西北工业大学

数字图像处理—论文翻译

原论文标题：Towards More Unified In-context Visual Understanding

蔡啟健

计算机学院

计算机科学与技术

2024 年 11 月

学号：2021302706

Northwestern Polytechnical University

Towards More Unified In-context Visual Understanding

Dianmo Sheng[1], Dongdong Chen[2], Zhentao Tan[1], Qiankun Liu[3], Qi Chu[1],
Jianmin Bao[2], Tao Gong[1,*], Bin Liu[1], Shengwei Xu[4], Nenghai Yu[1]
[1]School of Cyber Science and Technology, University of Science and Technology of China
Anhui Province Key Laboratory of Digital Security
the CAS Key Laboratory of Electromagnetic Space Information [2]Microsoft Research
[3]Beijing Institute of Technology [4]Beijing Electronic Science and Technology Institute

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 capable 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, 30, 52] 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, 43] present some early attempt at applying ICL into the vision-language (VL) tasks with the design of interleaved image and text data. For example, 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, 47, 48]. 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 [48] 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 fa-

*Tao Gong is the corresponding author

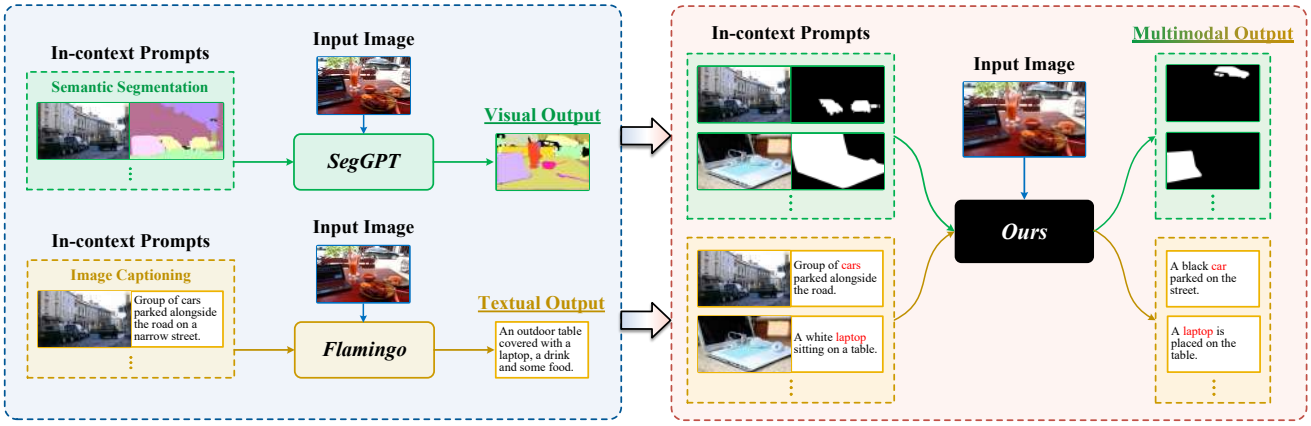

Figure 1. Motivation illustration of our method. In earlier efforts, existing in-context visual understanding models were confined to a particular output modality. For instance, SegGPT specialized in "**Image** → **Image**" applications, tailored for tasks involving image segmentation. Similarly, Flamingo 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.

cilitate 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 [14, 23] 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, 34], 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 [29, 49, 52].

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 [47] 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 [48] 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 [41] 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 [32, 33], recent works primarily concentrate on aligning visual features with the pre-trained LLMS for multimodal comprehension tasks [18, 24, 26, 27, 44, 45, 53, 57]. 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 [15, 21, 40, 42, 51, 54] are pioneering innovations in both understanding and generative capacities across modalities. IMAGEBIND [15] utilizes the image-paired data to connect five different modalities with a single joint embedding space, demonstrating impressive zero-shot capabilities across these modalities. Oth-

erwise, CoDi [42] introduces a composable generation strategy by bridging alignment in the diffusion process, facilitating the synchronized generation of any combination of output modalities, including language, image, video, or audio. Furthermore, NExT-GPT [51] integrates an LLM with multimodal 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 the multiple prompts.

Mixture of Experts models. Mixture of Experts (MoEs), which have demonstrated remarkable success in both computer vision [28, 35, 46] and natural language processing [11, 14, 22, 36, 59] 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]. [36] first provides compelling evidence for the efficacy of MoEs by incorporating MoE layers into LSTM models. Building upon this, subsequent studies [14, 20, 23, 37] extend the application of this approach to transformer architectures.

With different routing strategies, MoE models have also been studied for multitask learning [16, 22, 58] and multimodal learning [31, 38] as well. Recent work VL-MoE [38] 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 "$(i_1, o_1), \cdots, (i_{k+1}, o_{k+1})$" 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:* $[x_1, y_1, x_2, y_2]$. *Caption:* $<text>$." While $P = \{x_i, y_i\}_{i=1}^{N}$ represents points that locate the object. $<text>$ 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, 43], 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-

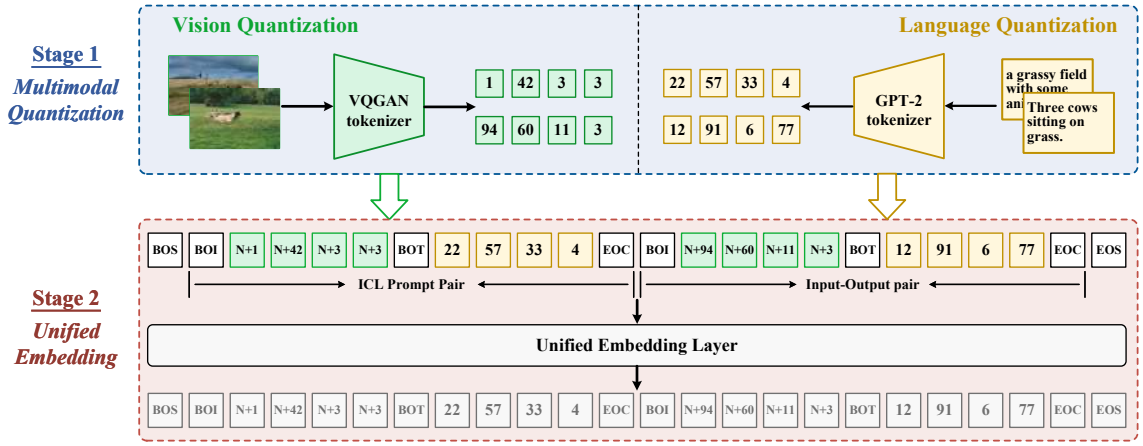

Figure 2. Overview of our unified multimodal representations pipeline with two stages. During the multimodal quantization phase, visual and linguistic inputs are encoded into discrete tokens via modality-specialized tokenizers: specifically, VQGAN's tokenizer for visual data and GPT-2's tokenizer for 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 [13]. 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:

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

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 [34] 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 <c_st> and <c_ed> denoting the start and end of this part. Compared with the class tokens proposed in [45], 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 $\{<\text{bin\_0}>, \cdots, <\text{bin\_1000}>\}$. Additional start and end tags <b_st>, <b_ed> 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 "$(z^1_{q\_i}, z^1_{q\_t}), \cdots, (z^{k+1}_{q\_i}, z^{k+1}_{q\_t})$" 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 [14, 58]. 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 [35]. 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: $x_{out} = \sum_{i=1}^{N} \mathcal{G}(x)_i \cdot \mathrm{FFN}_i(x)$.

Loss Function. Unlike previous vision generalists [4, 47, 48] 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:

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

We also utilize the auxiliary loss introduced in GShard [23] to optimize the gating network of MoEs, and the whole loss function can be represented as:

$$\mathcal{L} = \mathcal{L}_{out} + \lambda \cdot \mathcal{L}_{aux} \tag{3}$$

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 [47] and SegGPT [48], 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 Segmentaion. 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.

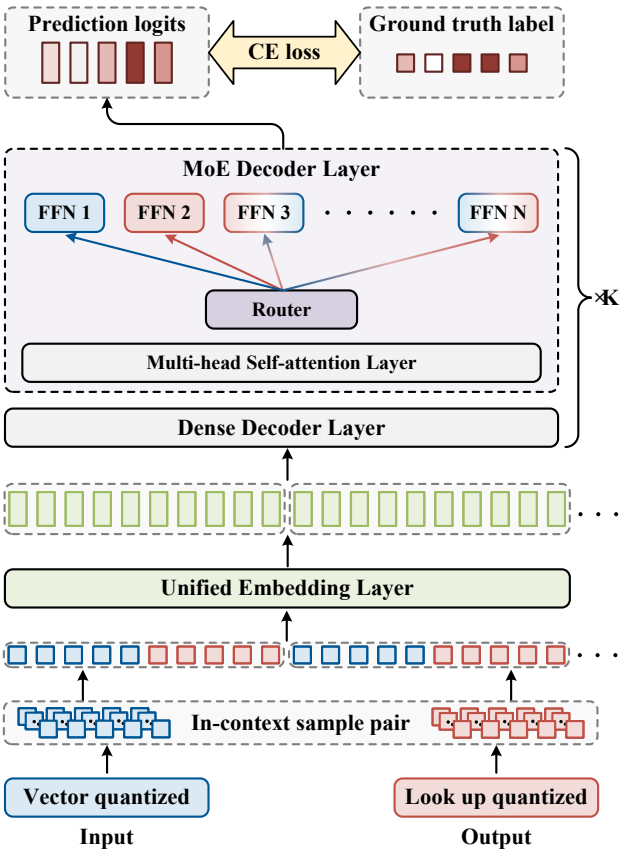

Figure 3. Overview of our pipeline. Here, we take the CA-ICL captioning task as an example. Multiple in-context samples and the input pair are first tokenized using modality-specific tokenizers and then projected into unified embedding representations. After undergoing interleaved concatenation, the 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 [19]. 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.

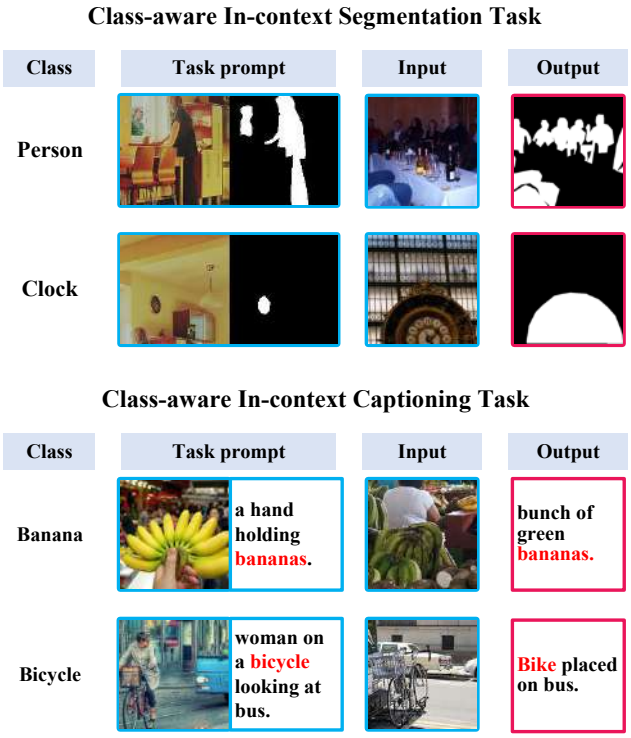

Figure 4. Class-aware in-context understanding task definitions. For the sake of easy demonstration, only one in-context sample is used here. The blue boxes □ on the left display the inputs of the model, while the red boxes □ on the right show the corresponding output. (In the absence of additional clarification,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

4.2. Implementation Details.

For the image tokenizer, we adopt VQ-GAN tokenizer [13] with a vocabulary size of 1024 and 16x downsampling ratio, which is pre-trained on the Imagenet dataset. The input image resolution is set to 256×256, leading to 256 tokens after quantization. For the text tokenizer, we employ GPT-2 BPE tokenizer [34] 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 [58]. 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 ↑	MAE ↓
✗	✗	31.82	0.176
✓	✗	33.54	0.172
✗	✓	42.87	0.133
✓	✓	**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 ↑	CIDEr ↑
✗	✗	**7.9**	104.4
✓	✗	0.0	2.7
✗	✓	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

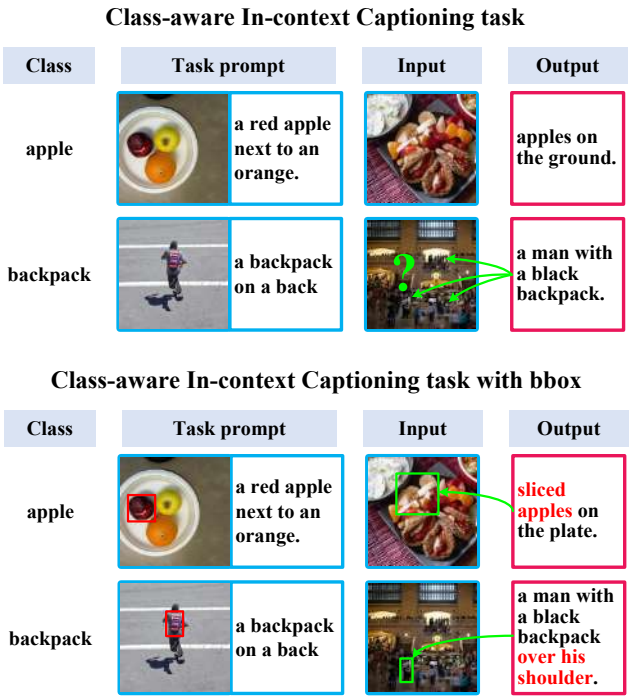

Figure 5. Analysis of the impact of including bbox information. For better visualization, the ground truth bboxes are indicated by rose boxes □, while the predicted bboxes are highlighted in green boxes □. With the bbox information in prompts, the model yields more precise descriptions that are aligned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

Model	CA-ICL segmentation	CA-ICL captioning
	MIoU ↑	CIDEr ↑
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.

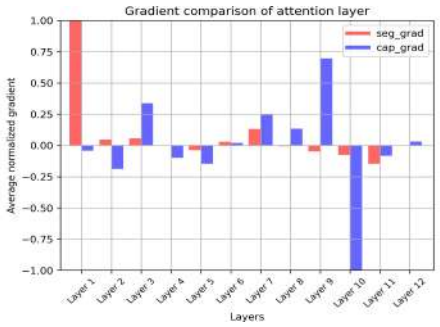

Figure 6. Gradient comparison of CA-ICL tasks. We utilize the normalized average gradient of each attention layer for comparison, while the symbol and the value represent the direction and magnitude of the gradient respectively.

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 [23] and the AG_MoE introduced in [58] 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 performance. 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 alleviate the task interference problems and, meanwhile, stabilize the training of MoEs. Drawing inspiration from

Models	Resolution	#Trainable Params	CA-ICL Segmentation		CA-ICL Captioning			
			MIoU ↑	MAE ↓	B@4 ↑	METEOR ↑	CIDEr ↑	mAP ↑
specialist model								
FPTrans [55]	480	139M	43.30	0.202	-	-	-	-
VAT [17]	417	27M	46.07	0.087	-	-	-	-
DCAMA [39]	384	89M	53.06	0.059	-	-	-	-
GRiT [50]	1024	197M	-	-	5.2	9.0	58.6	15.9
generalist model								
SegGPT [48]	448	307M	62.83	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 the MIoU and MAE scores for comparison. * indicates that we test the SegGPT with images in 256 resolution. The previous state-of-the-art results are underlined.

Uni-Perceiver v2 [25], 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 [48] and specialist few-shot segmentation models like FRTrans [55], VAT [17] and DCAMA [39]. 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. The CA-ICL captioning task most analogous to it is that of dense captioning, as both tasks necessitate the prediction of not only the caption but also the corresponding bbox. Therefore, we compare with the sota dense captioning model GRiT [50]. 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. We believe this is because they utilize 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 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.

向更统一的上下文视觉理解

摘要

大型语言模型（LLM）的快速发展加速了上下文学习（ICL）作为自然语言处理领域的一种前沿方法的出现。近年来，ICL 被应用于语义分割和图像字幕等视觉理解任务，取得了很好的效果。然而，现有的可视化 ICL 框架不能支持跨多种模式生成内容，这限制了它们潜在的使用场景。为了解决这个问题，我们提出了一个新的 ICL 框架，用于支持多模态输出的可视化理解。首先，我们将文本和视觉提示量化并嵌入到一个统一的表征空间中，结构为交错的上下文序列。然后采用仅解码器的稀疏转换器架构对其进行生成建模，促进上下文学习。多亏了这种设计，该模型能够通过统一管道中的多模态输出来处理上下文视觉理解任务。实验结果表明，与专门的模型和以往的 ICL 基线相比，我们的模型具有良好的性能。总的来说，我们的研究朝着统一多模态情境学习进一步。

1.引言

随着大型语言模型的快速发展，上下文学习（ICL）[5,30,52]逐渐成为自然语言处理（NLP）领域的一种新范式。正如 GPT-3 [5]中所介绍的那样，将给定的语言序列作为一个通用接口，该模型可以通过使用有限数量的提示和示例来快速适应不同的以语言为中心的任务。下面的一些作品[1,43]提出了一些早期尝试将 ICL 应用到视觉语言（VL）任务中，设计交错的图像和文本数据。例如，火烈鸟[1]将图像输入作为一个特殊的“<image>”标记，将交错输入提示作为文本进行，并将视觉信息注入到具有门控交叉注意密集块的预先训练过的 llm 中。它展示了一种处理各种视觉语言任务的非凡能力。然而，仅使用语言的 LLM 解码器设计使它只能输出文本输出。

最近，一些作品开始将类似的 ICL 思想应用到仅限视觉的任务中，通过将学习目标制定为图像内绘制[4,47,48]。通过收集良好的多任务视觉数据集和统一的网格图像提示设计，这些工作利用预先训练好的掩蔽图像建模模型来提供一个关于视觉中通用任务提示的视角。通过收集良好的多任务视觉数据集和统一的网格图像提示设计，这些工作利用预先训练好的掩蔽图像建模模型来提供一个关于视觉中通用任务提示的视角。例如，SegGPT [48]研究了基本的视觉理解问题，即分割任务，作为一个上下文着色问题，以实现上下文分割能力。然而，预先训练的以视觉为中心的内画框架将输出模态限制为仅限图像。因此，一个简单的问题是“如何在一个统一的框架中使用多模态输出的视觉理解来执行上下文学习？”

站在前人的肩膀上，在本文中，我们提出了多模态语境学习的第一次尝试。中心概念的目的是通过特定模式的量化和共享嵌入来统一视觉语言数据，然后对组织良好的上下文样本的交错序列进行下一个令牌预测。

详细地说，我们首先开发了详细和全面的视觉和语言提示，精心设计，以代表各种视觉理解任务。用来代表各种视觉理解任务。然后，我们使用特定于模态的量化器，分别将格式化的上下文提示和视觉输入转换为离散的标记。之后，使用统一的嵌入层将这些标记映射到共享的表征空间中。一旦模型输出带有特定提示的预测标记，特定于模式的解码器就会自动将它们解码到预期的域中。这种设计有效地允许了多模态的输入和输出。为了便于对统一表示的上下文学习，我们进一步将自回归变压器与专家混合器（MoEs）结合起来。自回归转换器基于连接标记预测产生自然的自然上下文关联，而 MoEs [14,23]是一种很有前途的多任务学习解决方案，通过动态激活子网络，而不需要特定任务的模块。按照之前的上下文提示格式，我们将语义分割和密集字幕作为示例图像理解任务，并将语义类别信息作为多个语义样本的线索。通过广泛的实验和分析，我们证明了我们的模型可以促进在视觉理解任务上的

上下文学习，并在一个统一的模型中实现多模态输出。

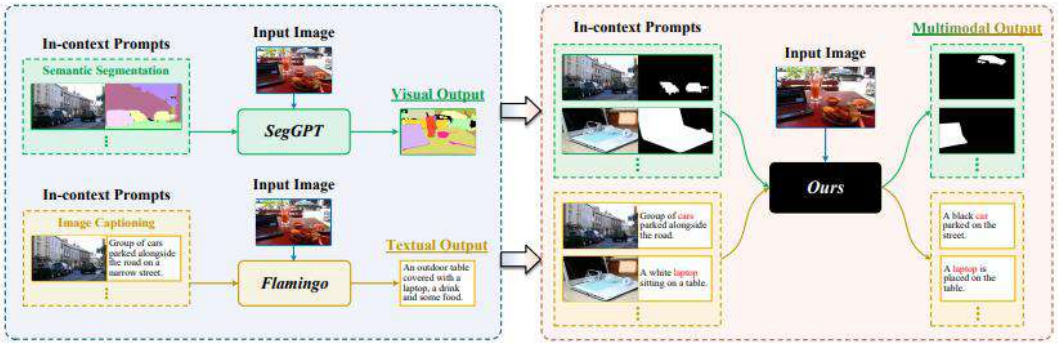

图 1.我们的方法的动机说明。在早期的工作中，现有的上下文视觉理解模型仅限于一个特定的输出模态。例如，SegGPT 专门从事“image→图像”应用程序，为涉及图像分割的任务量身定制。类似地，火烈鸟也是专门为“图像→Text”场景而设计的，专注于以语言为中心的任务，如图像字幕。相比之下，我们进一步尝试设计一个统一的模型，能够处理“Image →图像/文本”场景的多模态上下文视觉理解任务。

2. 相关工作

上下文学习。随着模型大小和语料库大小的维度不断升级，大型语言模型（LLMs）表现出了上下文学习（ICL）的能力，即从有限的上下文示例中提取知识的能力。例如，GPT-3 [5]率先将各种自然语言处理（NLP）任务作为文本完成难题，这是一种基于提供提示和示例的策略。这种新的方法通过修改演示和模板，大大简化了将任务知识集成到 llm 中的过程，这一概念已被各种研究[29,49,52]证实。

在计算机视觉领域，[4]研究最初提出了一种上下文训练范式，利用来自视觉相关文献的图像插图，显示了基本 CV 任务的能力。此外，Painter [47]的研究采用了连续像素上的掩蔽图像建模，在 7 个任务中使用自组织监督数据集进行上下文训练，并在它们上产生了高度竞争的结果。随后，SegGPT [48]是一种专门的方法，试图解决在类似的框架下的多样化和无限的分割任务。最近的研究主要集中在如何提高 ICL 的视觉能力，如提示选择[41]和利用记忆库[3]执行最近邻检索。

以前的工作通常都局限于特定的领域。相比之下，我们的研究是在视觉和语言领域进行的，因为我们渴望实现多模态情境学习的潜力。

多模式的理解和生成。多模态理解和生成代表了人工智能的一个新兴前沿，它试图通过各种形式的数据来解释和合成信息，如文本、图像、声音，甚至更多的模式。受 ChatGPT 和 GPT-4 [32,33]成功的启发，最近的工作主要集中在将视觉特征与预先训练好的 LLMS 进行多模态理解任务[18,24,26,27,44,45,53,57]的对齐上。虽然预先训练过的 llm 已经授权系统能够遵循视觉-语言交互的人类指令，但它们的应用程序仅限于生成文本输出。

扩展多模式能力的视野，新兴的研究范围[15,21,40,42,51,54]在跨模式的理解和生成能力方面开创创新。图像绑定[15]利用图像配对数据将五种不同的模式与一个单一的关节嵌入空间连接起来，在这些模式中展示了令人印象深刻的零射击能力。否则，CoDi [42]引入了一种可组合的生成策略，通过连接扩散过程中的对齐，促进了任何输出模式组合的同步生成，包括语言、图像、视频或音频。此外，NExT-GPT [51]集成了 LLM 与多模态适配器和不同的扩散解码器，使其能够通过理解和推理感知文本、图像、视频和音频的任意组合中的输入并生成输出。

然而，这些模型并不是为上下文学习而设计的，因为没有多个提示的好处。

专家模型的混合。专家混合物（MoEs）在计算机视觉[28,35,46]和自然语言处理[11,14,22,36,59]方面都取得了显著的成功。条件计算的目的是通过基于输入依赖因子[6,9]选择性地激活模型

的相关部分，在不显著增加计算成本的情况下增加模型参数的数量。[36]首先通过将 MoE 层纳入到 LSTM 模型中，为 MoEs 的有效性提供了令人信服的证据。在此基础上，随后的研究[14,20,23,37]将这种方法的应用扩展到变压器体系结构中。

在不同的路由策略下，我们还研究了多任务学习[16,22,58]和多模态学习[31,38]的 MoE 模型。最近的工作 VL-MoE [38]是第一个将特定于模态的 MoEs 与生成建模相结合的视觉语言预训练的工作。在这项工作中，我们进一步研究了结合自回归变换与 MoE 在视觉语言上下文学习中的潜力。

3. 方法

在本节中，我们提出了一个多模态的上下文内框架，它可以无缝地将语言模型的优势与上下文内学习的视觉-语言任务的具体需求集成起来。我们首先介绍了组织良好的视觉语言提示来描述基本的视觉理解任务，如分割和字幕（第 3.1 节）。在将输入转换为预定义的提示格式之后，我们使用特定模式的标记器将输入对的上下文中的提示量化为离散代码，然后用通用嵌入网络嵌入到统一表示中（第 3.2 节）。然后引入一种带有稀疏 MoEs 的仅解码器变压器，对交错的统一表示进行生成建模（第 3.3 节）。在下一段中，我们将详细阐述每一个部分。

3.1. 视觉语言提示式设计

我们首先实现统一的视觉语言提示来描述不同类型的视觉语言任务。我们将 k 个具有输入和输出的上下文样本，如“（i_1，o_1），・・・，（i_{k+1}，o_{k+1}）”作为交错数据，并将它们嵌入到离散标记空间中。这种创新的设计提供了根据特定的需求和偏好来定制视觉或视觉语言任务所需的灵活性。

仅限于视觉的任务。在之前的工作之后，我们将所有的视觉任务作为绘画任务。但是，内画是在标记空间中执行的。对于每一个由原始图像及其相应的任务输出组成的图像对，我们首先利用预先训练好的图像量化器将它们量化为离散的令牌。在每个图像的标记表示的前面插入一个特殊的标记“[BOI]”。然后，我们按照优先级的顺序连接每一对的视觉标记。这种结构在两个上下文对之间创建了一个内聚关系，将它们都构建为可视化标记组件。

视觉语言任务。对于视觉语言任务，这里我们以密集字幕任务为例。这些提示很清晰，与自然语言处理（NLP）任务非常相似。与现有的方法[1]类似，多个字幕样本可以被视为交错的图像和文本数据。对于每一幅图像，我们以与纯视觉任务相同的方式来量化它们，并使用特殊的“[BOI]”标签。对于文本部分，我们描述了带有相应的实例类别和边界框（bbox）的区域标题，如“类别： <c>.”。Bboxes: [x_1, y_1, x_2, y_2].标题：<text>。”而 $P = \{x_i, y_i\}_{i=1}^{N}$ 表示定位对象的点。<text>表示标题令牌的占位符。我们还在每个标题的开头添加了一个特殊的标签“[BOT]”。在通过查找词汇表进行标记化之后，我们使用类似的连接策略来获得上下文内的标记表示。

在上下文标记的每个片段结束时，我们加入了一个“[EOC]”标签来表示上下文样本的完成。

3.2. 统一的多模态表示法

基于第 3.1 节中讨论的多模态上下文内提示的基础，如何促进模型以统一的方式理解多模态输入是一个具有挑战性的问题。回顾之前的视觉语言模型[1,43]，我们决定使用离散标记方法作为各种输入和模型嵌入空间之间的桥梁。在本节中，我们将通过统一基于模态特定量化的表示来演示使用多模态上下文输入的通用训练配方的准备工作。

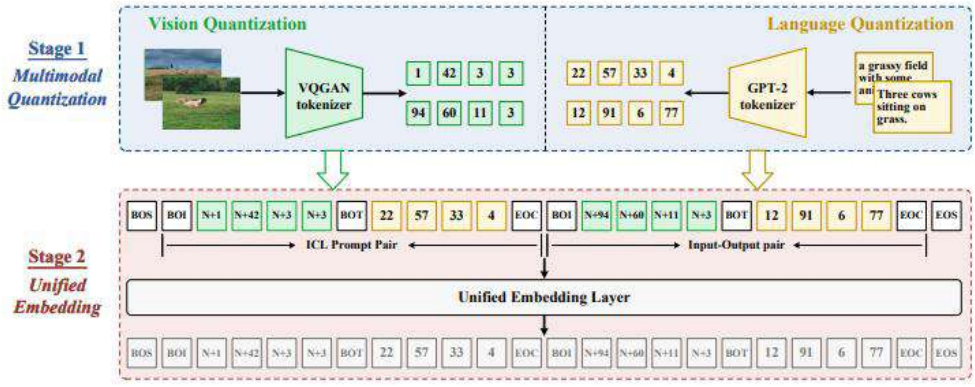

图 2.我们具有两个阶段的统一多模态表示管道的概述。在多模态量化阶段，视觉和语言输入通过模态专门的标记器被编码成离散的标记：具体来说，VQGAN 的标记器和文本的 GPT-2 的标记器。然后，在统一嵌入阶段，将多模态离散标记格式化为带有特殊标记的交错序列。然后，一个统一的嵌入层将序列投影为一般的表示形式。

多模态量化阶段。我们利用现有的众所周知的模态特定的量化器将多模态数据编码为离散标记。如图 2 所示，对于图像数据，我们采用了 VQ GAN [13]中使用的矢量量化器。给定一个图像 $x_{img} \in R^{H \times W \times 3}$，量化步骤通过搜索学习的离散码本 $Z=\{Z_k\}^K_{k=1} \subset R^{n_z}$ 中最近的嵌入，其中 nz 为码本大小，可以表示为：

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

其中 $z_{q\text{-}i}$ 是 x_{img} 的量化编码，E 表示卷积编码器。我们将视觉标记添加到文本词汇表中。

对于文本部分，使用了 GPT-2 [34]中的子字字节对编码（BPE）标记发生器。在编码信息的上下文中，BPE 标记化器通过查找词汇表将 x_{text} 量化为标记 $z_{q\text{-}t}$。我们将类别标签 c 视为自然语言格式，用两个特殊的标签<c_st>和<c_ed>表示这部分的开始和结束。与[45]中提出的类标记相比，语言中的类别标签提供了泛化到看不见的类的潜力。对于 bbox 信息，我们在[7]中采用了类似的方法。根据图像的大小将坐标 P 归一化后，我们将其映射到预定义的标记{<bin_0>，・・・，<bin_1000>}。额外的开始和结束标签<b_st>，<b_ed>被放置在 bbox 的两端。因此，我们可以用比数值表示更少的标记来控制坐标的精度。

统一嵌入阶段。在将每个模态数据量化为离散标记后，我们采取嵌入步骤。在这里，我们平等地对待两种模式下的数据，因为所有的标记都将通过一个线性层映射到一个统一的表示嵌入空间中。然后，所有上下文内的标记嵌入将按顺序连接为“（$z^1_{q\text{-}i}$，$z^2_{q\text{-}t}$）、・・・，（$z^{k+1}_{q\text{-}i}$，$z^{k+1}_{q\text{-}t}$）”，并输入到模型中。该设计为多模态知识转移提供了通用性和可扩展性。因此，该模型可以处理交错的图像和文本输入，如火烈鸟[1]。

3.3. 模型架构和培训目标

在统一了各种模态数据之后，我们现在将讨论如何在一个一般的框架中执行上下文内学习。我们使用 GPT-2 风格的解码器变压器架构构建我们的模型，稀疏 MoEs 用于多模态上下文学习。如图 3 所示，整个框架非常简单和直接。对于交错的输入表示，我们利用下一个标记预测来建模上下文信息。模型的预测对数将经过一个采样过程，将它们转换回标记，随后由每个模态的各自标记器解码。因此，该模型可以实现多模态输入提示和预测，而不是由于预先训练的主干而被限制在特定的输出领域。

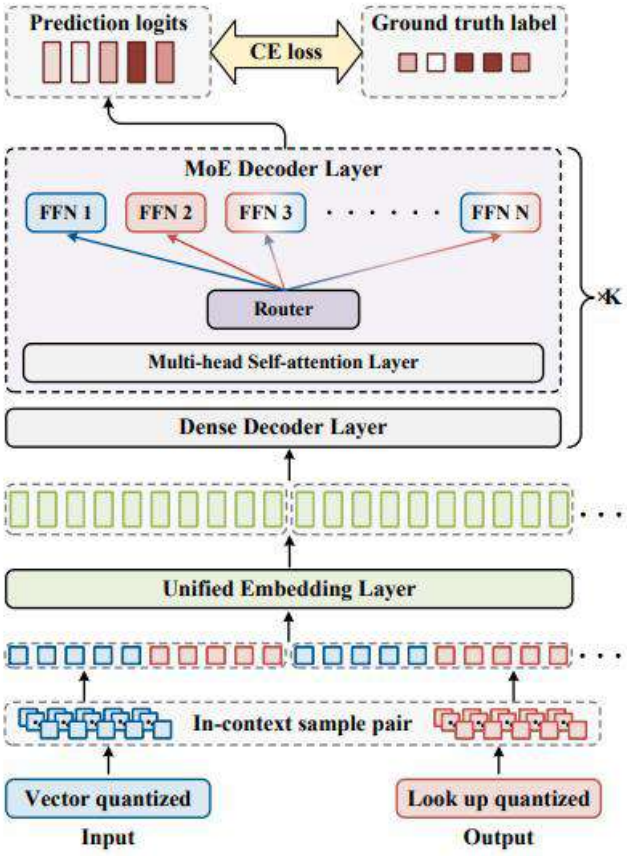

图 3.我们的管道的概述。在这里，我们以 CA-ICL 字幕任务为例。多个上下文样本和输入对首先使用特定模式的标记器进行标记，然后投影到统一的嵌入表示中。经过交错连接后，标记被输入到模型中进行生成建模

属性路由 MoE。具有共享参数的不同任务可能会相互冲突，如前面的工作[14,58]中所述。为了减轻任务干扰问题，我们使用了 MoE 层，它允许不同的模式和任务使用单独的参数。具体来说，我们将每个 MoE 解码器层中的 FFN 块替换为稀疏的 MoE 层，该层是[35]中引入的 N 个专家。在单感知-moe 之后，我们调整了上下文令牌的属性路由策略，并实现了 top-k 门控来决定每个令牌 $x \in R^D$ 嵌入的门控决策。因此，门控的计算公式为： $G(x) = top_k(softmax(W_g(x)))$，其中 W_g 是路由器的可学习权重，$top_k(\cdot)$表示选择最大 k 值的选择器。门控后，稀疏 MoE 层的输出是被激活专家计算的加权组合： $x_{out} = \sum_{i=1}^{N} G(x)_i \cdot FFN_i(x)$。

损失函数。与之前的视觉多面手[4,47,48]使用掩蔽图像建模作为学习目标不同，我们对火烈鸟[1]等交叉的上下文表示执行生成建模，通过利用下一个标记预测，受益于自然上下文理解。

交叉熵损失用于每个上下文对的输出标记和输入对，这限制了模型预测 P_{pred} 和地面真值标记 P_{gt} 之间的相似性，表示为：

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

我们还利用 GShard [23]中引入的辅助损耗来优化 MoEs 的门控网络，整个损耗函数可以表示为：

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

其中， λ 为辅助损失的重量。

4.实验

4.1 数据集和基准测试

先前在视觉语境学习方面的工作主要是将 NLP 的概念整合到传统的视觉任务中。正如 MAE-VQGAN [4]、Painter [47]和 SegGPT [48]所述，每个任务都涉及创建一个网格结构的图像。然而，这些方法忽略了特定任务的理解，将所有任务合并为一个单一的提示。因此，我们提出了一种重新定义语义线索的传统视觉任务方法，强调视觉语言理解任务，如语义分割和图像字幕，分别命名为类感知（CA-ICL）分割和字幕。

CA-ICL Segmentaion.如图 4 所示，对于分割特定类的实例，每个上下文中的样本都只提供了所需的类分割掩码。我们使用整个 MS-COCO 数据集进行数据处理，其中包含 80 个对象类。对于每个类别，都构建了一个用于上下文采样的掩码池。最后，我们收集了大约 35 万 k 类面具用于训练，15k 类面具进行验证。**评价度量**：我们采用传统的语义分割度量对 Union（MIoU）进行评价。假设输出是一个二进制掩码，我们也提出了平均绝对误差（MAE）分数。

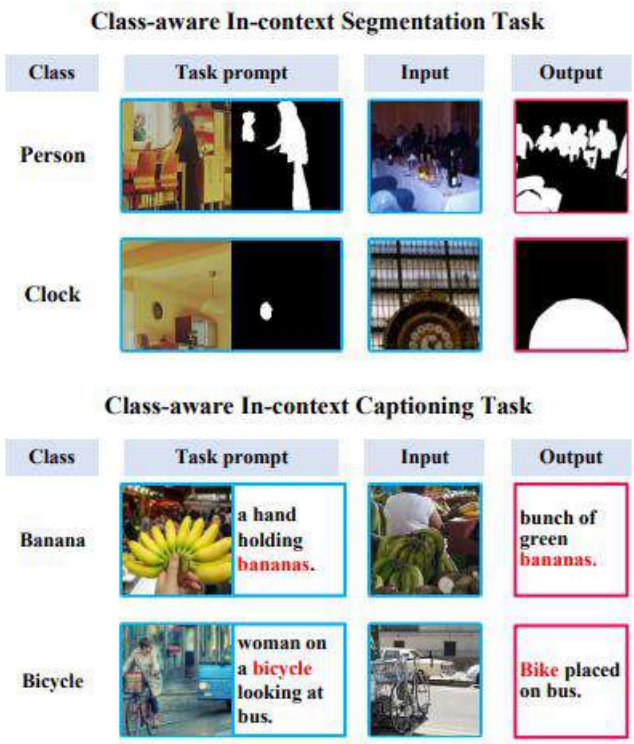

图 4.类感知的上下文理解任务定义。为了便于演示，这里只使用了一个上下文内的示例。左边的蓝框□显示模型的输入，而右边的红框□显示相应的输出。（在没有额外澄清的情况下，随后的符号传达了相同的含义。）

CA-ICL Captioning.对于 CA-ICL 字幕，我们还将类信息作为上下文线索，每个上下文样本都包含所需类别的标题。在这里，我们使用视觉基因组数据集，其中每个图像都有多个注释，包括针对图像的每个区域的对象标签和标题注释。我们有选择地使用与 MS-COCO 数据集中的类别相对应的类别，以确保每个类有超过 100 个描述。最后，我们收集了大约 460k 的区域描述的训练和 2k 的区域描述的测试集。**评估指标**：字幕性能使用 BLEU4、流星和 CIDEr

指标进行评估，这是图像字幕任务的标准。当在提示中合并 bbox 信息时，我们还在[19]之后显示平均平均精度（mAP）度量。通过对离子单元和流星的预定义阈值过滤预测，得到两个阈值的所有成对组合的 APs 的平均值，以评估定位和描述精度。

4.2.实施细节

对于图像标记器，我们采用 VQ-GAN 标记器[13]，词汇量大小为 1024 和 16 倍，这是在图像集数据集上进行预训练的。输入图像分辨率设置为 256×256，量化后得到 256 个令牌。对于文本标记器，我们使用 GPT-2 BPE 标记器[34]，词汇表大小为 50257。我们用 gpt-小模型架构实现我们的模型，同时用[58]中引入的属性路由 MoEs 替换部分解码器层中的 FFN。有关详细的体系结构超参数，请参阅补充部分。

在每次训练迭代中，上下文内样本的数量默认设置为 3。所有的参数都是从头开始训练的。权重 λ 设置为 0.02。为了进行优化，我们采用了基本学习率为 1e-4 的 AdamW 算法，辅以权重衰减为 0.05。我们利用 0.5 的梯度剪切来稳定训练过程，确保整个一致的性能。除非另有说明，训练在 8 个 NVIDIA A6000gpu 上运行 40 个周期，批处理大小为 512。

4.3.消融研究

在本节中，我们从任务定义、模型定义和多任务共训练策略三个角度对我们的方法进行了消融研究。在没有额外陈述的情况下，实验使用 128 分辨率的图像和 20 个时期进行。

具有类感知的上下文内的任务定义.在我们探索两个提出的上下文学习任务时，我们严格地检查了任务的定义。如表 1 所示，我们研究了 CA-ICL 分割任务中每个上下文样本中的对象大小和规模。我们的研究结果表明，包括一个小的物体和一个大的物体规模会产生最佳的结果。我们推测，跨越多个尺度的对象提供了更详细的见解和显著的上下文样本，导致了更丰富的信息多样性，这有利于分割。

在我们对 CA-ICL 字幕的研究中，我们探讨了上下文输入图像与其相应描述之间的相关性。我们从密集的字幕和视觉基础中获得灵感，研究合并对象位置信息是否有利于模型捕获上下文样本所传递的语义线索。

diverse sizes	large scale	MIoU ↑	MAE ↓
✗	✗	31.82	0.176
✓	✗	33.54	0.172
✗	✓	42.87	0.133
✓	✓	**45.04**	**0.128**

表 1.类感知上下文分割任务中对象大小和尺度的消融情况。对于对象的大小，我们采用 MS-COCO 的定义，即是否包含对象面积小于 322 平方单位的小实例。为了考虑对象规模，我们考虑了作物区域。高亮显示的行表示最佳选择。（在没有额外澄清的情况下，随后的符号传达了相同的含义。）

如表 2 所示，与基线相比，引入一个图像类型的输出会导致性能的显著下降。为了解决这个问题，我们探索了以文本格式编码 bbox 信息的方法，如第 3.1 节所述。虽然结果比“bbox 图像”好得多的方法，甚至优于 CIDEr 度量的基线。图 5 演示了使用“bbox 文本”类型的提示可以获得与预期区域对应的更精确的预测标题。这种对齐大大有助于在测试阶段准确和方便地验证模型的性能。在提供适当的示例的情况下，这个证据支持了模型有效地生成类感知标题的能力。

bbox_image	bbox_text	B@4 ↑	CIDEr ↑
✗	✗	**7.9**	104.4
✓	✗	0.0	2.7
✗	✓	7.8	**112.0**

表 2。bbox 信息在类感知上下文标题任务中的影响研究。“bbox 图像”和“bbox 文本”表示边界框为图像类型或文本格式。

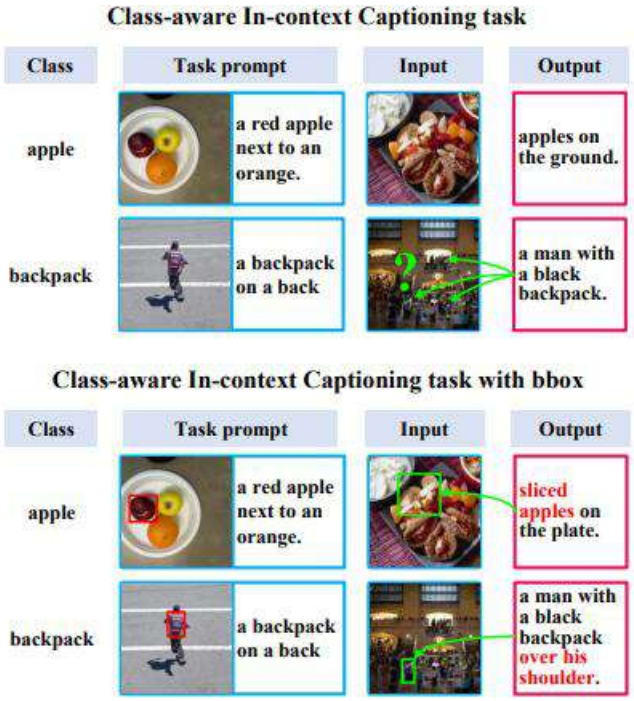

图 5.包含 bbox 信息的影响分析。为了更好地可视化，地面真实的 bbox 用玫瑰框□表示，而预测的 bbox 用绿色框□突出显示。使用提示中的 bbox 信息，模型会生成与指定区域位置对齐的更精确的描述。

模型变量定义。我们在更高分辨率 256 的各种模型配置进行了实验，以确定最佳选择。这些实验的参考是单任务表现，使用标准的 GPT-2 小架构建立任务共同训练，称为“所有任务”。我们用[23]中提出的 MoE 层和[58]中提出的 AG MoE 层代替了部分变压器块中的 FFN。表 3 中给出的结果显示，基线设置导致显著不平衡性能急剧分割性能下降，而模型与教育部配置超过基线分割性能 18.74 分，但仍然有一个显著的不足 10.8 分数的字幕性能。AG_MoE 结构的采用进一步缩小了这种性能差距。考虑到图像标记比文本标记占主导地位，以及不同任务复杂性的不同梯度方向（如图 6 所示），标题性能下降。具有共享参数的模型可能难以有效地管理两个任务之间的标记表示的显著差异，这突出了 moe 的优势。在下一节中，我们将解决与多任务协同培训相关的挑战。

Model	CA-ICL segmentation	CA-ICL captioning
	MIoU ↑	CIDEr ↑
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)

表 3.模型变量的消融和多任务学习策略。我们分别提出了 CA-ICL 分割和字幕任务的 MIoU 和 CIDEr 指标。在括号中，我们分析了与“所有任务”设置相比的差距。我们使用绿色和红色来表示性能的下降和提高。

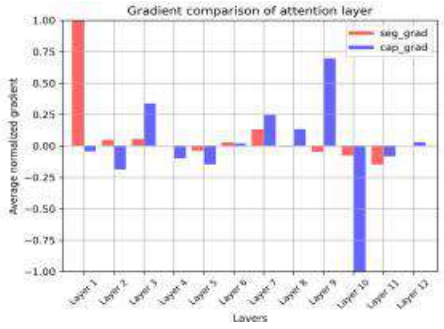

图 6.CA-ICL 任务的梯度比较。我们利用每个注意层的归一化平均梯度进行比较，而符号和值分别表示梯度的方向和大小。

多任务协同训练策略。在本节中，我们将探讨多任务联合训练的影响。如表 3 所示，使用标准的 GPT-2 小架构进行协同训练会导致显著的性能下降，这表明在处理涉及不同数据模式的任务方面存在相当大的差异。AG_MoE 体系结构的实现导致了跨任务的更平衡的性能，但与单任务场景相比，仍然存在显著的性能差距。为了进一步提高 AG_MoE 模型的性能，我们采用了多任务学习范式来缓解任务干扰问题，同时稳定了 MoE 模型的训练。从单感知器 v2 [25]中获得灵感，我们利用它们的非混合批量采样策略和相关优化器。在这里，每个数据集的采样权值 sk 被配置为与数据集大小的平方根成正比。对于缩放因子 wk，我们统一地为所有任务分配一个值为 1。如表 3 所示，将 AG MoE 架构与我们的多任务学习策略集成后，这两个任务的性能都超过了基线。这在 CA-ICL 分割任务中尤其值得注意，在该任务中，我们显著获得了 28.17 分。这表明，多任务策略有效地防止了批处理内潜在的任务冲突。

4.4.与最先进的方法进行比较

我们尝试了具有类感知能力的上下文任务，以与现有的最先进的专家模型和多面手进行比较。对于任务定义，我们采用了在烧蚀”（第 4.3 节）中讨论的最佳设置。对于模型和训练策略，我们利用 AG MoE 体系结构和多任务学习策略。

对于 CA-ICL 分割，我们与多面手分割模型 SegGPT [48]和专家少数分割模型如 FRTrans [55]、VAT [17]和 DCAMA [39]进行了比较。如表 4 所示，我们的模型在 256 的分辨率下训练，超过了在相同分辨率下评估的 SegGPT——MIoU 提高了 6.92，MAE 评分提高了 0.006。然而，在具有更多训练数据和更高分辨率输入的 SegGPT 的 448 版本之间仍然存在差距。其性能也可以与最先进的专家 DCAMA 相媲美，它的更高的分辨率是 384。

在 CA-ICL 字幕领域，评估的通才基线是开放火烈鸟[2]，这是一个大型的视觉语言模型，擅长于展示强大的上下文字幕能力。CA-ICL 字幕任务最类似的是密集字幕，因为这两个任务不仅需要预测标题，还需要预测相应的 bbox。因此，我们与 sota 密集字幕模型 GRiT [50]进行了比较。我们利用测试集中的图像来评估 GRiT。然后将生成的预测分配给我们的地面-真

实区域注释，利用它们各自的 bboxs 的 IoU 度量作为分配的基础。如表 4 所示，我们的方法在传统的图像字幕指标中达到了最先进的性能。与参数增加了 10 倍的开放火烈鸟相比，我们的 BLEU4 也增加了 0.7 点，CIDEr 的 25.6-point 也显著改善。然而，与 GRiT 相比，该结果在 mAP 评分上仍然存在差距。我们认为这是因为它们利用了一个前景对象提取器。

Models	Resolution	#Trainable Params	CA-ICL Segmentation		CA-ICL Captioning			
			MIoU ↑	MAE ↓	B@4 ↑	METEOR ↑	CIDEr ↑	mAP ↑
specialist model								
FPTrans [55]	480	139M	43.30	0.202	-	-	-	-
VAT [17]	417	27M	46.07	0.087	-	-	-	-
DCAMA [39]	384	89M	53.06	0.059	-	-	-	-
GRiT [50]	1024	197M	-	-	5.2	9.0	58.6	15.9
generalist model								
SegGPT [48]	448	307M	62.83	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

表 4.与最先进的专家和通才模型对课堂感知情境任务的比较。我们报告了 MIoU 和 MAE 分数以进行比较。*表示，我们用 256 个分辨率的图像来测试 SegGPT。我们强调了之前最先进的研究结果。

5.结论

在这项工作中，我们提出了一个统一的框架的上下文视觉理解。通过利用多模态量化和统一嵌入，我们的模型能够在一般的标记嵌入空间中联合学习多模态数据。通过将自回归变压器与 MoEs 框架协同集成，我们实现了稳定的多任务协同训练，同时受益于每个任务的平衡贡献。总的来说，我们的研究展示了跨各种模式和任务的情境学习的潜力。