F-LMM: Grounding Frozen Large Multimodal Models

Size Wu[1] Sheng Jin[2] Wenwei Zhang[3] Lumin Xu[4] Wentao Liu[2,3] Wei Li[1] Chen Change Loy[1]
[1] S-Lab, Nanyang Technological University [2] SenseTime Research and Tetras.AI
[3] Shanghai AI Laboratory [4] The Chinese University of Hong Kong
size001@e.ntu.edu.sg ccloy@ntu.edu.sg

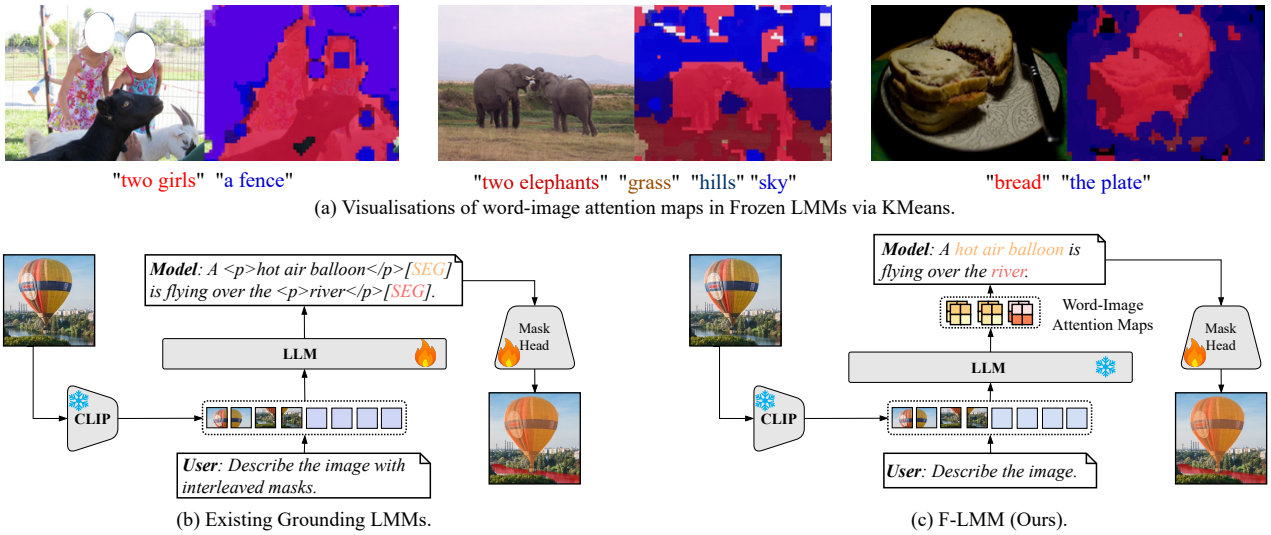
Figure 1. An example of user-AI conversation around an image. **Left**: The current state-of-the-art grounding model GLaMM [60] is effective for grounded conversation when prompted by *"answer with interleaved masks"*, but fails to follow user instruction to answer a single word (yes or no) and misunderstands the question as a referring segmentation prompt. **Right**: Our F-LMM preserves instruction-following ability while being able to perform visual grounding.

Abstract

Endowing Large Multimodal Models (LMMs) with visual grounding capability can significantly enhance AIs' understanding of the visual world and their interaction with humans. However, existing methods typically fine-tune the parameters of LMMs to learn additional segmentation tokens and overfit grounding and segmentation datasets. Such a design would inevitably cause a catastrophic diminution in the indispensable conversational capability of general AI assistants. In this paper, we comprehensively evaluate state-of-the-art grounding LMMs across a suite of multimodal question-answering benchmarks, observing drastic performance drops that indicate vanishing general knowledge comprehension and weakened instruction following ability. To address this issue, we present F-LMM—grounding frozen *off-the-shelf LMMs in human-AI conversations—a straightforward yet effective design based on the fact that word-pixel correspondences conducive to visual grounding inherently exist in the attention mechanism of well-trained LMMs. Using only a few trainable CNN layers, we can translate word-pixel attention weights to mask logits, which a SAM-based mask refiner can further optimise. Our F-LMM neither learns special segmentation tokens nor utilises high-quality grounded instruction-tuning data, but achieves competitive performance on referring expression segmentation and panoptic narrative grounding benchmarks while completely preserving LMMs' original conversational ability. Additionally, with instruction-following ability preserved and grounding ability obtained, F-LMM can be directly applied to complex tasks like reasoning segmentation, grounded conversation generation and visual chain-of-thought reasoning. Our code can be found at https://github.com/wusize/F-LMM.*

1. Introduction

Large Multimodal Models (LMMs), which integrate Large Language Models (LLMs) with visual signals, have demonstrated remarkable success in multimodal understanding, reasoning and interaction [34, 38–40, 44, 65, 78]. To further advance LMMs with better perception capability, a recent line of research [28, 60, 61, 70, 80, 84] that visually grounds language contents in user-model conversations has drawn increasing attention. This explicit association between key phrases/words and visual objects greatly enhances LMMs' understanding of the visual world and allows for more intuitive and meaningful human-AI interactions.



(a) Visualisations of word-image attention maps in Frozen LMMs via KMeans.

(b) Existing Grounding LMMs.

(c) F-LMM (Ours).

Figure 2. (a) Geometric and spatial cues conducive to visual grounding are observed in the visualisations of word-image attention maps in frozen LMMs. (b) Existing grounding LMMs are fine-tuned to generate a special mask token (*e.g.*, `[SEG]`) for visual grounding purposes, which ruins the original conversational ability. (c) Our F-LMM translates word-image attention maps from frozen LMMs to grounding masks, while fully preserving the general-purpose chat capability.

By design, one commonly adopted build (Figure 2(b)) for visually grounding language contents is connecting LMMs with a mask head (*e.g.*, Segment Anything Model (SAM) [26]), wherein both the LLM backbone and the mask head are fine-tuned with well-prepared visual grounding data that contains segmentation annotations. Also, some additional learnable tokens (*e.g.*, `[SEG]`) are introduced to the LMMs' vocabulary, to directly associate key phrases or words with visual objects in conversations. However, this design will inevitably provoke a *catastrophic diminution* in general knowledge comprehension and instruction-following ability due to the following reasons. First, existing segmentation and visual grounding data only contain *elementary* patterns for answering simple grounding prompts. Second, during the fine-tuning stage, the LMMs are mainly optimised for effectively modelling the relationship between key phrases or words and special segmentation tokens, *i.e.*, overfitting the segmentation and grounding data. Therefore, the conversational ability is sacrificed. For instance, the state-of-the-art grounding model GLaMM [60] fails to answer a simple yes-or-no question (Figure 1). Moreover, quantitative evaluations of existing grounding LMMs in conversational ability are presented in Table 1, with zero or near-zero scores on general multimodal question-answering benchmarks necessitating instruction-following ability.

One possible option to deal with this dilemma is to collect high-quality training data encompassing both meaningful conversations and mask annotations. For example, LLaVA-G [80] annotates the 150k LLaVA-Instruct data samples [39] with segmentation masks so that the LMMs simultaneously learn to chat and segment. Nonetheless, annotating high-quality grounded conversation data is costly and hard to scale. Despite being trained on costly annotated data, LLaVA-G still lags behind general-purpose LMMs on multimodal understanding tasks. Furthermore, training on large-scale annotated data normally consumes significant computational resources, which is, obviously not a resource-efficient solution.

In this paper, we propose a simple yet effective design, *i.e.*, grounding frozen LMMs (dubbed as F-LMM) in human-AI conversations. We argue that freezing the parameters of well-trained LMMs is *the most practical* design choice for fully preserving the original excellent conversational ability when building general-purpose grounding LMMs. In particular, we take inspiration from the built-in interpretability of the attention mechanism in transformers [6, 68] that represents interrelations between text and image contents in design. We observe that off-the-shelf LMMs already produce word-pixel correspondences necessary for visual grounding, despite they were not explicitly pre-trained with region or pixel annotations. As illustrated in Figure 2(a), we visualise word-image attention maps from frozen LMMs via K-Means clustering, revealing notable geometric and spatial cues of the objects [1]. For example, coarse visual grounding masks for key phrases (*e.g.*, "two girls", "two elephants", and "the plate") in language sentences emerge from attention maps in LMMs. Therefore, our F-LMM takes these visual-language correspondences as useful segmentation priors for decoding ground-

[1]For better visibility, we perform K-Means clustering on the stack-up of all attention maps collected in a forward pass instead of selecting a single attention map.

ing masks, without further tuning the LMMs' weights or learning a special segmentation token to model object locations, as shown in Figure 2(c).

The only trainable part of our F-LMM is a mask head plus a keyword selector. The mask head comprises a CNN-based mask decoder (a tiny U-Net [62]) that translates stacked attention maps to mask logits and a light-weight mask refiner (retrofitted from SAM [26]'s mask head) that uses additional image and language cues to refine the semantic-agnostic masks from the mask decoder. The keyword selector is a linear layer that discovers object nouns in text sequences, automating the process of grounded human-AI conversation. Moreover, we only use the RefCOCO(+/g) [24, 46] and PNG [16] datasets as our training data, enabling LMMs to segment user-described objects and ground key phrases or words in a text sequence. Unlike previous works [60, 61, 80], our F-LMM eliminates the necessity for high-quality conversation data that are annotated with masks to preserve conversational ability when learning grounding.

Our experiments demonstrate that F-LMM maintains the original excellence of off-the-shelf LMMs on general question-answering benchmarks, while achieving competitive results on referring segmentation and phrase grounding. In more complex tasks like reasoning segmentation, grounded conversation generation and visual chain-of-thought reasoning, F-LMM achieves better or comparable results when contrasted with models specially trained for such tasks. Compared with existing grounding LMMs, F-LMM offers the best balance between grounding and chat capabilities.

2. Related Work

Large Multimodal Models. Recent advancements in LMMs [2, 3, 13, 29–31, 34, 36, 38–40, 44, 47, 65, 75] have been fueled by the success of LLMs [1, 4, 12, 22, 48, 51, 66, 67, 81] since the debut of GPT series [1, 4, 56, 57] that feature an auto-regressive framework based on transformer decoders [68]. These LLMs possess general world knowledge and excellent conversational ability to follow human instructions, thanks to large-scale generative pre-training [4] and supervised finetuning on instruction-tuning data [71] or human feedback [50]. By integrating image representations from vision encoders [58, 79] to LLMs, LMMs enable visual understanding and reasoning in AI assistants. This integration is usually established by a multilayer perceptron (MLP) that directly maps image features to the LLMs' input embedding space [34, 38–40, 44, 65, 78] or a cross-attention module that abstracts the image contents with a set of query embeddings [2, 3, 31, 75]. In our research, we build F-LMM on LMMs of the former type (MLP-based), which preserves images' 2-D topological structure in the cross-modal integration.

Visual Segmentation. The task of predicting 2D masks for visual objects is known as image segmentation, which can be categorised into semantic segmentation [5, 8, 10, 87], instance segmentation [11, 19, 83] and panoptic segmentation [9, 25, 32, 73] depending on whether the goal is to differentiate pixel semantics or object instances. These standard segmentation approaches rely on a pre-defined set of object classes for recognition. In contrast, referring expression segmentation (RES) [24, 37, 45, 46, 49, 74, 88] involves segmenting objects based on free-form human language descriptions, allowing for enhanced human-model interaction. Additionally, panoptic narrative grounding (PNG) [14, 16, 17, 69] requires segmenting masks for key phrases or words in a sentence. In this study, we mainly leverage RES and PNG tasks to evaluate the grounding capability of LMMs. Besides, we also test LMMs' segmentation ability in complex scenarios that necessitate reasoning [28, 60, 63]. Moreover, the prompt-based SAM [26] pre-trained on billion-scale high-quality mask data has become a constituent component in many grounding LMMs to boost segmentation performance. We also adopt SAM's mask head to initialise our mask refiner.

Grounding Large Multimodal Models. Grounding Large Multimodal Models [3, 7, 28, 53, 54, 60, 61, 70, 72, 76, 78, 80, 84, 85] can localise language contents during user-model conversations. Some approaches [3, 7, 53, 76] represent coordinates of bounding boxes as texts and train LMMs to predict the coordinates in a generative manner. Several recent works [28, 54, 60, 61, 70, 80] train LMMs to predict a special segmentation token for encoding the grounded object and utilise a segmentation head (*e.g.*, SAM [26]) to decode object masks. This study mainly focuses on grounding LMMs with segmentation ability for visual perception. To obtain competitive visual grounding performance, existing works extensively fine-tune the parameters of LMMs on a large amount of segmentation [5, 18, 25, 59, 87] and grounding [16, 23, 24, 27, 46, 55] datasets. And to balance the LMMs' grounding and conversational abilities, there are efforts [61, 80, 84] to collect high-quality instruction-tuning data annotated with segmentation masks. In contrast, we make the first attempt to build grounding LMMs on top of off-the-shelf LMMs without fine-tuning their parameters. Furthermore, we bypass the need for grounded instruction-tuning data to preserve decent chat ability.

3. Method

In this section, we introduce our F-LMM by first probing the causal attention mechanism in LMMs with visualisations of word-image attention maps in Sec 3.1. Then, we elaborate on F-LMM exploiting segmentation priors from frozen LMMs for visual grounding using the mask head in Sec 3.2. Finally, we show how to automate the process of grounded conversation with a linear keyword selector to in-

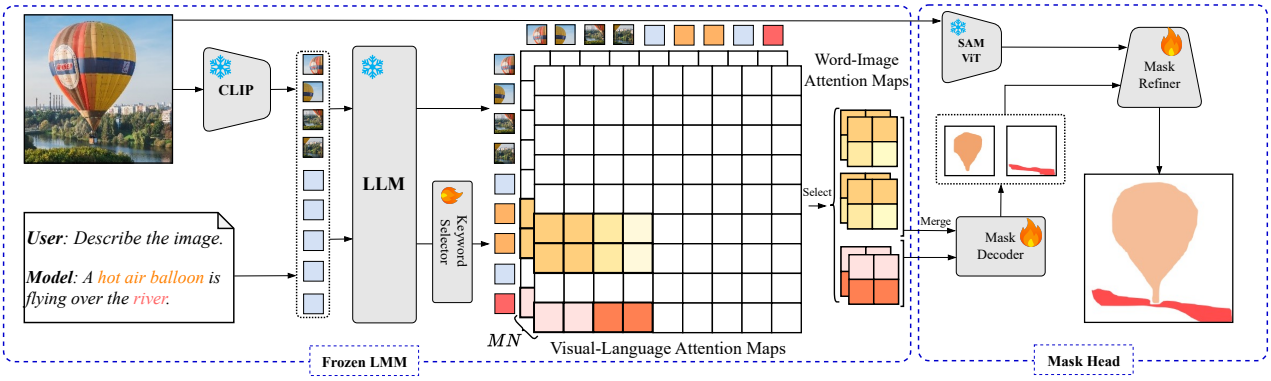


Figure 3. The overall pipeline of F-LMM. The word-image attention maps from the frozen LLM serve as segmentation priors for the mask head. The keyword selector discovers object nouns in the text sequence. The mask head encompasses a mask decoder that translates attention weights to mask logits and a mask refiner that optimises the mask decoder's predictions. M and N represent the numbers of transformer layers and attention heads.

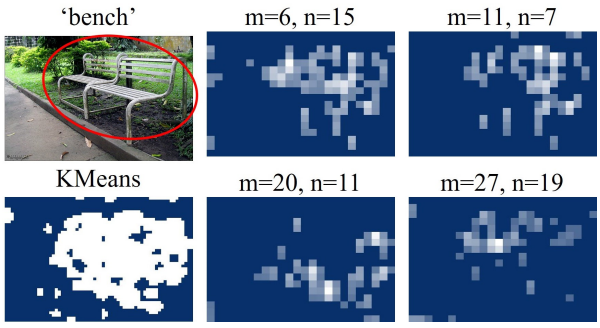


Figure 4. Visualisations of word-image attention maps. The letters m and n indicate that the attention map is derived from the n-th attention head of the m-th transformer layer.

dicate words of grounding targets in Sec 3.3. The overall pipeline is illustrated in Figure 3.

3.1. Segmentation Priors from Frozen LMM

Vision-Language Sequence. A typical build of a LMM [2] comprises an image encoder f_v (*e.g.*, CLIP [58] [3]), a vision-language projector f_p, and a LLM, denoted as f_{llm}. The inputs to an LMM are usually an image $\mathbf{X}_v \in \mathbb{R}^{3\times H\times W}$ and the associated text $\mathbf{X}_t$. The input image is first encoded by the vision encoder f_v and then mapped to the input space of the LLM by the projector f_p:

$$\mathbf{Z}_v = f_p(\texttt{Flatten}(f_v(\mathbf{X}_v))) \in \mathbb{R}^{hw\times d},$$

where h and w are the height and width of projected feature maps via f_v. The `Flatten` operation unfolds the 2-D image feature map to a 1-D sequence. The constant d is the hidden state dimension of the LLM. Likewise, the text input is first encoded as discrete tokens and then mapped to text embeddings:

$$\mathbf{Z}_t = \texttt{Embed}(\texttt{Tokenize}(\mathbf{X}_t)) \in \mathbb{R}^{L\times d},$$

where L denotes the length of text embeddings. The visual-language sequence input to the LLM is a concatenation of image and text embeddings: $\mathbf{Z} = \{\mathbf{Z}_v, \mathbf{Z}_t\} \in \mathbb{R}^{(hw+L)\times d}$.

Segmentation Priors in Self-Attention. The vision-language sequence is mainly processed by causal self-attentions [56, 68] in the LLM, including inner product and weighted-sum operations. Specifically, for a word token with position index i in the vision-language sequence $\mathbf{Z}$, its embedding $\boldsymbol{z}^i$ is updated by the weighted sum of the first i embeddings: $\hat{\boldsymbol{z}}^i = \texttt{SoftMax}(\frac{\boldsymbol{z}^i\cdot\mathbf{Z}[:i]}{d})\cdot\mathbf{Z}[:i]$, where $\texttt{SoftMax}(\frac{\boldsymbol{z}^i\cdot\mathbf{Z}[:i]}{d})$ is the attention weights. Here, we omit the residual layers and feedforward layers for brevity.

Considering the word-image interaction in the multi-modal scenario, we can select the word token's attention weights with the image embeddings from the overall vision-language attention weights:

$$\boldsymbol{a}^i = \texttt{Unflatten}(\texttt{SoftMax}(\frac{\boldsymbol{z}^i\cdot\mathbf{Z}[:i]}{d})[:hw]) \in \mathbb{R}^{h\times w},$$

where `Unflatten` restores the 2-D spatial structure from the 1-D sequence to form an attention map. In Figure 4, we visualise such word-image attention maps from various transformers layers and attention heads in an LMM (*i.e.*, DeepSeekVL-7B [44]). The objects' shape and location can be observed in word-image attention maps of certain layers or heads. The visibility is further enhanced when we stack the attention maps from all layers and heads and perform K-Means clustering. It can be observed that the attention

[2] In this paper, the term 'multimodal' stands for vision and language modalities.

[3] The image encoder might be any vision model that is pre-trained on image-text pairs. We use the classic term 'CLIP' in this paper to represent all such models for brevity.

Table 1. The main evaluation results on question-answering benchmarks, referring expression segmentation (RES) benchmark and panoptic narrative grounding (PNG) benchmark. LLaVA$^{\text{W}}$: LLaVA-In-the-Wild. LLaVA-1.6 and MGM-HD take high-resolution image inputs. LLaVA-1.6-M-7B means the model is based on Mistral-7B [22]. GLaMM-FS-7B means we use the 'FullScope' version of GLaMM.

Model	Multimodal Question Answering				RES			PNG		
	MME	MMBench	MMVet	LLaVA$^{\text{W}}$	RefCOCO	RefCOCO+	RefCOCOg	All	Thing	Stuff
PixelLM-7B [61]	309/135	17.4	15.9	46.4	73.0	66.3	69.3	43.1	41.0	47.9
LISA-7B [28]	1/1	0.4	19.1	47.5	74.9	65.1	67.9	-	-	-
PerceptionGPT-7B [54]	-	-	-	-	75.1	68.5	70.3	-	-	-
LLaVA-G-7B [80]	-	-	-	55.8	77.1	68.8	71.5	-	-	-
GroundHog-7B [84]	-	-	-	-	78.5	70.5	74.1	66.8	65.0	69.4
GLaMM-FS-7B [60]	14/9	36.8	10.3	32.0	78.6	70.5	74.8	55.8	52.9	62.3
GSVA-7B [72]	446/18	17.8	19.4	38.3	77.7	68.2	73.2	41.8	39.6	46.6
LaSagnA-7B [70]	0/0	0.0	16.7	34.5	76.8	66.4	70.6	-	-	-
F-LMM (DeepSeekVL-1.3B [44])	1307/225	64.6	34.8	51.1	75.0	62.8	68.2	64.9	63.4	68.3
F-LMM (MGM-2B [34])	1341/312	59.8	31.1	65.9	75.0	63.7	67.3	65.6	64.4	68.4
F-LMM (LLaVA-1.5-7B [38])	1511/348	64.3	30.5	69.0	75.2	63.7	67.1	64.8	63.4	68.2
F-LMM (HPT-Air-6B [65])	1010/ 258	69.8	31.3	59.2	74.3	64.0	67.5	65.5	64.0	68.8
F-LMM (HPT-Air-1.5-8B [65])	1476/308	75.2	36.3	62.1	76.3	64.5	68.5	65.4	64.1	68.5
F-LMM (MGM-7B [34])	1523/316	69.3	40.8	75.8	75.7	64.8	68.3	66.3	65.3	68.6
F-LMM (DeepSeekVL-7B [44])	1468/298	73.2	41.5	77.8	76.1	66.4	70.1	65.7	64.5	68.5
F-LMM (LLaVA-1.6-7B [40])	1519/322	68.1	44.1	72.3	75.8	65.8	70.1	66.3	65.1	69.0
F-LMM (LLaVA-1.6-M-7B [40])	1501/324	69.5	47.8	71.7	75.7	66.5	70.1	66.5	65.4	69.1
F-LMM (MGM-HD-7B [34])	1546/319	65.8	41.3	74.0	76.1	65.2	68.5	66.7	65.6	69.1

maps offer meaningful *segmentation priors* with spatial and geometric cues for grounding objects visually.

Language Cues. In addition to the spatial and geometric cues from word-image attention maps, F-LMM can also capitalise on the object's corresponding text embeddings from the LLM f_{llm}, which provide extra language cues for the grounding of visual objects.

3.2. Visual Grounding with Mask Head

We use the segmentation priors from the frozen LMM for pixel-level grounding, with the help of a mask head consisting of a mask decoder and a mask refiner.

Mask Decoder. The mask decoder f_d is a 2-D CNN model that transforms the word-image attention maps of grounded objects into mask logits, which is instantiated by a 3-stage U-Net [62]. Please refer to the supplemental material for details on the mask decoder. The extraction of word-image attention map $\boldsymbol{a}^i$ for a word token with position index i is illustrated in Eq. 3.1 and Figure 3. For an object described by multiple words, we merge their corresponding word-image attention maps to a single attention map $\boldsymbol{a}$ via element-wise average or max operation. The attention map $\boldsymbol{a}$ is further normalised as $\boldsymbol{a}/\text{sum}(\boldsymbol{a})$ so that all elements sum to 1. Considering M layers and N attention heads, we stack the MN attention maps as $\boldsymbol{A} \in \mathbb{R}^{MN\times h\times w}$, which forms the input to a mask decoder. Given the importance of high input resolution for segmentation models, we upsample the stacked attention maps $\boldsymbol{A}$ to $h' \times w'$ by bilinear interpolation before feeding it to a mask decoder, where $h' > h$ and $w' > w$. In practice, we set $h' = w' = 64$. Then, the mask decoder maps $\boldsymbol{A}$ into mask logits: $\mathbf{M}_{\text{logits}} = f_d(\boldsymbol{A})$. We derive the corresponding binary mask via $\mathbf{M}_{\text{pred}} = \mathbf{M}_{\text{logits}} > 0$. During training, the mask decoder is optimised with BCE and DICE losses [64].

Mask Refiner. The mask refiner f_r is retrofitted from the mask head of SAM [26], which predicts masks based on prompts as well as image embeddings from SAM's ViT-based image encoder. To refine the output of the mask decoder f_d, we re-use SAM's prompt encoder to transform $\mathbf{M}_{\text{logits}}$ into dense prompt embeddings (*i.e.*, a 2-D feature map) and the bounding box of $\mathbf{M}_{\text{pred}}$ to box embeddings. In addition to the location cues from the mask and the box, the language cues, *i.e.*, the object's corresponding text embeddings, are also utilised by f_r. Considering text embeddings from the M transformer layers, we train M learnable scalars to calculate a weighted sum of these text embeddings. The weighted-summed text embeddings are processed by a linear layer and then concatenated with the box embeddings to form sparse prompt embeddings. The dense and sparse prompt embeddings, together with SAM's image embeddings, are fed to the mask refiner f_r for finer-grained mask predictions $\mathbf{M}'_{\text{pred}}$. During training, we keep the ViT-based image encoder of SAM frozen and optimise the mask refiner f_r using BCE loss and DICE loss [64]. For more details on the SAM's prompt-based mask head, please refer to the original SAM paper [26].

3.3. Keyword Selector for Grounded Conversation

In grounded conversation [60] with interleaved segmentation masks and words, existing grounding LMMs [60, 61] expand LLMs' vocabularies with special tokens that indicate the start and end of grounding targets, which is infeasi-

Table 2. Reasoning Segmentation.

Model	Val	Test		
	-	Short	Long	All
X-Decoder [88]	22.6	20.4	22.2	21.7
SEEM [89]	25.5	20.1	25.6	24.3
GroundingSAM [41]	26.0	17.8	22.4	21.3
OVSeg [35]	28.5	18.0	28.7	26.1
LISA [28]	44.4	37.6	36.6	36.8
F-LMM	46.7	36.9	49.1	46.2

Table 3. Grounded Conversation Generation (GCG). M. stands for METEOR.

Model	GCG	Val			Test		
	Training	M.	mIoU	Recall	M.	mIoU	Recall
LISA [28]	✓	13.0	62.0	36.3	12.9	61.7	35.5
OMG-LLaVA [82]	✓	14.9	65.5	-	14.5	64.7	-
GLaMM [60]	✓	16.2	66.3	41.8	15.8	65.6	40.8
BuboGPT [86]	✗	17.2	54.0	29.4	17.1	54.1	27.0
KOSMOS-2 [52]	✗	16.1	55.6	28.3	15.8	56.8	29.0
F-LMM	✗	17.6	63.5	42.0	17.4	63.6	38.6

Table 4. Unleashing visual chain-of-thought reasoning with both excellent grounding and instruction-following ability.

Model	Visual	VisCoT Benchmark						POPE	
	CoT	DocVQA	TextCaps	TextVQA	DUDE	SROIE	Infographics	Acc	F1
VisCoT-7B [63]	✓	47.6	67.5	77.5	38.6	47.0	32.4	86.5	-
F-LMM	✗	43.2	63.5	74.5	32.0	28.4	43.2	87.0	86.0
F-LMM	✓	53.8	67.9	78.4	42.3	44.1	49.1	88.0	87.7

ble in F-LMM given its 'frozen' nature. A common practice to discover visual objects in text sequences offline is using external tools such as SpaCy [20], which parse all nouns from a sentence including unwanted non-object words. Instead of adopting such offline tools that produce noisy results, we automate generating interleaved words and masks by training a linear layer to directly predict whether a word is to be grounded or not.

Specifically, the linear layer (keyword selector) is placed on top of the LLM's transformer layers, projecting d-dimension hidden state vectors into one-dimension scores, followed by a sigmoid function that normalizes the scores to $[0, 1]$. During training, the score prediction is supervised by a BCE loss. During inference, word tokens with scores exceeding a threshold λ are regarded as positive for visual grounding. In practice, we set $\lambda = 0.3$. Adjacent positive word tokens are grouped to indicate a single visual object. After the word tokens of visual objects are selected, the corresponding attention maps and text embeddings are fed to the mask head for visual grounding.

4. Experiments

4.1. Implementation Details

Model Architectures. We build F-LMM on several open-sourced LMMs, including LLaVA-1.5 [38], LLaVA-Next [40], MiniGemini [34], DeepSeekVL [44] and HPT-Air [65]. The main experiment covers 10 LMMs with model sizes ranging from 1.3B to 8B. We employ a lightweight 3-stage U-Net [62] as the mask decoder to transform segmentation priors from frozen LLMs. The U-Net architecture features an encoder-decoder structure with skip connections, wherein feature maps are downsampled in the encoder and upsampled in the decoder. Please check the supplementary material for more details on the mask decoder. As for the SAM-based mask refiner, we choose SAM ViT-L [26] that balances cost and performance well.

Model Training. We train F-LMM on RefCOCO(+/g) [24, 46] and PNG [16] datasets with about 190k data samples on a single machine with 8 NVIDIA A800-40G GPUs, which costs about 20 hours for each round of model training. We set the batch size to 8 and train models for 8 epochs, with gradient clipping at a max norm of 1.0. The AdamW [43] optimiser is used with a learning rate of 1e-4, a weight decay of 0.01, and betas as (0.9, 0.999). We choose a warm-up ratio of 0.03 to stabilise model optimisation.

4.2. Standard QA and Grounding Tasks

For a comprehensive study of LMMs' conversational and grounding capabilities, we first evaluate models under standard question-answering and grounding benchmarks separately. We summarise the evaluation results of grounding LMMs in Table 1. Please refer to the supplementary material for more detailed results.

Benchmarks. For comprehensive *conversational ability* evaluation, we choose four widely used general question-answering benchmarks including MME [15], MMBench [42], LLaVA-In-the-Wild [39] and MMVet [77]. The MME and MMBench require an LMM to strictly follow the instruction to reply with single words (yes or no) or answer MCQs with alphabetical letters (*i.e.*, answering A, B, C, or D). The LLaVA-In-the-Wild and MMVet benchmarks ask a model to respond with open-ended texts while demanding general world knowledge comprehension. In terms of *grounding ability* evaluation, we assess the LMMs' ability to segment user-described objects on referring expression segmentation (RES) [24, 46] benchmarks including RefCOCO, RefCOCO+, and RefCOCOg, using the cIoU metric. Due to limited space, we only report results on the `Val` splits of RefCOCO(+/g) in Table 1. We also test the LMMs' ability to ground key phrases or words in user-model conversations on the Panoptic Narrative Grounding (PNG) [16] benchmark, measuring individual mask recalls on thing/stuff objects and overall recall scores.

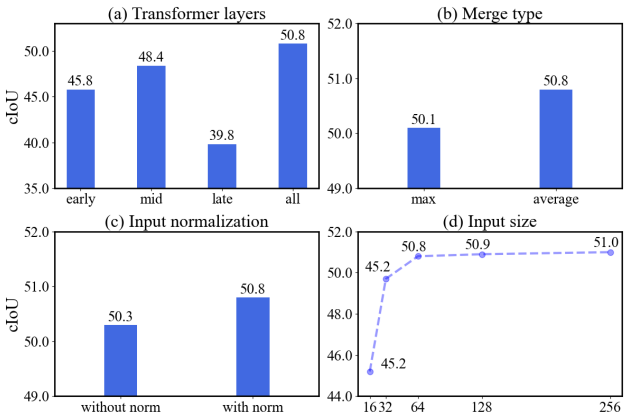


Figure 5. Ablation study of the mask decoder.

Comparisons with Existing Methods. We compare F-LMM with existing grounding LMMs. As shown in Table 1, our F-LMM provides the best balance with conversational and grounding abilities among compared methods. On the question-answering benchmarks, existing grounding LMMs obtain zero or near-zero scores on MMBench and MME while lagging significantly behind general-purpose LMMs on MMVet and LLaVA-In-the-Wild benchmarks, indicating compromised instruction-following ability and weakened general knowledge comprehension. On the RES and PNG benchmarks, our F-LMM achieves comparable results despite not having the parameters of LMMs fine-tuned for grounding purposes.

4.3. Complex Scenarios

In this section, we evaluate grounding LMMs under more complex scenarios that typically require the LMMs to perform both reasoning and segmentation. The base LMM we use is DeepSeekVL-7B [44], considering its flexibility in supporting both single and multiple image inputs. And the size of models compared in this section is also 7B.

Reasoning Segmentation is proposed in LISA [28] that requires a model to infer what object to segment from common-sense knowledge or via logical reasoning. The evaluation results are provided in Table 2 and the metric is gIoU. F-LMM can effectively perform reasoning segmentation even though it is not trained on such type of data. It is remarkable that F-LMM significantly outperforms existing models on the subset of long sentences, reflecting the advantage of F-LMM in handling complex contexts.

Grounded Conversation Generation (GCG) is proposed in GLaMM [60], which requires a model to generate interleaved segmentation masks and texts. For performance evaluation, METEOR (M.) and mIoU are used to measure the quality of generated texts and masks, respectively. In addition, we report the recall of object masks. As shown in Table 3, F-LMM exhibits the best zero-shot performance while being comparable with models fine-tuned on GCG training dataset.

Table 5. Ablation study of the mask refiner on PNG benchmark.

(a) Effects of Mask Refiner

Mask Refiner			PNG		
mask	box	text	All	Thing	Stuff
✗	✗	✗	50.8	48.6	55.9
✓	✗	✗	63.4	62.0	66.8
✓	✓	✗	63.7	62.2	67.1
✓	✓	✓	64.9	63.4	68.3

(b) SAM Variants

SAM	PNG		
	All	Thing	Stuff
ViT-B	63.0	61.4	66.8
ViT-L	64.9	63.4	68.3
ViT-H	65.0	63.5	68.3

Table 6. Ablation study of keyword selection.

Method	F1	Recall	Precision
SpaCy Parser [20]	57.8	97.3	41.1
Linear Keyword Selector	82.8	96.6	72.5

Visual Chain-of-Thought Reasoning. In human-AI conversations that involve Visual Chain-of-Thought Reasoning (VisCoT) [63], an LMM first localises the region/object relevant to the human's question and then generates the final answer by zooming in on the question-related region. Here we evaluate F-LMM and VisCoT-7B on the VisCoT benchmark [63]. As shown in Table 4, F-LMM achieves remarkable performance gains when prompted in the VisCoT manner. It is noticeable that F-LMM even outperforms VisCoT-7B [63] that has been well-tuned on the VisCoT training data [63]. Furthermore, we perform VisCoT on the object hallucination benchmark POPE [33] and observe significant performance gain in resisting object hallucinations.

4.4. Ablation Study

We investigate the effects of design choices of F-LMM. All the ablation studies are conducted on the PNG dataset and we use DeepSeekVL-1.3B [44] unless otherwise stated.

Mask Decoder. We summarise our analyses of the mask decoder in Figure 5. Note that the mask refiner is not involved in this part. We first consider attention maps from different transformer layers, *i.e.*, early (6^{th}), mid (12^{th}) and late (24^{th}) layers. As shown in Figure 5(a), attention maps from late layers perform the worst, conforming to prior studies [21] indicating that deeper transformer layers tend to focus on abstract concepts instead of visual details. And using attention maps from all layers achieves the best performance. Next, we study how to merge the word-image attention maps of multi-word objects, as shown in Figure 5(b). We find that the average operation outperforms the max operation by a margin of 0.7. In Figure 5(c), we show that normalizing the inputs to the mask decoder provides a 0.5 performance gain. Finally, we experiment with different input sizes for the mask decoder. As shown in Figure 5(d), using 64×64 yields the best performance-cost trade-off.

Mask Refiner. We study the effects of the mask refiner on segmentation in Table 5a. The performance of using only the mask decoder is shown in the first row of Table 5a. With the masks fed to the mask refiner, we observe a significant 12.6 performance gain on the PNG benchmark. Adding

Model	Chat ↓	Ground ↓
DeepseekVL-1.3B [44]	7.75	8.33
MGM-2B [34]	6.00	8.33
LLaVA-1.5-7B [38]	6.75	7.83
HPT-Air-6B [65]	9.00	7.16
HPT-Air-1.5-8B [65]	6.50	7.00
MGM-7B [34]	5.75	4.83
DeepseekVL-7B [44]	3.75	4.00
LLaVA-1.6-7B [40]	2.75	3.00
LLaVA-1.6-M-7B [40]	3.25	1.66
MGM-HD-7B [34]	3.50	2.83

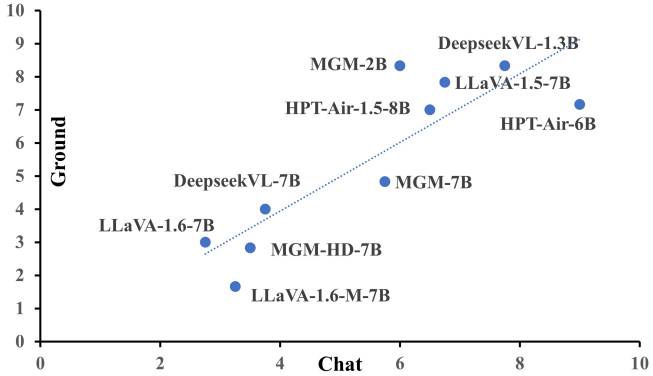


Figure 6. The relevance between an LMM's chat ability and the grounding ability of the F-LMM built on it. The left table shows the average ranks of each LMM on question-answering ('Chat') and grounding benchmarks ('Ground'). ↓ means the lower the better. The dashed line in the right figure is the linear fit of the rank data points, indicating a positive correlation between abilities to chat and ground.

box and text prompts to the mask refiner further improves the performance by 0.3 and 1.2. Then we experiment with different SAM [26] model variants, *i.e.*, ViT-B(ase), ViT-L(arge) and ViT-H(uge) as mask refiners. As shown in Table 5b, the performance grows with model sizes. For a good trade-off between cost and performance, we select the ViT-L variant of SAM as the default mask refiner.

Keyword Selector. We analyse the keyword selector, implemented as a linear layer, on the PNG dataset, using the F1 score as the main metric. As shown in Table 6, our keyword selector achieves significantly higher F1 scores than the external SpaCy tool [20]. We also report recall and precision scores. Our keyword selector achieves comparable recall while being much more precise compared to SpaCy, which enumerates all nouns in a sentence.

4.5. Analysis & Visualisation

Scalability: Does Better Chatting Lead to Better Grounding? We study the relevance between an LMM's chat ability and the grounding ability of the F-LMM built on it. Specifically, we examine the correlation between performance on the question-answering and grounding benchmarks. For the ten models reported in Table 1, we calculate their average ranks in each benchmark category. In Figure 6, we plot these ranks as 2D coordinates, *i.e.*, (`Chat Rank`, `Ground Rank`), and apply a linear fit to the data points. As indicated by the blue dashed line, frozen LMMs with stronger conversational ability can serve as better backbones for grounding. We also observe that larger LMMs generally excel in both conversation and grounding tasks, and LMMs with larger input resolutions (*e.g.*, LLaVA-1.6 and MGM-HD) can handle both tasks better.

From Attention Maps to Segmentation Masks. We visualise the word-image attention maps by applying KMeans clustering to the stacked attention maps that are collected from all transformer layers and attention heads. The attention maps of multiple-word objects are merged by element-wise average. As shown in Figure 7, we observe that the pixels of objects are roughly clustered together (top-left). With the CNN-Based mask decoder, the attention weights are mapped to 2D binary masks (top-right), which are then further optimised by the SAM-based mask refiner (bottom-left). The model used is DeepSeekVL-1.3B [44].

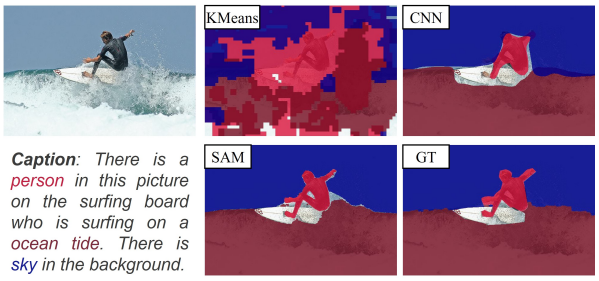


Figure 7. Visualisations of KMeans clustering on attention maps and segmentation results from the CNN-based mask decoder and the SAM-based mask refiner.

5. Conclusion

In this work, we have studied the limitation of existing grounding LMMs, *i.e.*, the loss of general world knowledge and instruction-following ability. To address this issue, we make the first attempt to ground fully frozen LMMs, which are already well-trained for user-model conversation, based on the insight that the geometric and spatial cues needed for visual grounding are inherently present within the self-attention mechanism of LMMs. By incorporating a CNN-based mask decoder and a SAM-based mask refiner, we achieve competitive visual grounding performance without sacrificing any conversational abilities of pre-trained LMMs. With this combination of strong conversational and visual grounding capabilities, these LMMs show promise for complex perception and reasoning tasks, such as segmentation with reasoning, grounded conversation generation, and visual chain-of-thought reasoning.

Acknowledgement. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-048T), the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018.

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020.

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Zihan Ding, Zi-han Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Xiaolin Wei, and Si Liu. Ppmn: Pixel-phrase matching network for one-stage panoptic narrative grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5537–5546, 2022.

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[16] Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic narrative grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1364–1373, 2021.

[17] Tianyu Guo, Haowei Wang, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. Improving panoptic narrative grounding by harnessing semantic relationships and visual confirmation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1985–1993, 2024.

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

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[21] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[22] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lam-

ple, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

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

[25] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *CVPR*, 2019.

[26] 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. *arXiv preprint arXiv:2304.02643*, 2023.

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[28] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[29] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

[30] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[32] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2021.

[33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[34] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*, 2023.

[35] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.

[36] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023.

[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*, pages 23592–23601, 2023.

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[41] 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 preprint arXiv:2303.05499*, 2023.

[42] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[44] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

[45] 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*, pages 10034–10043, 2020.

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

[47] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

[48] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[49] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European*

Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pages 792–807. Springer, 2016.

[50] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[51] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[52] 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 preprint arXiv:2306.14824*, 2023.

[53] 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 preprint arXiv:2306.14824*, 2023.

[54] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. *arXiv preprint arXiv:2311.06612*, 2023.

[55] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.

[56] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[58] 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. 2021.

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

[60] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[61] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023.

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[63] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

[64] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[65] HyperGAI Team. Hpt 1.5 air: Best open-sourced 8b multimodal llm with llama 3, 2024.

[66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[67] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[69] Haowei Wang, Jiayi Ji, Yiyi Zhou, Yongjian Wu, and Xiaoshuai Sun. Towards real-time panoptic narrative grounding by an end-to-end grounding network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2528–2536, 2023.

[70] Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2024.

[71] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

[73] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8818–8826, 2019.

[74] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[75] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

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

[77] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024.

[78] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*, 2023.

[79] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[80] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models, 2023.

[81] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[82] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024.

[83] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.

[84] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. *arXiv preprint arXiv:2402.16846*, 2024.

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

[86] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

[87] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[88] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.

[89] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.