [86]	DN53	-	81.40	78.60	-	69.10	56.10	-	68.90	68.30	-	-	-
VGTR [12]	RN50	78.29	81.49	72.38	63.29	70.01	55.64	61.64	64.19	64.01	63.63	75.44	-
SeqTR [87]	DN53	83.72	86.51	81.24	71.45	76.26	64.88	71.50	74.86	74.21	69.66	81.23	50
VLTVG [70]	RN50	84.53	87.69	79.22	73.60	78.37	64.53	72.53	74.90	73.88	71.60	79.18	79*
TransCP [60]	RN50	84.25	87.38	79.78	73.07	78.05	63.35	72.60	-	-	72.05	80.04	74*
Dyn.MDETR [57]	ViT-B/16	85.97	88.82	80.12	74.83	81.70	63.44	72.21	74.14	74.49	70.37	81.89	-
SimVG-TB (ours)	ViT-B/32	87.07	89.04	83.57	78.84	83.46	70.67	77.66	79.82	79.93	74.59	81.59	44
SimVG-DB (ours)	ViT-B/32	87.63	90.22	84.04	78.65	83.36	71.82	78.81	80.37	80.51	74.83	82.04	52
SimVG-TB (ours)	ViT-L/32	**90.61**	**92.53**	**87.68**	**85.36**	**89.61**	**79.74**	79.34	**85.99**	**86.83**	**79.30**	82.61	101
SimVG-DB (ours)	ViT-L/32	90.51	92.37	87.07	84.88	88.50	78.66	**80.42**	85.72	86.70	78.75	**83.15**	116

As shown in Table 1, on the ReferItGame and Flickr30K Entities datasets which mostly contain short noun phrases, the performance boosts to 74.83 and 82.04 with a large margin over the previous one-stage method [73]. Compared to existing transformer-based methods [8, 32, 87], SimVG still significantly outperforms most SOTA methods by approximately 3.4 points on the ReferItGame dataset, and it also slightly outperforms Dynamic MDETR [57] on the Flickr30k dataset. Furthermore, scaling the model from base to large has led to significant improvements across all the datasets.

SimVG can be seamlessly extended to GREC without any network modification. As shown in Table 2, SimVG achieves a significant improvement over existing publicly available methods on the GRefCOCO dataset, with an average increase of 9 points, surpassing UNINEXT [69].

Table 3 demonstrates that when pre-training on a large corpus of image-text pairs, SimVG exhibits greater data efficiency as compared with most of the existing SOTA methods. Despite utilizing only 28K images, which is nearly six times fewer than MDETR [25], and three times fewer than RefTR [32], SimVG still achieves SOTA performance, surpassing most existing methods by a significant margin. Compared to MDETR, SimVG demonstrates an average improvement of 5 points, and compared to the recent SOTA model GroundingDINO [43], it achieves an average improvement of 2 points. Moreover, increasing the volume of pre-training data further enhances performance. Additionally, SimVG applies a lighter transformer structure in the head. Specifically, SimVG-TB only uses 1.58 million parameters, which is smaller than some lightweight models [87, 32]. Lastly, we observe that scaling the multimodal encoder from ViT-B to ViT-L results in the performance of the lightweight token branch surpassing that of its teacher model. We hypothesize that as the model size increases, the reliability of the decoder branch's performance improves, helping to mitigate the impact of mislabeled ground truth data. This, in turn, enhances the generalization ability of the token branch, further demonstrating the effectiveness of the DWBD method.

4.4 Ablation Studies

4.4.1 Multi-Modality Encoder Architecture

To investigate the advantages of decoupling multimodal fusion from visual grounding, we design three architectures for experimental verification. To ensure fairness, we consistently employ the ViT-B/32 model for feature extraction and the VGTR [12] head for prediction. "CLIP" represents a

Table 2: GREC benchmark results on GRefCOCO dataset. Threshold is set to 0.7 for all the methods.

Methods	Visual Encoder	Textual Encoder	val		testA		testB	
			Prec@(F_1@0.5)	N-acc.	Prec@(F_1@0.5)	N-acc.	Prec@(F_1@0.5)	N-acc.
MCN [46]	DN53	GRU	28.0	30.6	32.3	32.0	26.8	30.3
VLT [10]	DN53	GRU	36.6	35.2	40.2	34.1	30.2	32.5
MDETR [25]	RN101	RoBERTa	42.7	36.3	50.0	34.5	36.5	31.0
UNINEXT [69]	RN50	BERT	58.2	50.6	46.4	49.3	42.9	48.2
SimVG-TB (ours)	ViT-B/32	/	<u>61.3</u>	**56.1**	<u>61.7</u>	**58.0**	<u>53.1</u>	**57.5**
SimVG-DB (ours)	ViT-B/32	/	**62.1**	<u>54.7</u>	**64.6**	<u>57.2</u>	**54.8**	<u>57.2</u>

Table 3: Comparison with pre-trained models on RefCOCO [80], RefCOCO+ [80], and RefCOCOg [48] datasets. We only count the parameters of transformer architecture in head.

Models	Visual Encoder	Params (M)	Pre-train images	RefCOCO			RefCOCO+			RefCOCOg		Time (ms)
				val	testA	testB	val	testA	testB	val-u	test-u	
UNITER$_L$ [6]	RN101	-	4.6M	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77	-
VILLA$_L$ [14]	RN101	-	4.6M	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	-
MDETR [25]	RN101	17.36	200K	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	108
RefTR [32]	RN101	17.86	100K	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01	**40**
SeqTR [87]	DN53	7.90	174K	87.00	90.15	83.59	78.69	84.51	71.87	82.69	83.37	50
UniTAB [74]	RN101	-	200K	88.59	91.06	83.75	80.97	85.36	71.55	84.58	84.70	-
DQ-DETR [41]	RN101	-	200K	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44	-
GroundingDINO [43]	Swin-T	-	200K	89.19	91.86	85.99	81.09	87.40	74.71	84.15	84.94	120
PolyFormer [40]	Swin-B	-	174K	89.73	91.73	86.03	83.73	88.60	76.38	84.46	84.96	-
PolyFormer [40]	Swin-L	-	174K	90.38	92.89	87.16	84.98	89.77	77.97	85.83	85.91	-
OFA-L [62]	RN152	-	20M	90.05	92.93	85.26	85.80	89.87	79.22	85.89	86.55	-
mPLUG-2 [67]	ViT-L/14	-	14M	92.40	94.51	88.42	86.02	90.17	78.17	85.88	86.42	-
SimVG-DB (ours)	ViT-B/32	6.32	28K	90.98	92.68	87.94	84.17	88.58	78.53	85.90	86.23	52
SimVG-TB (ours)	ViT-B/32	1.58	174K	90.59	92.80	87.04	83.54	88.05	77.50	85.38	86.28	<u>44</u>
SimVG-DB (ours)	ViT-B/32	6.32	174K	91.47	93.65	87.94	84.83	88.85	79.12	86.30	87.26	52
SimVG-TB (ours)	ViT-L/32	1.58	28K	**92.99**	**94.86**	<u>90.12</u>	**87.43**	<u>91.02</u>	<u>82.10</u>	<u>87.95</u>	<u>88.96</u>	101
SimVG-DB (ours)	ViT-L/32	6.32	28K	<u>92.93</u>	<u>94.70</u>	**90.28**	<u>87.28</u>	**91.64**	**82.41**	**87.99**	**89.15**	116

typical dual-stream multimodal pretraining structure. "ViLT" represents a one-stream multimodal fusion method. "BEiT-3" represents a dual-stream method with a fusion encoder. The experimental results are reported in Table 4. Approaches like ViLT and BEiT-3, which decouple the multimodal fusion process from downstream, show significant improvements compared to the methods that adopt a multimodal independent encoder architecture.

However, this experiment does not aim to demonstrate the superiority of architectures like BEiT-3. Our focus is to highlight that, by decoupling multimodal fusion and leveraging readily available multimodal pretrained weights, we can significantly enhance the convergence speed and performance of visual grounding. As depicted in Fig. 4, ViLT and BEiT-3 demonstrate notably accelerated convergence by decoupling multimodal fusion. In contrast, although CLIP leverages a large amount of image-text data for pre-training, it only performs cross-modal alignment and does not integrate the information from the image and text models to achieve a fused representation.

As shown in Table 4, building upon BEiT-3. We observe that increasing the stride size of the original visual embedding from 16 to 32 and applying bilinear interpolation to the convolutional kernel significantly enhances performance. This is because bilinear interpolation preserves the original feature distribution after compression, thereby accelerating convergence. Furthermore, experimental results from the decoder and token branches reveal a notable performance gap, highlighting the necessity of designing dynamic weight-balance distillation to mitigate this disparity.

Method (ViT-B/32)	RefCOCO		
	val	testA	testB
CLIP [51]	73.93	77.14	67.43
ViLT [27]	78.54	82.31	72.47
BEiT-3 [64]	82.35	84.66	78.38
Baseline (BEiT-3)	82.35	84.66	78.38
+VE Interp.	85.37(**+3.02**)	86.67(**+2.01**)	81.57(**+3.19**)
Token Branch	85.47	86.75	81.66
Decoder Branch	86.78	88.19	82.83

Table 4: Some ablation experiments on different multimodal fusion architectures. VE Interp. refers to the downsampling convolution kernel in Visual Embed that performs bilinear interpolation from pre-trained weights.
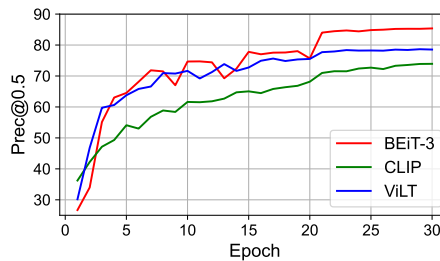

Figure 4: The convergence speed of three different multimodal pretraining architecture models.

Method	RefCOCO		
	val	testA	testB
Token Branch			
Baseline	85.47	86.75	81.66
TQG	86.20(**+0.73**)	88.11(**+1.36**)	82.43(**+0.77**)
Decoder Branch			
Baseline	86.78	88.19	82.83
Mask Max Pool	87.21	88.20	83.28
TQG	87.44(**+0.66**)	88.84(**+0.65**)	83.61(**+0.78**)

Table 5: Ablation studies related to TQG module in the token and decoder branches.

Figure 5: The impact of TQG transformer layers, average refers to the mean value of three sets.

Method	RefCOCO		
	val	testA	testB
Baseline	85.47	86.75	81.66
One Stage Distill			
DETR Distill[3]	86.14	87.50	81.54
Merge Distill	85.98	87.27	82.09
DWB Distill	86.57(**+1.10**)	87.80(**+1.05**)	82.71(**+1.05**)
Two Stage Distill			
DETR Distill[3]	86.49	88.25	82.30
Merge Distill	86.02	88.03	82.56
DWB Distill	86.96(**+1.49**)	88.22(**+1.47**)	83.16(**+1.50**)

Table 6: Ablation studies related to DWBD module, including one/two stage distillation results.

Figure 6: Curves of $\mathcal{L}_{dwbd}$ and W_{dt}, along with the accuracy of the decoder and token branches.

4.4.2 Text-guided Query Generation

As indicated in Table 5, experimental results demonstrate a clear positive impact of the TQG module on both the token and decoder branches, achieving an average absolute improvement of 0.8 points. This guidance mechanism aligns with the concept of DAB-DETR [42], which injects textual priors into queries to imbue them with target-pointing properties. "Mask Max Pool" involves using a text mask to select valid text tokens and then performing max pooling to compress the dimensions, as represented in Sec. 3.3. Furthermore, Fig. 5 illustrates the impact of transformer layers on TQG. The 2-layer transformer structure is adopted to balance both efficiency and performance.

4.4.3 Dynamic Weight-Balance Distillation

The distillation experiments are shown in Table 6, where "DETR Distill" adopts the settings from [3] and uses the predictions from the decoder branch as the teacher for learning. "Merge Distill" combines the ground truth with the decoder prediction, enabling the token branch to select matching targets adaptively. It can be observed that all the three distillation methods improve the performance of the token branch, with the two-stage distillation method further enhancing its performance. Ultimately, our proposed DWBD achieves an average improvement of 1.5 points compared to the baseline. From Table 1 and Table 3, we observe that when employing the ViT-L as teacher model, the performance of the lightweight token branch can even surpass that of the decoder branch on certain metrics during synchronous learning. We hypothesize that this is primarily because the token branch distills more robust feature representations as the teacher's cognitive capabilities improve. Additionally, Fig. 6 illustrates the dynamic balance process of DWBD during training. We can observe that as the confidence of the decoder branch increases, the value of W_{dt} rises correspondingly, indicating that the decoder branch provides more guidance to the token branch. This mechanism allows for the dynamic adjustment of guidance distribution between the ground truth and the decoder prediction.

5 Visualization

We conduct an attention analysis of SimVG from two perspectives, as shown in Fig. 7. First, we visualize the multimodal representations of BEiT-3 using GradCAM [56] to generate heatmaps, revealing that BEiT-3 primarily focuses on global foreground information. Additionally, we visualize the attention maps of the decoder, which highlight the model's focus on regions referred to by the text. More qualitative results can be found in Appendix G.
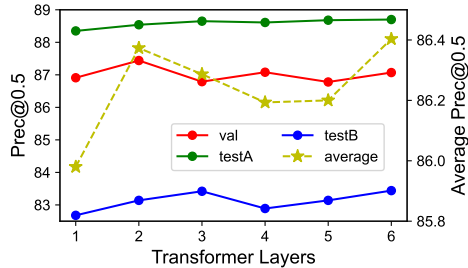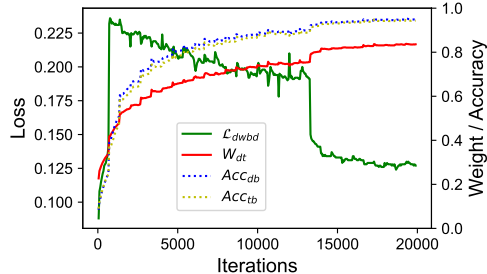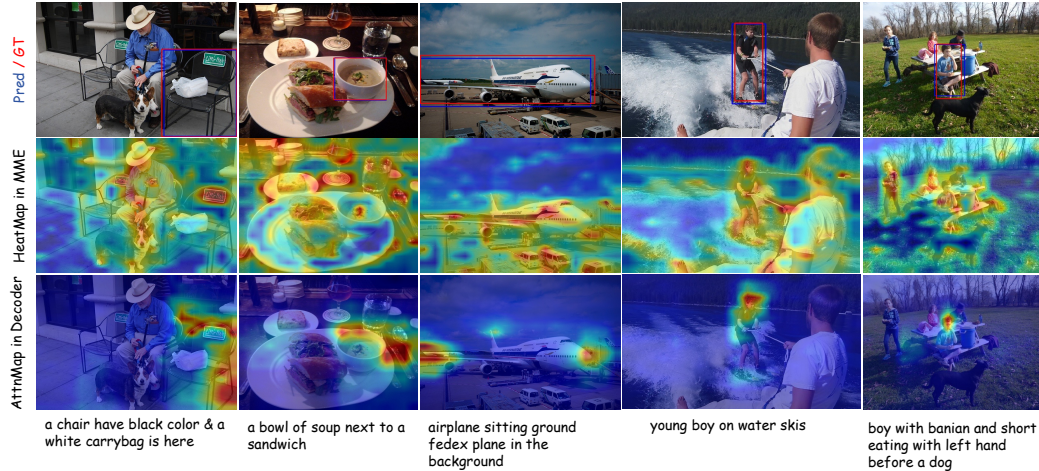
Figure 7: Visualization of the feature maps in MME and the decoder. The heatmap in MME is generated using GradCAM [56], while the attention response of the decoder is obtained from the attention map.

6 Conclusion

In this paper, we re-examine the visual grounding task by decoupling image-text mutual understanding from the downstream task. We construct a simple yet powerful model architecture named SimVG, which leverages the existing research in multimodal fusion to fully explore the contextual associations between modalities. Additionally, to simplify the whole pipeline while maintaining performance, we adopt dynamic weight-balance distillation (DWBD) to let the stronger decoder branch guide the lightweight token branch while learning synchronously. Furthermore, we propose a text-guided query generation (TQG) module to provide textual prior knowledge for object queries. Experimental results demonstrate that SimVG not only achieves improvements in efficiency and convergence speed but also attains new state-of-the-art performance across various benchmarks.

Acknowledgements

References

[1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[3] Jiahao Chang, Shuo Wang, Hai-Ming Xu, Zehui Chen, Chenhongyi Yang, and Feng Zhao. Detrdistill: A universal knowledge distillation framework for detr-families. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6898–6908, 2023.

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv*, 2023.

[5] Xiaokang Chen, Jiahui Chen, Yan Liu, and Gang Zeng. D3etr: Decoder distillation for detection transformer. *arXiv*, 2022.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, 2020.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255. Ieee, 2009.

[8] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1769–1779, 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018.

[10] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16321–16330, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020.

[12] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, 2022.

[13] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding (CVIU)*, 114(4):419–428, 2010.

[14] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6616–6628, 2020.

[15] Yushuo Guan, Pengyu Zhao, Bingxuan Wang, Yuanxing Zhang, Cong Yao, Kaigui Bian, and Jian Tang. Differentiable feature aggregation search for knowledge distillation. In *European Conference on Computer Vision (ECCV)*, pages 469–484. Springer, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension. *arXiv*, 2023.

[18] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[19] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1115–1124, 2017.

[20] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16888–16897, 2021.

[21] Linjiang Huang, Kaixin Lu, Guanglu Song, Liang Wang, Si Liu, Yu Liu, and Hongsheng Li. Teach-detr: Better training detr with teachers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[22] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 4904–4916, 2021.

[24] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv*, 2023.

[25] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1780–1790, 2021.

[26] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.

[27] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning (ICML)*, pages 5583–5594. PMLR, 2021.

[28] 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 (CVPR)*, pages 4015–4026, 2023.

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[30] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning (ICML)*, 2022.

[32] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[33] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[34] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision (ECCV)*, pages 280–296. Springer, 2022.

[35] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 6657–6678, 2024.

[36] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10880–10889, 2020.

[37] 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 (CVPR)*, 2023.

[38] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4673–4682, 2019.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems (NeurIPS)*, 36, 2024.

[40] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18663, 2023.

[41] Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 1728–1736, 2023.

[42] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations (ICLR)*, 2022.

[43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv*, 2023.

[44] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1950–1959, 2019.

[45] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2604–2613, 2019.

[46] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10034–10043, 2020.

[47] 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 (CVPR)*, pages 11–20, 2016.

[48] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 792–807. Springer, 2016.

[49] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv*, 2023.

[50] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Internatioanl Journal of Computer Vision (IJCV)*, 123(1):74–93, 2017.

[51] 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 (ICML)*, pages 8748–8763. PMLR, 2021.

[52] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

[54] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 658–666, 2019.

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision (IJCV)*, 115:211–252, 2015.

[56] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[57] Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[58] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[59] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[60] Wei Tang, Liang Li, Xuejing Liu, Lu Jin, Jinhui Tang, and Zechao Li. Context disentangling and prototype inheriting for robust visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[62] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv*, 2022.

[63] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4933–4942, 2019.

[64] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023.

[65] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 346–362. Springer, 2020.

[66] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10687–10698, 2020.

[67] Haiyang Xu, Qinghao Ye, Mingshi Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qiuchen Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Feiran Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning (ICML)*, 2023.

[68] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13030–13039, 2024.

[69] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15325–15336, 2023.

[70] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9499–9508, 2022.

[71] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

[72] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4644–4653, 2019.

[73] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–404, 2020.

[74] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision (ECCV)*, pages 521–539. Springer, 2022.

[75] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4683–4693, 2019.

[76] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024.

[77] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv*, 2023.

[78] 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*, 2023.

[79] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[80] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 69–85, 2016.

[81] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3903–3911, 2020.

[82] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv*, 2023.

[83] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4158–4166, 2018.

[84] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations (ICLR)*, 2020.

[85] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

[86] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2074–2084, 2021.

[87] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision (ECCV)*, pages 598–615. Springer, 2022.

[88] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4252–4261, 2018.

Appendix

A Datasets

RefCOCO/+/g. RefCOCO [80] contains 142,210 referring expressions, 50,000 referred objects, and 19,994 images. The testA set primarily describes people, while the testB set mainly describes objects other than people. Similarly, RefCOCO+ [80] contains 141,564 expressions, 49,856 referred objects, and 19,992 images. RefCOCO+ referring expressions focus more on attributes of the referent, such as color, shape, and digits, and avoid using words indicating absolute spatial location. RefCOCOg [47, 48] is divided into two partitions: the *google* split [47] and the *umd* split [48]. Each split includes 95,010 referring expressions, 49,822 referred objects, and 25,799 images.

ReferItGame. ReferItGame [26] comprises 120,072 referring expressions and 99,220 referents corresponding to 19,997 images sourced from the SAIAPR-12 [13] dataset. The dataset is partitioned using the cleaned Berkeley split, with 54,127, 5,842, and 60,103 referring expressions allocated to the train, validation, and test sets, respectively.

Flickr30K. Flickr30K Entities [50] is characterized by short region phrases used as language queries, rather than complete sentences, which may describe multiple objects. The dataset consists of 31,783 images with a total of 427,000 referred entities across the train, validation, and test sets.

GRefCOCO. GRefCOCO [37] comprises 278,232 expressions, which include 80,022 multi-target expressions and 32,202 no-target expressions, referring to 60,287 distinct instances in 19,994 images. Some single-target expressions are inherited from RefCOCO [80].

Pre-Training Dataset. Following the approach in [87], we combine region descriptions from the Visual Genome [29] dataset, annotations from RefCOCO/+/g [80, 47, 48], ReferItGame [26], and language queries from the Flickr30K Entities [50]. Our pre-training is configured in two modes. One uses only the lightweight COCO series dataset [80, 47, 48], which includes approximately 321k distinct language expressions and 28k images. The other mode is configured to be consistent with SeqTR [87], containing approximately 6.1M distinct language expressions and 174k images.

B Evaluation Metrics

Precision@0.5 [Prec@0.5]: REC and phrase localization, we evaluate the performance using Precision@0.5. The prediction is deemed correct if its IoU with ground-truth box is larger than 0.5.

Precision@(F_1=1, IoU≥0.5) [Prec@(F_1@0.5)]: The percentage of samples achieving an F_1 score of 1 with an IoU threshold of 0.5 is computed. A predicted bounding box is considered a TP if it has a matching ground-truth bounding box with an IoU≥0.5. If multiple predicted bounding boxes match one ground-truth bounding box, only the one with the highest IoU is considered TP, and the others are FP. Ground-truth bounding boxes with no matched bounding box are FN, while predicted bounding boxes with no matched ground-truth bounding box are FP. The F_1 score for a sample is calculated as $F_1 = \frac{2TP}{2TP+FN+FP}$. A sample is considered successfully predicted if its F_1 score is 1. For samples with no target, the F_1 score is 1 if there is no predicted bounding box, otherwise 0. The ratio of successfully predicted samples is then computed as Precision@(F_1=1, IoU≥0.5).

N-acc: No-target accuracy (N-acc) evaluates the model's ability to identify samples with no target. In a no-target sample, predicting no bounding box is a TP, otherwise it's a FN. N-acc is calculated as $\frac{TP}{TP+FN}$, reflecting the model's performance in identifying samples with no target.

C More Implementation Details

In this section, we provide additional details about the experimental settings described in the main text. For all ablation experiments, we use 512×512 sized images as input. For ViT-B experiments, the two-stage distillation experiments use an additional 20 epochs of training. The description in Sec. 4.2 applies to all SimVG-base models; however, we make some adjustments for the large models. Due to the higher memory usage of the large models, all large models are trained with a batch size of 4. For the large model, the decoder's projection input dimension is increased from 768 to 1024, while all other settings remain consistent with the base model. Additionally, in the pre-training experiments

presented in Table 3, the number of training epochs for the large models is reduced from 30 to 20 due to the increased training cost and the number of token branch distillation epochs is reduced from 20 to 10. In the experiments, all results for SimVG-TB are obtained using DWBD in a two-stage distillation process. For SimVG-DB, the results are obtained by supervising the decoder branch with ground truth. In the GREC experiment as shown in Table 2, we set the number of object queries to 10. Furthermore, the distillation parameters for all base models are set to $\gamma_1 = 2$ and $\gamma_2 = 1$, while for all large models, they are set to $\gamma_1 = 1$ and $\gamma_2 = 0.4$. All training is performed without using the exponential moving average (EMA) strategy. Lastly, it is important to emphasize that the BEiT-3 pre-trained model used in this paper was not trained on the six datasets used for validation in this study.

D   Additional Exploration Studies

D.1   Number of Layers and Query

The number of transformer layers in the decoder branch and MLP layers in the token branch can impact performance and efficiency. Excessive layers can reduce computational efficiency and increase parameter count, while too few layers may lead to subpar performance. To determine the optimal number of layers, we conducted experiments under varying settings, with results presented in Fig. 8 and Fig. 9. Fig. 8 indicates that increasing the number of MLP layers does not yield significant gains. Therefore, we selected single MLP layer. In contrast, Fig. 9 shows that increasing the number of transformer layers results in performance improvement. Considering efficiency, we opted for 3 transformer layers.
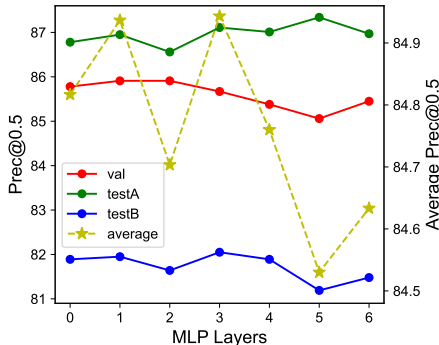

Figure 8: Ablation study on number of token MLP layers in RefCOCO dataset.
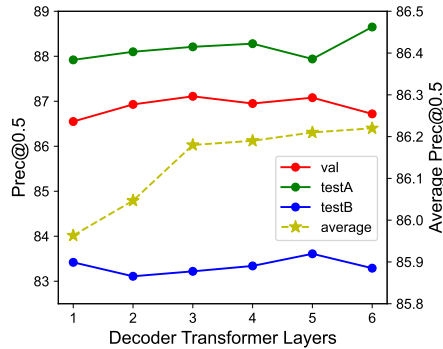
Figure 9: Ablation study on number of decoder transformer layers in RefCOCO dataset.

D.2   Efficiency of Training

We analyze the training efficiency of SimVG from two aspects. As shown in Table 10, SimVG significantly outperforms mainstream methods in terms of the required number of epochs and training time. On the RefCOCO+ dataset, SimVG only requires 30 epochs and 5.5 hours, faster than SeqTR's 60 epochs and 9 hours, demonstrating the accelerated convergence brought by multimodal pre-training. As shown in Table 11, SimVG adopts a more lightweight head design and, despite using less pre-training data, achieves notable performance improvements.

D.3   Compare with Multi-modal Large Language Models

With the surge of large models across various domains, there have been methods proposed for the visual grounding task that leverage multimodal large language models. These methods combine autoregressive and prompt-based approaches to locate targets. Most of these models have billions of parameters, benefiting from their base pretrained models being pretrained on very large datasets, thus exhibiting strong robust performance. We compare these large model approaches with our proposed SimVG, as shown in Table 7. Despite our model having an order of magnitude fewer parameters compared to large models, it still achieves competitive performance, thanks to its decoupled multimodal

Method	Visual Encoder	Epoch	Training Time
TransVG	RN50	180	~90h
VGTR	RN50	120	~50h
Dynamic MDETR	ViT-B/16	90	-
SeqTR	DN53	60	~9h
SimVG	ViT-B/32	30	~5.5h

Figure 10: Training time on a single RTX 4090 GPU and the number of epochs required for convergence on the RefCOCO+ dataset.

Method	Head Params (M)	Pretrain Images	RefCOCO		
			val	testA	testB
RefTR	17.86	100K	85.65	88.73	81.16
MDETR	17.36	200K	86.75	89.58	81.41
SeqTR	7.90	174K	87.00	90.15	83.59
SimVG-DB	6.32	28K	90.98	92.68	87.94
SimVG-TB	1.58	28K	90.18	92.41	87.21

Figure 11: A comparison of the number of parameters in the head and the number of pre-training images for some publicly available methods.

Models	LLM Size	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
KOSMOS-2 [49]	1.6B	52.32	57.42	47.26	45.48	50.73	42.24	60.57	61.65
Shikra [4]	7B	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
NExT-Chat* [82]	7B	85.50	90.00	77.90	77.20	84.50	68.00	80.10	79.80
Ferret* [78]	7B	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76
GroundingGPT [35]	7B	88.02	91.55	82.47	81.61	87.18	73.18	81.67	81.99
PixelLLM [68]	4B	89.80	92.20	86.40	83.20	87.00	78.90	84.60	86.00
COMM [24]	7B	91.73	94.06	88.85	87.21	**91.74**	81.39	87.32	88.33
SimVG-DB-Base (ours)	**0.18B**	91.47	93.65	87.94	84.83	88.85	79.12	86.30	87.26
SimVG-DB-Large (ours)	0.61B	**92.87**	**94.35**	**89.46**	**87.28**	91.64	**82.41**	**87.99**	**89.15**

Table 7: Performance comparison on the REC task. "*" indicates that the model employs additional image region perception modules.

understanding. Moreover, from the experiments, we also observe that as the model's parameter count increases, SimVG's performance shows an increasing trend.

D.4   Analysis of Error Bars

We conducted experiments with VGTR [12], SeqTR [87], and our proposed SimVG, each repeated five times with different random seeds. As illustrated in Fig. 12, our approach exhibits a notably higher median accuracy compared to the VGTR [12] method. Furthermore, in contrast to SeqTR [87], our method demonstrates a more tightly clustered distribution of results. These results indicate that the proposed SimVG not only enhances accuracy but also increases stability.
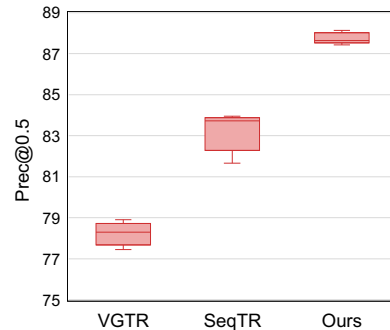

Figure 12: The box-plot on validation set of RefCOCO dataset.

E   Limitations

Our method does not fully explore or utilize hierarchical information in features. Approaches like the FPN in ViTDet [34] that expand feature hierarchy could be considered to further enhance the model's ability to capture targets of different scales. Our method can be applied not only to detection-related tasks but also to segmentation-related tasks. Further validation on more downstream tasks is warranted to demonstrate its stable effectiveness.

F   Broader Impacts

Further research and careful consideration are necessary when utilizing this technology, as the presented proposed method relies on statistics derived from training datasets that may possess biases and could potentially result in negative societal impacts.

Figure 13: Examples predicted by SimVG on the validation set of RefCOCO/+/g datasets.

G Qualitative Results

In this section, we present the visualizations of the SimVG results for the referring expression comprehension (REC) and general referring expression comprehension task (GREC). We present both the caption and image. The red and blue rectangles on images refer to groundtruth and predict boxes, respectively.

Firstly, in Fig. 13, we present some visual examples of our proposed SimVG on the RefCOCO, RefCOCO+, and RefCOCOg datasets. It can be seen that SimVG can accurately perceive and locate objects even in long texts or complex images.

2nd green bottle from left and left green bottle	25 and batter guy whos swinging	A baby holding a sandwich in front and a kid sucking his finger on the left	A hand holding a photo and a lady standing beside a lake
All the circular pastries with a hole in the center inside the plate	A man standing in the right and a black umbrella	All sofas that are white	All goats

(a) Good Cases

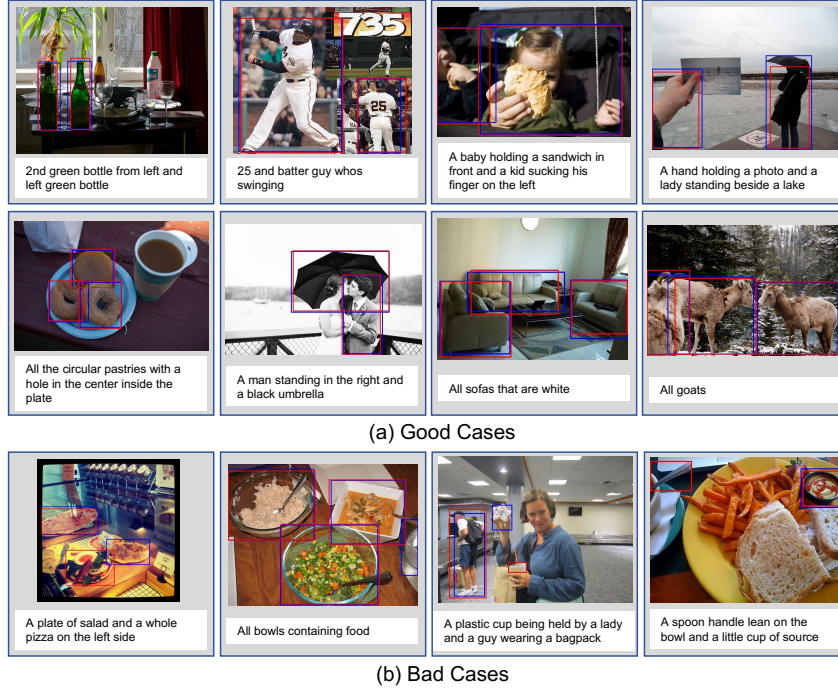|			
A plate of salad and a whole pizza on the left side	All bowls containing food	A plastic cup being held by a lady and a guy wearing a bagpack	A spoon handle lean on the bowl and a little cup of source

(b) Bad Cases

Figure 14: Examples predicted by SimVG on the validation set of GRefCOCO datasets.

Then, in Fig. 14, we show some examples of correct and incorrect results of SimVG on the GRef-COCO dataset. The GREC task requires a deeper semantic understanding of image-text relationships, and in the error cases, the instances of missed detections and false positives significantly increase.

Finally, in Fig. 15, we compare the output results of SimVG with the comparable SeqTR model. We find that the proposed SimVG model can better understand the interrelationships between images and texts.

Figure 15: Some examples comparing SeqTR and SimVG models on the RefCOCOg dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We list the main contributions of this paper at the end of the introduction, and explain the scrope and differences between our method and previous methods in both the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe the limitations of our approach in Sec. E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We present comprehensive experimental results conducted across multiple datasets, comparing our method with prior works to validate its effectiveness.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the structure of our proposed model throughout Sec. 3 and implementation details of the experiment in Sec. 4.2. To further ensure reproducibility, we will soon open source the code of our method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Once the paper is accepted, we will open source our code. In addition, the data we use are all publicly available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation Details in Sec. 4.2 and Appendix Sec. C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We independently set up repeated experiments to analyze the error bar in Sec. D.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Implementation Details in Sec. 4.2 and Appendix Sec. C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We have carefully read the NeurIPS Code of Ethics and have determined that our approach adheres to the relevant ethical guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Although our work is only for academic research purpose, we also discuss the potential positive societal impacts and negative societal impacts in Sec. F.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification:

1. RefCOCO:https://paperswithcode.com/dataset/refcoco
2. GRefCOCO:https://paperswithcode.com/dataset/grefcoco
3. ReferItGame:https://paperswithcode.com/dataset/referitgame
4. Flickr30K Entities:https://paperswithcode.com/dataset/flickr30k-entities

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our work does not release new assets. The data and models used in our work are publicly released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.