

Towards More Unified In-context Visual Understanding

Dianmo Sheng^{1,2}, Dongdong Chen³, Zhentao Tan^{1,2}, Qiankun Liu⁴, Qi Chu^{1,2},
 Jianmin Bao³, Tao Gong^{1,2,†}, Bin Liu^{1,2}, Shengwei Xu⁵, Nenghai Yu^{1,2}

¹School of Cyber Science and Technology, University of Science and Technology of China

²Anhui Province Key Laboratory of Digital Security ³Microsoft

⁴Beijing Institute of Technology ⁵Beijing Electronic Science and Technology Institute

dmsheng@mail.ustc.edu.cn, tgong@ustc.edu.cn, xusw@besti.edu.cn

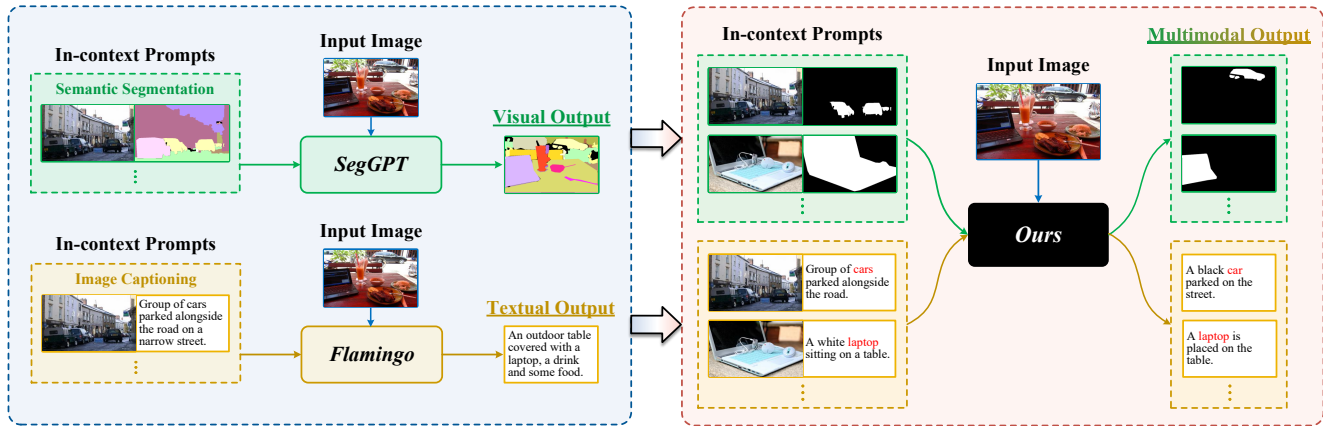


Figure 1. Motivation illustration of our method. In earlier efforts, existing in-context visual understanding models were confined to a particular output modality. For instance, SegGPT specialized in “**Image** → **Image**” applications, tailored for tasks involving image segmentation. Similarly, Flamingo was purpose-built for “**Image** → **Text**” scenarios, focusing on language-centric tasks such as image captioning. In contrast, we take a further attempt to design a unified model capable of handling multimodal in-context visual understanding tasks for “**Image** → **Image / Text**” scenarios.

Abstract

The rapid advancement of large language models (LLMs) has accelerated the emergence of in-context learning (ICL) as a cutting-edge approach in the natural language processing domain. Recently, ICL has been employed in visual understanding tasks, such as semantic segmentation and image captioning, yielding promising results. However, existing visual ICL framework can not enable producing content across multiple modalities, which limits their potential usage scenarios. To address this issue, we present a new ICL framework for visual understanding with multi-modal output enabled. First, we quantize and embed both text and visual prompt into a unified representational space, structured as interleaved in-context sequences. Then a decoder-only sparse transformer architecture is employed to perform generative modeling on them, facilitating in-context learning. Thanks to this design, the model is capa-

ble of handling in-context vision understanding tasks with multimodal output in a unified pipeline. Experimental results demonstrate that our model achieves competitive performance compared with specialized models and previous ICL baselines. Overall, our research takes a further step toward unified multimodal in-context learning.

1. Introduction

With the rapid progress of large language models, in-context learning (ICL) [5, 29, 51] has gradually become a new paradigm in the field of natural language processing (NLP). As introduced in GPT-3 [5], given language sequences as a universal interface, the model can quickly adapt to different language-centric tasks by utilizing a limited number of prompts and examples.

Some following works [1, 42] present some early attempt at applying ICL into the vision-language (VL) tasks with the design of interleaved image and text data. For example,

[†] Corresponding author.

Flamingo [1] takes the image input as a special “<image>” token to conduct the interleaved input prompt as text, and injects visual information into pre-trained LLMs with gated cross-attention dense block. It demonstrates a remarkable capability to address various vision-language tasks. However, the language-only LLM decoder design makes it only able to output text outputs.

More recently, some works start to apply the similar ICL idea into the vision-only tasks via formulating the learning goal as image inpainting [4, 46, 47]. With the well-collected multi-task vision datasets and unified grid image prompt design, these works utilize pre-trained masked image modeling models to give a perspective of what can be general-purpose task prompts in vision. For instance, SegGPT [47] studies the fundamental visual understanding problem, segmentation task, as an in-context coloring problem to achieve the in-context segmentation capability. Yet, the pre-trained vision-centric inpainting framework confines the output modality to be image only. Therefore, a straightforward question is “*How to perform in-context learning with multimodal output enabled for visual understanding in a unified framework?*”

Standing on the shoulders of predecessors, in this paper, we present the first attempt at multimodal in-context learning. The central concept aims to unify vision-language data via modality-specific quantization and shared embedding, then perform next-token prediction on the well-organized interleaved sequences of in-context samples.

In detail, we first develop detailed and comprehensive vision and language prompts, carefully designed to represent various vision understanding tasks. Then we employ modality-specific quantizers to transform the formatted in-context prompts and the visual input into discrete tokens respectively. Following this, a unified embedding layer is used to map these tokens into a shared representational space. Once the model outputs prediction tokens with specific prompts, the modality-specific decoders automatically decode them into the intended domains. This design effectively allows for multimodal input and output. To facilitate the in-context learning on unified representations, we further combine the autoregressive transformer with the Mixture of Experts (MoEs). The autoregressive transformer produces a natural contextual association based on the next-token prediction, while MoEs [13, 22] serve as a promising solution for multi-task learning by dynamically activating sub-networks without the need for task-specific modules. Following previous in-context prompts formats, we take semantic segmentation and dense captioning as the example image understanding tasks, and formatting semantic category information as the clue across multiple in-context samples. Through extensive experiments and analysis, we demonstrate that our model can facilitate in-context learning on vision understanding tasks and enable multimodal

outputs within a unified model.

2. Related Works

In-Context Learning. As the dimensions of both model size and corpus size escalate [5, 8, 10, 33], large language models (LLMs) exhibit an aptitude for in-context learning (ICL), namely, the capacity to distill knowledge from a limited array of contextual examples. GPT-3 [5], for instance, pioneers the articulation of various natural language processing (NLP) tasks as text completion conundrums, a strategy predicated on the provision of prompts and examples. This novel methodology considerably simplifies the integration of task knowledge into LLMs by modifying the demonstrations and templates, a concept substantiated by various studies [28, 48, 51].

Within the field of computer vision, the study [4] initially advances an in-context training paradigm utilizing image inpainting on illustrations and infographics derived from vision-related literature, which shows competencies in fundamental CV tasks. Additionally, the study by Painter [46] employs masked image modeling on continuous pixels to conduct in-context training with self-organized supervised datasets in seven tasks, and yields highly competitive outcomes on them. Subsequently, SegGPT [47] is a dedicated method trying to solve diverse and unlimited segmentation tasks with a similar framework. Recent studies have concentrated on how to enhance the ICL capability in vision, such as prompt selection [40] and the execution of nearest neighbor retrieval utilizing a memory bank [3].

Prior works have typically been confined to specific domains. In contrast, our study is conducted across both vision and language domains, as we aspire to realize the potential of multimodal in-context learning.

Multimodal Understanding and Generation. Multimodal understanding and generation represent an emerging frontier in artificial intelligence that seeks to interpret and synthesize information across various forms of data, such as text, images, sounds, and even more modalities. Inspired by the success of ChatGPT as well as GPT-4 [31, 32], recent works primarily concentrate on aligning visual features with the pre-trained LLMs for multimodal comprehension tasks [17, 23, 25, 26, 43, 44, 52, 55]. While pre-trained LLMs have empowered systems to follow human instructions for vision-language interactions, their application has been confined to generating textual outputs.

Expanding the horizons of multimodal capabilities, a burgeoning spectrum of studies [14, 20, 39, 41, 50, 53] are pioneering innovations in both understanding and generative capacities across modalities. IMAGEBIND [14] utilizes image-paired data to connect five different modalities with a single joint embedding space, demonstrating impressive zero-shot capabilities across these modalities. Other

wise, CoDi [41] introduces a composable generation strategy by bridging alignment in diffusion process, facilitating the synchronized generation of any combination of output modalities, including language, image, video, or audio. Furthermore, NExT-GPT [50] integrates an LLM with multi-modal adaptors and diverse diffusion decoders, enabling it to perceive inputs and generate outputs in arbitrary combinations of text, images, videos, and audio with understanding and reasoning. However, these models are not designed for in-context learning, without the benefit of multiple prompts.

Mixture of Experts models. Mixture of Experts (MoEs), which have demonstrated remarkable success in both computer vision [27, 34, 45] and natural language processing [11, 13, 21, 35, 57] with the context of conditional computation. Conditional computation aims to increase the number of model parameters without significantly increasing computational cost by selectively activating relevant parts of the model based on input-dependent factors [6, 9]. [35] first provides compelling evidence for the efficacy of MoEs by incorporating MoE layers into LSTM models. Building upon this, subsequent studies [13, 19, 22, 36] extend the application of this approach to transformer architectures.

With different routing strategies, MoE models have also been studied for multitask learning [15, 21, 56] and multi-modal learning [30, 37] as well. Recent work VL-MoE [37] is the first work to combine modality-specific MoEs with generative modeling for vision-language pretraining. In this work, we further study the potential of combining autoregressive transformer with MoE for vision-language in-context learning.

3. Method

In this section, We present a multimodal in-context framework that can seamlessly integrate the strengths of language models with the specific requirements of vision-language tasks for in-context learning. We first introduce well-organized vision-language prompts to describe foundational visual understanding tasks like segmentation and captioning (Section 3.1). After conducting the input into predefined prompts format, we quantize in-context prompts with the input pair into discrete codes using modality-specific tokenizers, and then embed them into unified representations with a general embedding network (Section 3.2). Then a decoder-only transformer with sparse MoEs is introduced to perform generative modeling on the interleaved unified representations (Section 3.3). In the following paragraph, we will elaborate on each part in detail.

3.1. Vision-Language Prompt Design

We begin by implementing unified vision-language prompts to depict different types of vision-language tasks. We

treat k in-context samples with input and output like $(i_1, o_1), \dots, (i_{k+1}, o_{k+1})$ as interleaved data, and embed them in the discrete token space. This innovative design provides the flexibility required for customizing vision or vision-language tasks according to specific needs and preferences.

Vision-Only Tasks. Following previous works, we conduct all vision-only tasks as an inpainting task. However, the inpainting is performed in token space. For every image pair that is composed of an original image and its corresponding task output, we first quantize them into discrete tokens utilizing a pre-trained image quantizer. A special tag “[BOI]” is inserted in front of each image’s token representation. Then we concatenate each pair’s visual tokens obeying the order of precedence. This structure creates a cohesive relationship between the two in-context pairs, framing them both as visual token components.

Vision-Language Tasks. For vision-language tasks, here we take the dense captioning task as an example. The prompts are clear and closely resemble those of natural language processing (NLP) tasks. Similar to existing methods [1], multiple captioning samples can be treated as interleaved image and text data. For each image, we quantize them the same way as in vision-only tasks, with the special “[BOI]” tag. For the text part, we describe the region caption with corresponding instance category and bounding box (bbox) like “*Category: <c>. Bboxes: [x₁, y₁, x₂, y₂]. Caption: <text>.*” While $P = \{x_i, y_i\}_{i=1}^N$ represents points that locate the object. $\langle text \rangle$ represents the placeholder of caption tokens. We also add a special tag “[BOT]” at the beginning of each caption. After being tokenized by looking up the vocabulary, we use a similar concatenation strategy to get the in-context token representations.

At the conclusion of each segment of in-context tokens, we incorporate an “[EOC]” tag to signify the completion of in-context samples.

3.2. Unified Multimodal Representations.

Building upon the foundation of multimodal in-context prompts discussed in Section 3.1, how to facilitate the model understanding multimodal input in a unified manner is a challenging problem. Revisiting previous vision-language models [1, 42], we decide to utilize the discrete token method as the bridge between the various input and the model embedding space. In this section, we will demonstrate the preparation for a general training recipe with multimodal in-context inputs by unifying representations based on modality-specific quantization.

Multimodal Quantization Stage. We leverage existing well-known modality-specific quantizers to encode multimodal data into discrete tokens. As illustrated in Figure 2, for image data, we adopt the vector quantizer used in VQ-

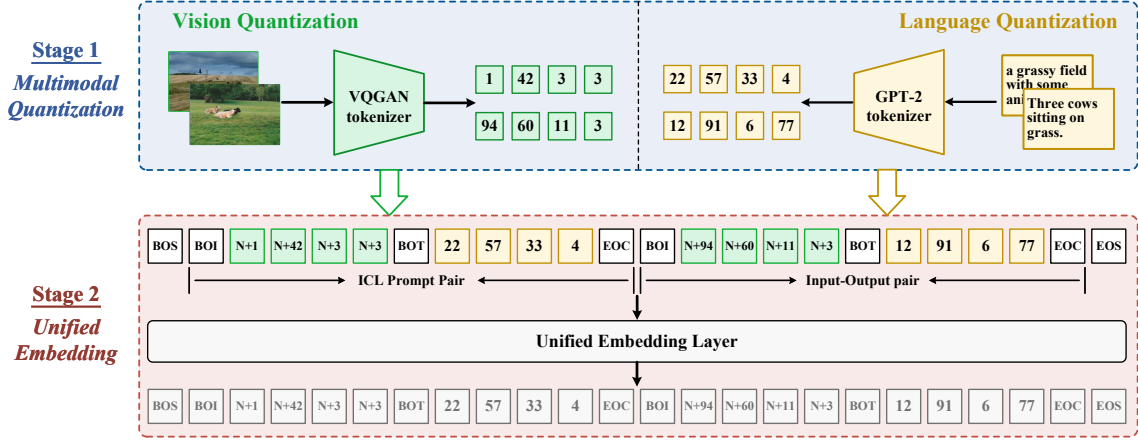


Figure 2. Overview of our unified multimodal representations pipeline with two stages. During the multimodal quantization phase, visual and linguistic inputs are encoded into discrete tokens via modality-specialized tokenizers: specifically, VQGAN’s tokenizer for visual data and GPT-2’s tokenizer for texts. After that, in the unified embedding stage, multimodal discrete tokens are formatted as an interleaved sequence with special tokens. Then a unified embedding layer projects the sequence into general representations.

GAN [12]. Given an image $x_{img} \in \mathbb{R}^{H \times W \times 3}$, the quantization step is performed by searching the nearest embedding in the learned, discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^K \subset \mathbb{R}^{n_z}$, where n_z is the codebook size, which can be formulated as:

$$z_{q,i} = \arg \min_{z_k \in \mathcal{Z}} \|E(x_{img}) - z_k\|_2. \quad (1)$$

where $z_{q,i}$ is the quantized encoding of x_{img} , and E represents for the convolution encoder. We add the visual tokens to the text vocabulary.

For the text part, the subword Byte-Pair Encoding (BPE) tokenizer in GPT-2 [33] is utilized. In the context of encoding information, BPE tokenizer quantizes x_{text} into tokens $z_{q,t}$ by looking up the vocabulary. We treat the category label c as the natural language format, with two special tags $\langle c_{st} \rangle$ and $\langle c_{ed} \rangle$ denoting the start and end of this part. Compared with the class tokens proposed in [44], category label in language offers the potential for generalization to unseen classes. For the bbox information, we adopt a similar method in [7]. After normalizing the coordinates P with 3 decimal places according to the size of the image, we map it to predefined tokens $\{\langle bin_0 \rangle, \dots, \langle bin_1000 \rangle\}$. Additional start and end tags $\langle b_{st} \rangle, \langle b_{ed} \rangle$ are placed at both ends of the bbox. Therefore, we can control the precision of coordinates with fewer tokens than the numerical representation.

Unified Embedding Stage. After quantizing each modality data into discrete tokens, we take the embedding step. Here, we treat data in both modalities equally, as all the tokens will be mapped into a unified representation embedding space by a linear layer. Then, all in-context token embeddings will be concatenated sequentially as

$(z_{q,i}^1, z_{q,t}^1), \dots, (z_{q,i}^{k+1}, z_{q,t}^{k+1})$ and fed into the model. This design offers generality and scalability for multimodal knowledge transfer. Thus, the model can handle interleaved image and text inputs like Flamingo [1].

3.3. Model Architecture and Training Objective

After the unification of various modality data, we are now going to discuss how to perform in-context learning in a general framework. We construct our model using a GPT-2 style decoder-only transformer architecture with the sparse MoEs for multimodal in-context learning. As shown in Figure 3, the overall framework is very simple and straightforward. With the interleaved input representations, we utilize next-token prediction for modeling the contextual information. The model’s predictive logits will undergo a sampling process to convert them back into tokens, which are subsequently decoded by the respective tokenizer of each modality. Consequently, the model can achieve multimodal input prompts and prediction, rather than being limited to specific output domains owing to the pre-trained backbone.

Attribute Routing MoE. Different tasks with shared parameters may conflict with each other as described in previous works [13, 56]. To mitigate the task interference issue, we utilize MoE layers, which allow different modalities and tasks to use separate parameters. For details, we replace the FFN block in each MoE decoder layer with the sparse MoE layer with N experts introduced in [34]. Following Uni-Perceiver-MoE, we adapt the attribute routing strategy for in-context tokens, and top-k gating is implemented to decide the gating decision for the embedding of each token $x \in \mathbb{R}^D$. Therefore the calculation of gating is formulated as: $\mathcal{G}(x) = \text{top}_k(\text{softmax}(W_g(x)))$, where W_g is the learnable weights of the router, and $\text{top}_k(\cdot)$ represents oper-

ator that choose the largest k values. After gating, the output of sparse MoE layer is the weighted combination of the activated experts' computation: $x_{out} = \sum_{i=1}^N \mathcal{G}(x)_i \cdot \text{FFN}_i(x)$.

Loss Function. Unlike previous vision generalists [4, 46, 47] using masked image modeling as the learning objective, we perform generative modeling on interleaved in-context representations like Flamingo [1], benefiting from the natural context understanding by leveraging next token prediction.

The cross-entropy loss is employed on the output tokens of each in-context pair as well as the input pair, which constrains the similarity between model predictions \mathcal{P}_{pred} and ground-truth tokens \mathcal{P}_{gt} , represented as:

$$\mathcal{L}_{out} = \sum_{i=1}^{k+1} \text{CE}(\mathcal{P}_{pred}^i, \mathcal{P}_{gt}^i) \quad (2)$$

We also utilize the auxiliary loss introduced in GShard [22] to optimize the gating network of MoEs, and the whole loss function can be represented as:

$$\mathcal{L} = \mathcal{L}_{out} + \lambda \cdot \mathcal{L}_{aux} \quad (3)$$

where λ is the weight of auxiliary loss.

4. Experiments

4.1. Datasets and Benchmarks.

Prior works in visual in-context learning predominantly aimed to integrate concepts from NLP into conventional visual tasks. As detailed in MAE-VQGAN [4], Painter [46] and SegGPT [47], each task involves creating a grid-structured image. However, these approaches overlook task-specific comprehension, merging all tasks into a singular prompt. Consequently, we propose a redefined approach to traditional visual tasks with semantic clues, emphasizing vision-language understanding tasks such as semantic segmentation and image captioning, which are named class-aware in-context (short for CA-ICL) segmentation and captioning respectively.

CA-ICL Segmentation. As depicted in Figure 4, for segmenting instances of a particular class, each in-context sample is provided solely with the desired class segmentation mask. We conduct the data with the entire MS-COCO dataset, which contains 80 object classes. For each category, a mask pool is built for in-context sampling. Finally, we collect about 350k class masks for training and 15k class masks for validation. **Evaluation Metric:** We take the conventional semantic segmentation metric Mean Intersection over Union (MIoU) for evaluation. Given that the output is a binary mask, we also present the Mean Absolute Error (MAE) scores.

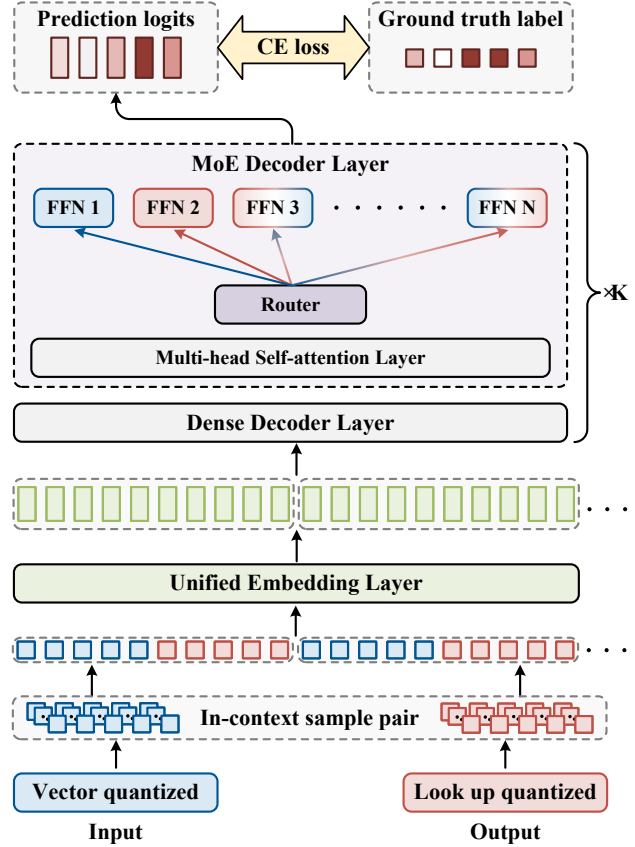



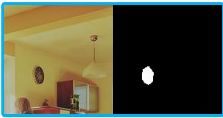




Figure 3. Overview of our pipeline. Here, we take the CA-ICL captioning task as an example. Multiple in-context samples and the input pair are first tokenized using modality-specific tokenizers and then projected into unified embedding representations. After undergoing interleaved concatenation, the tokens are inputted into the model for generative modeling.

CA-ICL Captioning. For the CA-ICL captioning, we also take the class information as the in-context clue, with each in-context sample containing the caption for the desired category. Here, we use the Visual Genome dataset, from which each image has multiple annotations, including object labels and caption annotations for each region of the image. We selectively use categories that correspond with those in the MS-COCO dataset, ensuring that each class has more than 100 descriptions. Finally, we collected about 460k region descriptions for training and 2k region descriptions for the test set. **Evaluation Metric:** Captioning performance is assessed using the BLEU4, METEOR, and CIDEr metrics, which are standard in image captioning tasks. When incorporating bbox information in prompts, we also present the mean Average Precision (mAP) metric following [18]. By filtering the prediction with predefined thresholds on IoU and METEOR, the average of the APs obtained for all pairwise combinations of the two thresholds to evaluate both localization and description accuracy.

Class-aware In-context Segmentation Task

Class	Task prompt	Input	Output
Person			
			

Class-aware In-context Captioning Task







Class	Task prompt	Input	Output
Banana	 a hand holding bananas .		
Bicycle	 woman on a bicycle looking at bus.		

Figure 4. Class-aware in-context understanding task definitions. For the sake of easy demonstration, only one in-context sample is used here. The blue boxes \square on the left display the inputs of the model, while the red boxes \square on the right show the corresponding output. (In the absence of additional clarification, subsequent notations convey the same meaning.)

4.2. Implementation Details.

For image tokenizer, we adopt VQ-GAN tokenizer [12] with a vocabulary size of 1024 and 16x downsampling ratio, which is pre-trained on Imagenet dataset. The input image resolution is set to 256×256 , leading to 256 tokens after quantization. For text tokenizer, we employ GPT-2 BPE tokenizer [33] with a vocabulary size of 50257. We implement our model with GPT-small model architecture while replacing the FFN in part of the decoder layers with attribute routing MoEs introduced in [56]. Please refer to the supplementary for detailed architecture hyperparameters.

During each training iteration, the number of in-context samples is set to 3 by default. All parameters are trained from scratch. The weight λ is set to 0.02. For optimization, we employ the AdamW algorithm with a base learning rate of $1e-4$, complemented by a weight decay of 0.05. We utilize gradient clipping at a value of 0.5 to stabilize the training process, ensuring consistent performance throughout. Unless otherwise specified, the training runs for 40 epochs with a batch size of 512 on 8 NVIDIA A6000 GPUs.

diverse sizes	large scale	MIoU \uparrow	MAE \downarrow
\times	\times	31.82	0.176
\checkmark	\times	33.54	0.172
\times	\checkmark	42.87	0.133
\checkmark	\checkmark	45.04	0.128

Table 1. Ablation of object size and scale in class-aware in-context segmentation task. Regarding object size, we adopt the MS-COCO definition, for whether to include small instances with an object area less than 32^2 square units. For object scale considerations, the crop region is taken into account. The highlighted row indicates the best choice. (In the absence of additional clarification, subsequent notations convey the same meaning.)

bbox_image	bbox_text	B@4 \uparrow	CIDEr \uparrow
\times	\times	7.9	104.4
\checkmark	\times	0.0	2.7
\times	\checkmark	7.8	112.0

Table 2. Ablation study on the impact of bbox information in class-aware in-context caption task. “bbox_image” and “bbox_text” indicate that the bounding box is in image type or in text format.

4.3. Ablation Studies



In this section, we conduct an ablation study of our method from three perspectives: task definition, model definition, and multi-task co-training strategy. Without additional statements, the experiments are conducted using images in 128 resolution with 20 epochs of training.

Class-aware In-context Task Definitions. In our exploration of two proposed in-context learning tasks, we rigorously examine the task definitions. As demonstrated in Table 1, we investigate the object size and scale within each in-context sample for the CA-ICL segmentation task. Our findings indicate that including small objects with a large object scale yields optimal results. We surmise that objects spanning multiple scales offer more detailed insights and salient in-context samples lead to a richer diversity of information, which is beneficial for segmentation.

In our research on CA-ICL captioning, We explore the correlation between in-context input images and their corresponding descriptions. We drew inspiration from dense captioning and visual grounding, examining if incorporating object location information is beneficial for the model to capture semantic cues conveyed by in-context samples.

As evidenced in Table 2, introducing an image-type output leads to a notable decline in performance compared to the baseline. To tackle this issue, we explored the method of encoding bbox information in a textual format, as outlined in Section 3.1. While the results were considerably

Class-aware In-context Captioning task

Class	Task prompt	Input	Output
apple	a red apple next to an orange.		apples on the ground.
backpack	a backpack on a back		a man with a black backpack.

Class-aware In-context Captioning task with bbox

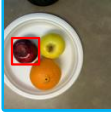
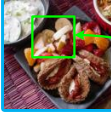
Class	Task prompt	Input	Output
apple	a red apple next to an orange.		sliced apples on the plate.
backpack	a backpack on a back		a man with a black backpack over his shoulder.

Figure 5. Analysis of the impact of including bbox information. For better visualization, the ground truth bboxes are indicated by rose boxes \square , while the predicted bboxes are highlighted in green boxes \square . With the bbox information in prompts, the model yields more precise descriptions that are aligned with the specified region locations.

better than the “bbox_image” approach, even outperformed the baseline in CIDEr metric. Figure 5 demonstrates that using prompts of the “bbox_text” type leads to more precise predicted captions that correspond with the intended region. This alignment significantly aids in the accurate and convenient verification of the model’s performance during testing phases. This evidence supports the model’s capability to effectively generate class-aware captions when supplied with appropriate examples.

Model Variants Definition. We conducted experiments using various model configurations at a higher resolution of 256 to identify the optimal choice. The reference for these experiments is the single task performance, with the baseline established as task co-training using the standard GPT-2 small architecture, referred to as “all tasks”. We replace the FFN in part of transformer blocks with the MoE layer proposed in [22] and the AG_MoE introduced in [56] for analysis. The results presented in Table 3 reveal that the baseline setting results in significant unbalanced performance with a sharp segmentation performance decrease, while models with MoE configurations surpass the baseline in segmentation performance by 18.74 scores, yet there remains a notable shortfall of 10.8 scores in captioning per-

Model	CA-ICL segmentation	CA-ICL captioning
	MIoU \uparrow	CIDEr \uparrow
single task	51.91	88.6
all tasks	21.74	77.3
w/ MoE	40.48 (+18.74)	66.5 (-10.8)
w/ AG_MoE	42.02 (+20.28)	67.9 (-9.4)
w/ MT	33.72 (+11.98)	81.1 (+3.8)
w/ AG_MoE and MT	49.91 (+28.17)	78.3 (+1.0)

Table 3. Ablation of model variants and multi-task learning strategy. We present the MIoU and CIDEr metrics for CA-ICL segmentation and captioning tasks, respectively. In the brackets, we analyze gaps compared to the “all tasks” setting. We use green and red to indicate the performance decreases and increases.

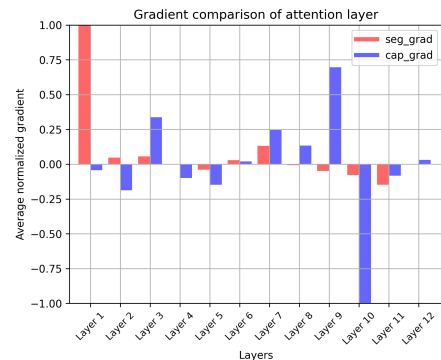


Figure 6. Gradient comparison of CA-ICL tasks. We utilize the normalized average gradient of each attention layer for comparison, while the symbol and the value represent the direction and magnitude of the gradient respectively.

formance. The adoption of the AG_MoE structure further narrows this performance gap. Considering the image tokens dominate compared with text tokens and the divergent gradient directions of differing task complexities (as shown in Figure 6), the caption performance drops. Models with shared parameters might struggle to effectively manage the significant difference in token representations between the two tasks, highlighting the advantages of MoEs. In the following section, we will address the challenges associated with multi-task co-training.

Multi-task Co-training Strategy. In this section, we explore the impact of multi-task joint training. As demonstrated in Table 3, employing the standard GPT-2 small architecture for co-training results in significant performance degradation, suggesting a considerable disparity in handling tasks involving different data modalities. The implementation of the AG_MoE architecture results in a more balanced performance across tasks, yet there remains a notable performance gap compared to single-task scenarios.

To further enhance the performance of the model with AG_MoE, we adopt a multi-task learning paradigm to alle-

Models	Resolution	#Trainable Params	CA-ICL Segmentation		CA-ICL Captioning			
			MIoU \uparrow	MAE \downarrow	B@4 \uparrow	METEOR \uparrow	CIDEr \uparrow	mAP \uparrow
specialist model								
FPTrans [54]	480	139M	43.30	0.202	-	-	-	-
VAT [16]	417	27M	46.07	0.087	-	-	-	-
DCAMA [38]	384	89M	53.06	<u>0.059</u>	-	-	-	-
GRiT [49]	1024	197M	-	-	5.2	9.0	58.6	<u>15.9</u>
generalist model								
SegGPT [47]	448	307M	<u>62.83</u>	0.092	-	-	-	-
SegGPT*	256	307M	51.12	0.116	-	-	-	-
OpenFlamingo [2]	224	3B	-	-	4.6	11.4	61.3	-
Ours	256	309M	58.04	0.110	5.3	14.3	86.9	10.9

Table 4. Comparison with state-of-the-art specialist and generalist models on class-aware in-context task. We report both MIoU and MAE scores for comparison. * indicates that we test the SegGPT with images in 256 resolution. Previous state-of-the-art results are underlined.

viate the task interference problems and, meanwhile, stabilize the training of MoEs. Drawing inspiration from Uni-Perceiver v2 [24], we utilize their unmixed batch sampling strategy and correlative optimizer. Here, the sampling weight s_k of each dataset is configured to be proportional to the square root of the dataset’s size. For the scaling factor w_k , we uniformly assign a value of 1 to all tasks. As evidenced in Table 3, the integration of the AG_MoE architecture with our multi-task learning strategy results in performance that exceeds the baseline for both tasks. This is particularly notable in the CA-ICL segmentation task, where an impressive gain of 28.17 points is observed. This indicates that the multi-task strategy effectively prevents potential task conflicts within a batch.

4.4. Comparison with State-of-the-art Methods

We experimented with class-aware in-context tasks to compare with existing state-of-the-art specialist models as well as generalists. For the task definition, we adopt the best settings as discussed in ablations (Section 4.3). For the model and training strategy, we utilize AG_MoE architecture with the multi-task learning strategy.

For CA-ICL segmentation, we compare with generalist segmentation model SegGPT [47] and specialist few-shot segmentation models like FRTrans [54], VAT [16] and DCAMA [38]. As indicated in Table 4, our model trained at a resolution of 256 surpasses SegGPT that evaluated at the same resolution—an improvement of 6.92 in MIoU and 0.006 in the MAE score. However, still a gap between the 448 version of SegGPT with more training data and higher resolution input. The performance is also notably comparable to the state-of-the-art specialist DCAMA, which operates at a higher resolution of 384 as well.

In the domain of CA-ICL captioning, the generalist baseline for evaluation is Openflamingo [2], a large vision-

language model that excels in demonstrating strong in-context captioning ability. For CA-ICL captioning, its closest counterpart is dense captioning, as both tasks necessitate the prediction of the caption and the corresponding bbox. Therefore, we compare with the sota dense captioning model GRiT [49]. We utilize the images in our test set to evaluate GRiT. Then allocate the generated predictions to our ground-truth regions annotations, utilizing the IoU metric of their respective bboxes as the basis for the assignment. As shown in Table 4, our method achieves state-of-the-art performance in traditional image captioning metrics. In comparison to Openflamingo, which has a parameter tenfold greater, we also achieve a 0.7-point increase in BLEU4 and a significant 25.6-point improvement in CIDEr. However, the result still has a gap in the mAP score compared with GRiT, which utilized a foreground object extractor.

5. Conclusion

In this work, we present a unified framework for in-context visual understanding. By leveraging multimodal quantization and unified embedding, our model is capable of jointly learning multimodal data in the general token embedding space. By further synergistically integrating autoregressive transformer with the MoEs framework, we achieve stable multi-task co-training while simultaneously benefiting from the balanced contributions of each task. Overall, our research showcases the potential of in-context learning across various modalities as well as tasks.

6. Acknowledgment

This work was partially supported by the Anhui Provincial Science and Technology Major Project (No.2023z020006) and National Natural Science Foundation of China (No.62121002).

References

- [1] 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. [1](#), [2](#), [3](#), [4](#), [5](#)
- [2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. [8](#)
- [3] Ivana Balažević, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier J Hénaff. Towards in-context scene understanding. *arXiv preprint arXiv:2306.01667*, 2023. [2](#)
- [4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. [2](#), [5](#)
- [5] 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. [1](#), [2](#)
- [6] Ke Chen, Lei Xu, and Huisheng Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural networks*, 12(9):1229–1252, 1999. [3](#)
- [7] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. [4](#)
- [8] 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. *arXiv preprint arXiv:2204.02311*, 2022. [2](#)
- [9] Andrew Davis and Itamar Arel. Low-rank approximations for conditional feedforward computation in deep neural networks. *arXiv preprint arXiv:1312.4461*, 2013. [3](#)
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [2](#)
- [11] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022. [3](#)
- [12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. [4](#), [6](#)
- [13] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022. [2](#), [3](#), [4](#)
- [14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. [2](#)
- [15] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347, 2021. [3](#)
- [16] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. [8](#)
- [17] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. [2](#)
- [18] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. [5](#)
- [19] Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. Scalable and efficient moe training for multi-task multilingual models. *arXiv preprint arXiv:2109.10465*, 2021. [3](#)
- [20] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023. [2](#)
- [21] Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*, 2021. [3](#)
- [22] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. [2](#), [3](#), [5](#), [7](#)
- [23] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. [2](#)
- [24] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhui Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2700, 2023. [8](#)

- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [2](#)
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. [2](#)
- [27] Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. Cross-token modeling with conditional computation. *arXiv preprint arXiv:2109.02008*, 2021. [3](#)
- [28] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021. [2](#)
- [29] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021. [1](#)
- [30] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with lmoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022. [3](#)
- [31] OpenAI. Introducing chatgpt. 2022. [2](#)
- [32] OpenAI. Gpt-4 technical report. 2023. [2](#)
- [33] 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. [2](#), [4](#), [6](#)
- [34] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. [3](#), [4](#)
- [35] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. [3](#)
- [36] Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, et al. Mesh-tensorflow: Deep learning for supercomputers. *Advances in neural information processing systems*, 31, 2018. [3](#)
- [37] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023. [3](#)
- [38] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, pages 151–168. Springer, 2022. [8](#)
- [39] Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. [2](#)
- [40] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*, 2023. [2](#)
- [41] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023. [2](#), [3](#)
- [42] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. [1](#), [3](#)
- [43] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. [2](#)
- [44] Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. [2](#), [4](#)
- [45] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562. PMLR, 2020. [3](#)
- [46] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. [2](#), [5](#)
- [47] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. [2](#), [5](#), [8](#)
- [48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. [2](#)
- [49] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. [8](#)
- [50] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. [2](#), [3](#)
- [51] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning. *arXiv preprint arXiv:2212.10375*, 2022. [1](#), [2](#)
- [52] 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. [2](#)
- [53] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian

- Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. [2](#)
- [54] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *Advances in Neural Information Processing Systems*, 35:6575–6588, 2022. [8](#)
- [55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [2](#)
- [56] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *Advances in Neural Information Processing Systems*, 35:2664–2678, 2022. [3](#), [4](#), [6](#), [7](#)
- [57] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. [3](#)