SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

October 2025, Vol. 68, Iss. 10, 200106:1–200106:15
https://doi.org/10.1007/s11432-024-4603-x

Special Topic: Large Multimodal Models

Consistent multimodal pre-training for visual tokenization

Ting PAN[1,2,3], Lulu TANG[3], Xinlong WANG[3], Xin LIU[4*] & Shiguang SHAN[1,2*]

[1]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100086, China*
[2]*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*Beijing Academy of Artificial Intelligence, Beijing 100083, China*
[4]*SeetaCloud, Nanjing 210000, China*

Received 15 May 2024/Revised 30 April 2025/Accepted 23 July 2025/Published online 28 September 2025

Abstract Multimodal large language models (MLLMs) have recently demonstrated notable progress in understanding diverse visual context. Nevertheless, the overall performance of these large vision-language connecting models is highly related to a smaller vision-language pre-trained (CLIP) model at low resolution. Currently, this nesting vision-language alignment paradigm has hindered the development of a distinct vision foundation model for domain-specific multimodal tasks (e.g., OCR and document perception). In this paper, we explore a native high-resolution vision foundation model that is specifically designed for both image-level and region-level multimodal language tasks, clearly substituting the low-resolution CLIP models. Specifically, we introduce TAP-v2, a novel visual tokenizer that encodes general-purpose contextual information to enable comprehensive perception across diverse visual content.

Keywords foundation model, multimodal, representation learning, visual tokenization

Citation Pan T, Tang L L, Wang X L, et al. Consistent multimodal pre-training for visual tokenization. Sci China Inf Sci, 2025, 68(10): 200106, https://doi.org/10.1007/s11432-024-4603-x

1 Introduction

A key design of multimodal understanding is to efficiently and effectively abstract the visual context for language models. It demands a single vision encoder that is capable of tokenizing excessive regions for any subsequent visual or language prompts. However, existing methods often focus on either transferring performant CLIP [1] variants (e.g., SigLIP [2], EVACLIP [3], and InternViT [4]) at low resolution [5–7], or additionally combing features from the vision-centric encoders (e.g., DINOv2 [8] and SAM [9]) at high resolution [10–12]. But nevertheless, the CLIP-style pre-trained vision encoders have been examined to be indispensable [13, 14] for visually-conditioned language models (VLMs); even the corresponding visual representation could hardly capture the small or crowded objects for language understanding.

To enable the fine-grained CLIP visual features at low resolution, recent studies [15–17] coincidentally focused on `split-encode-merge` operation to gather the features of cropped sub-images. However, this revision is cumbersome and poses a new challenge for large language models (LLMs). Specifically, Refs. [15, 17] concatenated features along the sequence dimension, resulting 2k–8k visual tokens to heavily occupy the multimodal context sequence. On the other hand, Ref. [16] concatenated features along the embedding dimension, demanding a larger number of attention heads to integrate the incomplete visual information.

In this work, we explore a native high-resolution and general-purpose visual tokenizer for both image-level and region-level multimodal language tasks, clearly substituting the coarse-grained CLIP models. This is achieved by simultaneously enhancing a pre-trained promptable model with curated region and image captions for language-guided visual tokenization. By leveraging the accurate visual prompting and versatile language prompting, the massive tokens ($\sim$4k) output from the vision encoder can naturally gather the high-density and comprehensive visual information for flexible visual and language prompts.

We start by introducing a novel promptable tokenization framework for segmentation, recognition, and captioning. This requires a unified model capable of abstracting two general-purpose representations, i.e.,

[*] Corresponding author (email: xin.liu@seetacloud.com, sgshan@ict.ac.cn)

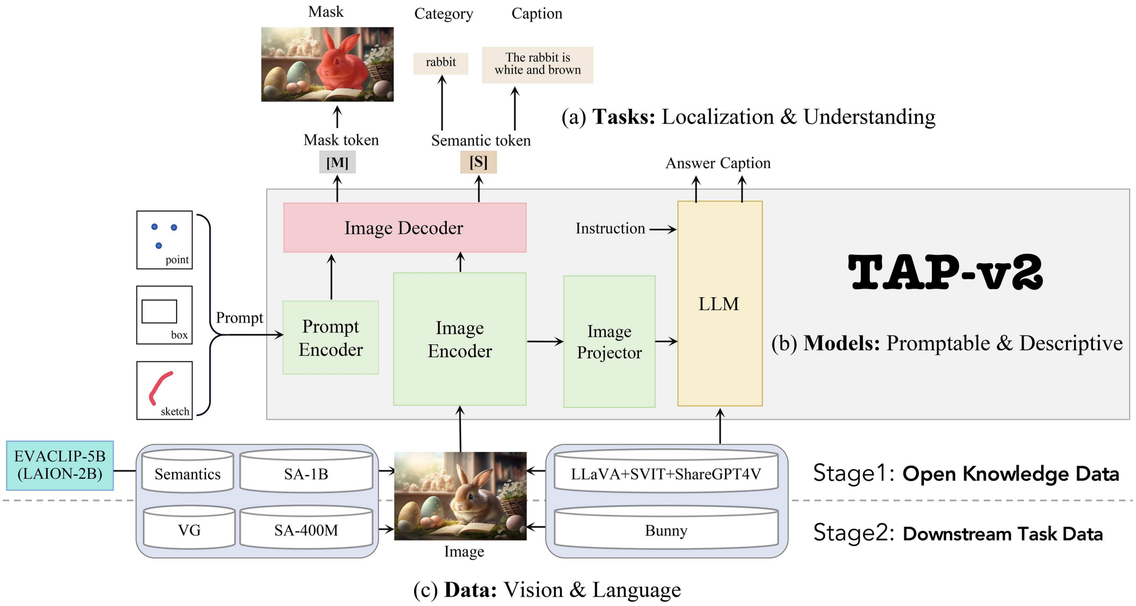


Figure 1 (Color online) TAPv2 is a unified foundation model capable of understanding any regions or images through flexible prompts, including points, bounding boxes, sketches, and texts. To achieve region-level understanding, we employ a versatile image decoder by integrating a single mask token and a corresponding semantic token per predicted mask. For image-level understanding, we utilize an image projector and an LLM to transform the model into a vision assistant. The model is pre-trained initially on SA-1B segmentation masks combined with semantic priors derived from an EVACLIP-5B, followed by fine-tuning on a combination of pre-trained and Visual Genome (VG) data. It is subsequently pre-trained using an LLM on synthetic image captions collected from LLaVA, SVIT, and ShareGPT4V, and finally fine-tuned on instruction-following data collected by Bunny.

the mask token and semantic token, given flexible prompting that cues any region of interest. We follow SAM's architecture, but upgrade the mask decoder into a generic decoder, where an additional semantic token is produced for each predicted mask. The mask token contributes to pixel-wise segmentation, similar to SAM, while the semantic token is responsible for region-level semantic prediction. By leveraging a semantic token, the model can concurrently address an open-vocabulary classification task through an MLP head, and a promptable captioning task with an auto-regressive text decoder. We refer to this model as TAP-v2-Base, an acronym for tokenize anything via prompting, as illustrated in Figure 1.

Training such a highly performant and generalizable model necessitates a diverse, large-scale dataset. Currently, there is no web-scale data source available for simultaneous segmentation and recognition. SA-1B [9] constructs 1.1 B high-quality mask annotations on 11M images for training a segmentation foundation model, e.g., SAM. On the other hand, LAION-2B [18] collects 2 B image-text pairs from the web, facilitating the training of generalizable recognition models, e.g., CLIP. To address the challenge posed by the lack of aligned data, we introduce the SemanticSA-1B dataset (see Figure 1(c)). This dataset implicitly integrates web-scale semantics from LAION-2B into SA-1B. Specifically, for each segmentation mask within SA-1B, we extract its concept distribution over a concept vocabulary as its semantic prior, which is predicted by a powerful CLIP model trained on massive LAION image-text pairs. As a result, the SA-1B data, along with LAION-2B priors, contribute to our initial pre-training dataset.

Using the SemanticSA-1B dataset, we firstly pre-train our model with the ground-truth masks and associated semantics from scratch, effectively integrating CLIP's capabilities within the TAPv2 architecture. This is achieved by pre-training a promptable tokenizer simultaneously for segmentation and concept prediction. To predict the semantic concept for each masked image, we further propose minimizing the KL divergence loss between the predicted concept distribution and the target distribution, aiming to maximize the transfer of CLIP knowledge. This joint training loss enables powerful generalization in both localization and recognition, thereby facilitating general-purpose vision tasks.

We then introduce a region-level multimodal pre-training stage, that enables end-to-end vision-language binding across supervised region descriptions (e.g., Visual Genome [19] dataset), weakly supervised CLIP semantics (e.g., EVACLIP-5B [3]), and semi-supervised segmentation masks (e.g., SA-1B [9] dataset). The mixed supervision can be evenly employed within TAPv2 architecture, performing the unified optimization through versatile and promptable tokenization experts. Consequently, we create the CaptionSA-400M dataset, a mixture of 40% SemanticSA-1B data and 400% Visual Genome data for pre-training.

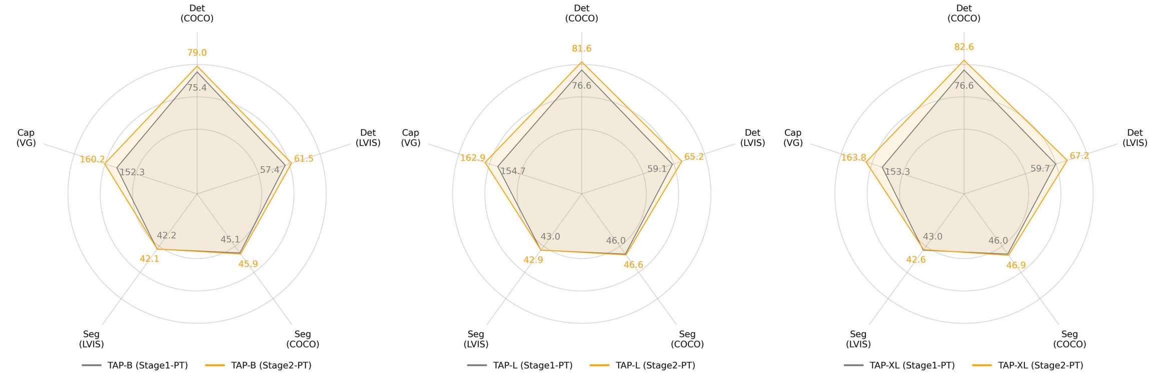


Figure 2 (Color online) The two-stage multimodal pre-training consistently improves TAPv2's semantic representation at scale.

With the help of CaptionSA-400M dataset, we significantly improve the open-vocabulary classification and region-level caption performance (Figure 2). This unlocks a promising pathway towards powerful zero-shot region understanding without unbalanced and handcrafted semantic tags [20, 21].

We introduce an image-level multimodal pre-training stage that enables visual reasoning capabilities for subsequent visual instruction-tuning (e.g., LLaVA [22]) stage. We refer to this model as `TAP-v2-Chat`. Building on recent practices in constructing high-quality image caption datasets [23, 24], we opt for pre-training TAPv2 on the detailed image caption data, but merely use the COCO [25], Visual Genome [19] and SA-1B [9] images. To be specific, we collect 232K image captions synthesized by GPT4 [26] or generated by GPT4-V [27], sourcing from [22, 23, 28]. We train an image captioning model by jointly optimizing the vision components and language model. This optimal captioning task effectively translates the region-level visual tokens for text tokens, demonstrating the redundancy of the pre-aligned CLIP model.

We extensively evaluated our TAPv2 model and its components. TAPv2 demonstrates the strong zero-shot performance in instance classification, e.g., 67.2 AP on the challenging LVIS [29] benchmark, while maintaining competitive zero-shot segmentation performance, e.g., 42.6 vs. 43.1 AP for TAPv2 and SAM. Notably, we set a new record with a CIDEr score of 163.8 in the region caption task on Visual Genome [19], using significantly fewer parameters compared to the prior studies. When connected with large language models, TAP shows comparable or even better performance as CLIP, e.g., 62.8 score on the GQA [30] benchmark and 63.7 score on the SEED-Bench [31]. Our findings demonstrate that tokenized region-level features are generalizable across both localization and semantic understanding, and can be directly transferred for image-level comprehension. Above all, we believe the TAPv2 model can be a versatile image tokenizer, capable of encoding context for a broad range of visual perception tasks (see Figure 3).

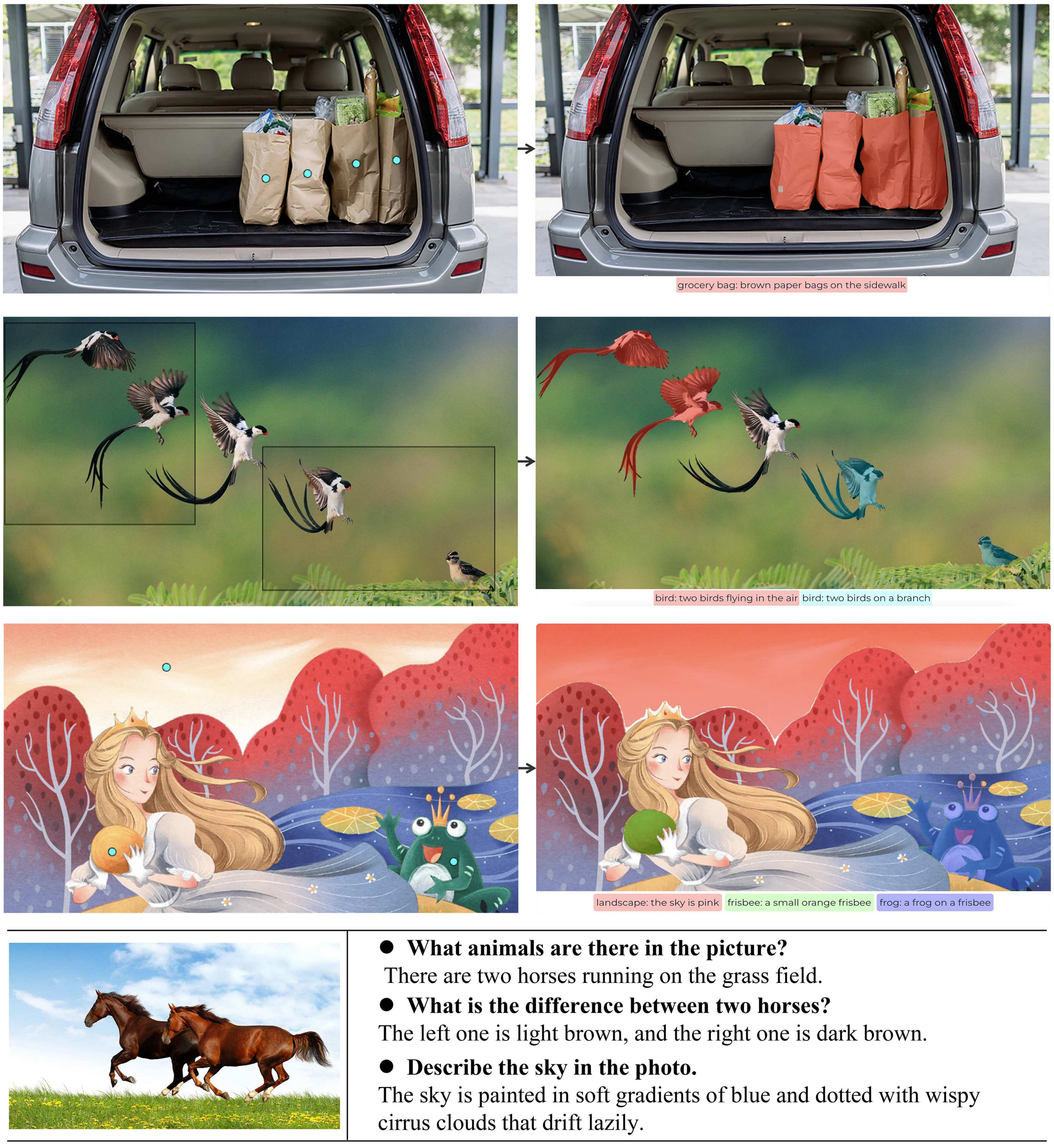
2 Related work

2.1 Vision foundation models

Vision foundation models aim to achieve strong zero and few-shot generalization capabilities across a broad range of vision tasks. Starting with CLIP [1], which simultaneously trains image and text encoders with massive image-text pairs to align two modalities, numerous efforts have emerged to train a general-purpose vision-language representation at scale [3, 32, 33]. In addition, some studies aim to build vision generalist models [9, 21, 34–36]. For example, SAM [9] introduces a large-scale dataset and trains a model for promptable segmentation. Concurrent to SAM, SegGPT [35] unifies a variety of segmentation tasks into one in-context segmentation problem. Some other studies seek to build a generalist model by leveraging multi-modality datasets [5, 37, 38]. In this work, we aim to build a vision foundation model that serves as a tokenizer to encode general-purpose visual context for a broad range of multimodal tasks.

2.2 Zero-shot region understanding

Recent studies [39, 40] aim to merge CLIP's proficiency in open-vocabulary classification with SAM's capability in segmentation. For instance, RegionSpot [40] unifies prompting by adding an adapter trained


grocery bag: brown paper bags on the sidewalk
bird: two birds flying in the air
bird: two birds on a branch
landscape: the sky is pink
frisbee: a small orange frisbee
frog: a frog on a frisbee

- **What animals are there in the picture?**
There are two horses running on the grass field.
- **What is the difference between two horses?**
The left one is light brown, and the right one is dark brown.
- **Describe the sky in the photo.**
The sky is painted in soft gradients of blue and dotted with wispy cirrus clouds that drift lazily.

Figure 3 (Color online) TAP-v2 accepts flexible prompts and outputs mask, category, caption, and chat response at once.

on detection datasets, enabling SAM's mask tokens to interact with CLIP's features derived from masked image segments. SAM-CLIP [39] distills knowledge from both SAM and CLIP by retraining the visual encoder with a portion of data sampled for two teachers, retaining the original strengths of both CLIP and SAM. Some other studies [21, 36, 41] attempted to construct unified models capable of recognizing objects in arbitrary regions. For example, SEEM [36] is built upon X-Decoder [42], excelling in handling various types of prompts, including clicks, bounding boxes, scribbles, text, and referring image segments. Following SAM [9], ASM [21] created a new dataset (AS-1B) for SA-1B [9], constructing rich annotations of semantic tags, question-answering pairs, and detailed captions. Leveraging this dataset, they develop a new model, ASM, for panoptic visual recognition. Unlike these models relying on multi-modal datasets, we leverage segmentation masks from SA-1B and semantic priors from a high-performing CLIP model, aiming to develop a promptable tokenizer that can understand semantic context for any given region.

2.3 Multimodal visual understanding

With the success of LLMs, many recent studies [12, 22–24, 37, 43, 44] shifted toward the multimodal large language models (MLLMs) for unified visual understanding and reasoning. Flamingo [37], BLIP-2 [43], and MiniGPT-4 [44] propose to fuse visual tokens to frozen large language models through gated attention or query transformers. Building upon the pre-trained LLMs with the affordable fine-tuning cost, e.g.,

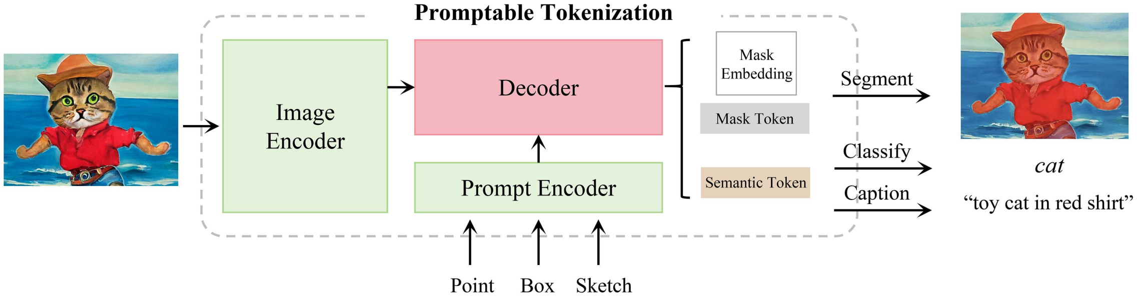


Figure 4 (Color online) Overview of promptable tokenization. Building upon SAM's architecture, we develop a decoder, adding a semantic token to each predicted mask. Our model is pre-trained on SemanticSA-1B, jointly optimized for concept prediction and promptable segmentation. Subsequently, the pre-trained tokenizer (in dotted box) is employed for region captioning.

Phi-2 [45] and Vicuna-7B [46], LLaVA [22] and its data-centric variants [12,23,24] directly concatenate the CLIP visual tokens with language instruction tokens into the LLM, demonstrating remarkable success. In addition, some other studies [13, 14] explored the key design of visual components in MLLMs, mainly including architecture and optimization. The purpose of this work is to reveal that a region-level visual tokenizer could naturally be pre-trained into a strong image-level tokenizer for various multimodal tasks.

3 Approach

We emphatically introduce a promptable framework that efficiently enables the segmentation, recognition, and captioning of arbitrary regions of interest. This is achieved by pre-training a promptable tokenizer that utilizes extensive segmentation masks with CLIP priors, and subsequently expanding the model's capabilities to include generative abilities for promptable captioning (Subsection 3.1). We concurrently introduce a descriptive framework that effectively enables the reasoning of entire regions. This is achieved by pre-training a captioner that describes massive visual elements with detailed captions, and subsequently expanding the captioner to follow the complex visual instructions (Subsection 3.2).

3.1 Promptable tokenization

Our primary objective is to align vision and language within a promptable segmentation model to enhance the model with region-level semantic awareness. To achieve this goal, we introduce our model architecture, pre-training dataset, and learning method involving concept prediction, promptable segmentation, and promptable captioning, as well as the pre-training loss, in this subsection.

Model architecture. The region-level tokenizer model comprises three essential modules: an image encoder, a prompt encoder, and an image decoder (see Figure 4). We maintain SAM's architecture but upgrade its mask decoder to a generic decoder. To more efficiently and effectively achieve our objectives, we make several modifications to SAM's architecture. Specifically, the image encoder adopts a standard vision transformer (ViT) [47], where a 16×16 non-overlapping attention window is employed. To alleviate computational intensity, we substitute the global attention in the image encoder with 3×3 residual convolution blocks [48] and replace the query-based relative position embedding [49] with an index-based version. Regarding the prompt encoder, we do not add the mask prediction from the previous stage to the image embeddings, as it introduces discrepancies between the prior prompts (e.g., sketch points) and advanced prompts (e.g., interactive points). Consequently, all mask embedding layers in the prompt encoder are removed. In the image decoder, we add an additional semantic token to each predicted mask, where the mask token is employed for pixel-wise segmentation, and the semantic token contributes to region-level recognition. Therefore, the image decoder produces a total of 4 masks and 9 tokens: 4 mask tokens, 4 semantic tokens, and an IoU token.

Pre-training dataset. In contrast to prior methods [21, 36, 50] reliant on paired or synthetic region-text data, we align the image segments with language mainly using segmentation data and CLIP priors. As SA-1B is a class-agnostic dataset, we utilize the high-performing CLIP model, EVA-CLIP [3], to compute the concept distribution P_{target} as the semantic prior for each image segment within SA-1B. We first create a label list consisting of 2560 categories collected from various popular image datasets. Then, we employ a simple prompt template: 'a {}' or 'a photo of a {}' to generate the text embeddings T_C

using CLIP. Meanwhile, for each masked image segment from SA-1B, we obtain its visual embeddings V_C generated by CLIP. The concept distribution can be defined as

$$P_{\text{target}} = \text{Softmax}\left(\frac{V_C \cdot T_C/\tau}{\|V_C\| \cdot \|T_C\|}\right), \tag{1}$$

where τ denotes the temperature parameter. Consequently, the segmentation data along with their off-the-shelf CLIP priors $(V_C, T_C, P_{\text{target}})$ are stored, constituting our first pre-training dataset, SemanticSA-1B. To leverage the high-quality but limited region-text data, we further construct another pre-training dataset for a two-stage pre-training. Specifically, we adopt a mixing recipe that merges $\sim$4M regions in the Visual Genome dataset with $\sim$96M regions in the SA-1B dataset, initiating a core subset CaptionSA-100M. We randomly sample the full SA-1B dataset to roll out 4 subsets, finally resulting in CaptionSA-400M for an appropriate pre-training cost (i.e., approximate one epoch SemnaticSA-1B pre-training).

Concept prediction. To enhance our model with semantic awareness, we propose to predict region concepts using the semantic token. Concretely, we employ the semantic token to obtain the predicted visual embedding V_P, which is further projected to the concept distribution P_{pred},

$$P_{\text{pred}} = \text{Softmax}\left(\frac{V_P \cdot T_C/\tau}{\|V_P\| \cdot \|T_C\|}\right). \tag{2}$$

We propose to align the concept distribution between the model's prediction and CLIP's target. The concept alignment loss can be defined as the KL divergence loss between P_{pred} and P_{target}, represented as

$$\mathcal{L}_{\text{concept}} = \mathcal{L}_{\text{KL}}(P_{\text{pred}} \,||\, P_{\text{target}}). \tag{3}$$

Different from feature alignment that typically minimizes the negative cosine similarity between the predicted visual embedding and CLIP visual embedding, formed as $\mathcal{L}_{\text{feat}}{=}-\frac{V_P \cdot V_C}{\|V_P\|_2 \cdot \|V_C\|_2}$, our concept alignment minimizes $\mathcal{L}_{\text{concept}}$ between two distributions. It measures the similarity between V_P and T_C, drawing V_P closer to positive T_C, while pushing it away from negative T_C. This encourages V_P to be orthogonal, maximizing the transfer of CLIP knowledge.

Promptable segmentation. The mask decoder in SAM's architecture responds to input prompts to generate segmentations. We thus consider promptable segmentation as a necessary prelude to unsealing the semantic capabilities. Following SAM, our model defaults to predicting four masks for each prompt; yet a routing strategy selects one to resolve the ambiguity. To improve training efficiency on the large-scale SA-1B dataset, we employ a two-stage sampling strategy with maximal 9 prompt points, as it is performed within 11 interactive stages in the original SAM. In the first stage, we sample a box or point with equal probability from the ground-truth mask. In the second stage, we uniformly sample 1–8 points from the error region between predicted and ground-truth masks. To enable mask as the prompt at the first stage, an aspect unexplored in SAM, we introduce a non-interactive sampling method with a 50% probability in the second stage. This sampling uniformly fetches 1–9 points from the ground-truth mask, providing a wider prompt space. As for mask supervision, a linear combination of focal loss [51] and dice loss [52] is employed at a 20:1 ratio, following SAM [9].

Promptable captioning. In order to employ the promptable semantic tokens to describe any region, we append an additional lightweight text decoder at the top of the model. An overview of our text generation architecture is depicted in Figure 5. We utilize a standard Transformer with an embedding dimension of 512 to yield the brief region descriptions. This lightweight text decoder is sufficient to perform mask-to-text translation if prompted with semantic context. Given semantic tokens generated by the promptable tokenizer (refer to Figure 4), we apply a linear projection to these semantic tokens, aligning their dimensions with text embeddings (see Figure 5). Subsequently, we place the semantic token at the leading position of a sequence, followed by a `[BOS]` token and word tokens. Rotary embedding [53] is utilized to integrate the positional encoding for multi-modal sequences. We adopt byte-pair encoding [54] with a 32k token vocabulary. Eventually, we perform the next token prediction through causal language modeling, employing the cross-entropy loss for the region-level captioning task, denoted as $\mathcal{L}_{\text{cap}}$.

Pre-training loss. Our pre-training loss is the joint loss combining concept prediction, promptable segmentation and captioning: $\mathcal{L} = \mathcal{L}_{\text{concept}} + \alpha\mathcal{L}_{\text{seg}} + \beta\mathcal{L}_{\text{cap}}$. With this loss, we initiate a promptable tokenizer on SemanticSA-1B from scratch at the first pre-training stage by setting β to 0. In the second stage, we experimentally set α, β to $(5, 1)$ on CaptionSA-400M. An overview of our method is illustrated in Figure 4.

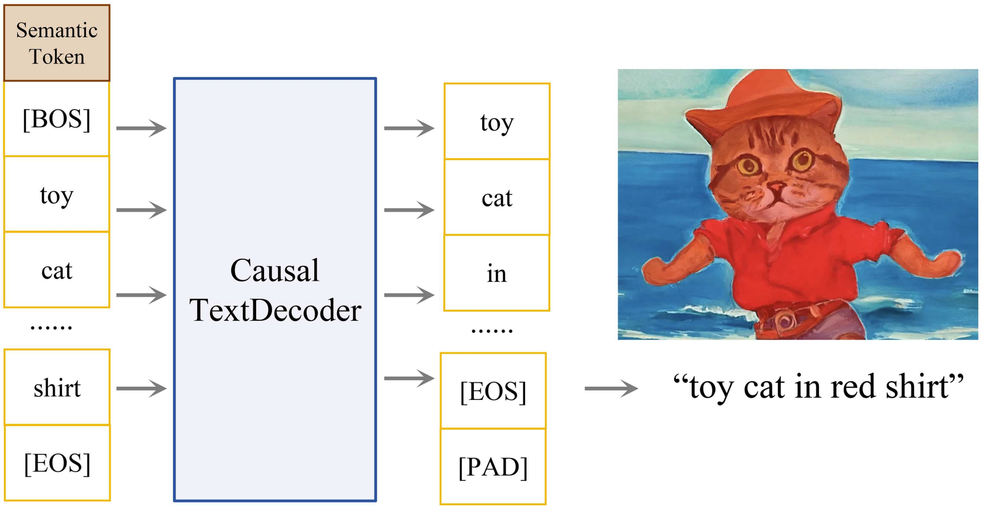
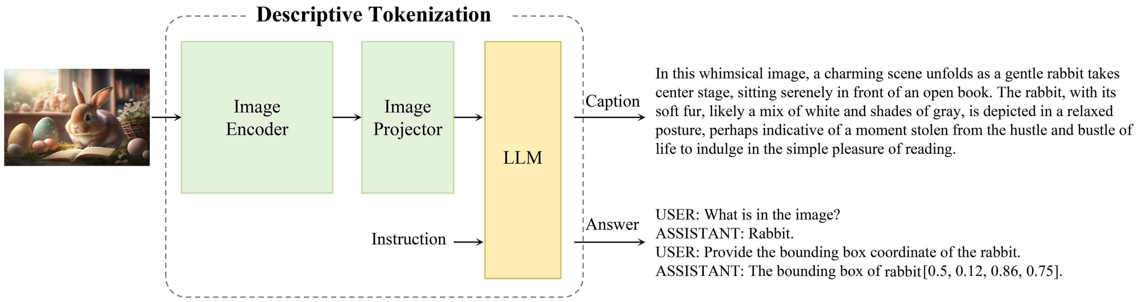


Figure 5 (Color online) Promptable captioning. A semantic token is used to prompt text generation.



In this whimsical image, a charming scene unfolds as a gentle rabbit takes center stage, sitting serenely in front of an open book. The rabbit, with its soft fur, likely a mix of white and shades of gray, is depicted in a relaxed posture, perhaps indicative of a moment stolen from the hustle and bustle of life to indulge in the simple pleasure of reading.

USER: What is in the image?
ASSISTANT: Rabbit.
USER: Provide the bounding box coordinate of the rabbit.
ASSISTANT: The bounding box of rabbit[0.5, 0.12, 0.86, 0.75].

Figure 6 (Color online) Overview of TAPv2-Chat architecture. We append an image projector on top of the image encoder to connect a large language model (LLM). The model is pre-trained on detailed image caption data, conversationally describing each visual element in the image. Subsequently, the pre-trained descriptive tokenizer (in dotted box) is employed for instruction-following.

3.2 Descriptive tokenization

Our secondary objective is to straightly align the pre-trained image encoder with a large language model. In other words, we pursue simultaneous vision and language prompts with two individual decoders, by seamlessly learning an image-level tokenization space from the original region-level tokenization space. In this subsection, we introduce the connecting architecture, training methods, and corresponding datasets.

Connecting architecture. Following [55, 56], we append a lightweight convolutional projector on top of the image encoder for image-level tokenization (Figure 6). This projector employs three 3×3 residual blocks to (i) parameter-efficiently merge the 2×2 neighboring patch tokens and (ii) progressively propagate information across the 4×4 partitioned patch windows. The last layer of each residual block is initialized as zero, such that the initial region-level tokenization is not compromised. The output layer of projector is a two-layer MLP that is introduced to match the embedding dimension of vision and language modalities. **Training datasets.** Since the image encoder is originally pre-trained on high-resolution images for perceiving diverse visual objects, it becomes clear that it should be further trained with similar data. In the pre-training stage, we follow this guideline and collect 232k detailed image captions from three publicly available data sources: LLaVA-Instruct-150K [22], SVIT-core-150K [28], and ShareGPT4V [23]. Our pre-training dataset includes 118k/70k/20k images from COCO [25], Visual Genome [19], and SA-1B [9], omitting the low quality data that are learned in EVACLIP-5B priors (e.g., LAION-2B [18]). Moving to the supervised fine-tuning stage, we utilize Bunny-695K [24] dataset to dive into the GPT-4's open-world knowledge on the abundant objects and regions in the Visual Genome dataset.

Training methods. We employ a two-stage training strategy. Initially, we prompt a large language model with only the visual tokens output from the projector, pre-training a universal captioning model. Subsequently, we concatenate visual tokens and task instruction tokens, tuning the pre-trained captioner into a generalist model to harness the multimodal capabilities across conversational and reasoning tasks. We adopt the cross-entropy loss during both two stages for the next token prediction. Following [6, 23, 56], we optimize the identical components for both the captioning model and generalist model, including (1) the last k layers of image-encoder, (2) the image projector, and (3) the large language model. Specifically, we leverage LoRA [57] to consistently fine-tune the large language model for these longer training schedules.

3.3 Capabilities for inference

After training visual perception on CaptionSA-400M and visual reasoning on Bunny-695K, our model is capable of conducting classification, segmentation, captioning, and instruction-following simultaneously. The following outlines the inference pipeline of all TAP's capabilities.

Mask selection. Given a visual prompt, our image decoder produces 4 masks and 9 tokens. The final mask and the associated semantic token are selected using a heuristic strategy. Specifically, we choose the first mask if prompted with a boundary box, and select the top-ranked remainder if prompted with loose points, akin to a simplified implementation of the mixture-of-experts (MoE) [58].

Concept prediction. The selected semantic token is then utilized for predicting concepts on a dataset-specific concept vocabulary (e.g., COCO and LVIS). Concretely, we employ the semantic token to obtain a 1024-dimensional visual embedding through a 3-layer MLP (256→1024→1024). This visual embedding is further projected to the concept distribution logits (i.e., P_{pred}) for classification.

Region captioning. We generate up to 40 word tokens with the greedy sampling strategy prompted by the selected semantic token. To speed up attention computation, we follow a standard practice for auto-regressive decoding, caching the key and value pairs for the previous generation.

Instruction following. If given the instruction, we prompt a large language model with pre-computed visual tokens and corresponding instruction tokens to generate answer tokens. We attach each pair of `instruction-answer` with the roles `USER` and `ASSISTANT` to organize the multi-turn conversations.

4 Experiments

4.1 Experiments setup

Pre-training. We pre-train the region-level tokenizer on SemanticSA-1B and CaptionSA-400M datasets, which include the SA-1B data along with associated CLIP priors and Visual Genome region captions. The SA-1B data comprise 11M high-resolution images with around 100 regions per image, totaling 1.1 B segmentation masks. To obtain CLIP priors for SA-1B data, inspired by [50, 59, 60], we utilize EVA-CLIP [3] to generate text embeddings on a curated label space, merging from COCO [25], ADE20K [61], LVIS [29], Objects365 [62], Visual Genome [19], and OpenImagesV4 [63]. This results in a concept list spanning 2560 categories. We pre-train the image-level tokenizer with 232k detailed image captions, sourcing the 208k images from [9, 22, 25] that contain common objects in their natural context.

Evaluation. We assess zero-shot instance segmentation performance on COCO and LVIS. For zero-shot instance classification, we prioritize LVIS due to its broader range of 1203 categories compared to COCO, which covers only 80 common categories, diverging from the open-world assumption. In the region-level captioning task, regarding the domain gap between SA-1B and Visual Genome (VG) [19], we evaluate the models that fully fine-tune the text decoder on VG v1.0 train set. We report the following metrics on the VG test set and RefCOCOg [64] validation set: BLEU@4, METEOR, ROUGE, and CIDEr. We evaluate multimodal understanding performance on popular benchmarks: VQA-v2 [65] test-dev split, GQA [30] test-dev-balanced split, TextVQA [66] validation split, MMBench [67] dev split, and SEED-Bench [31].

Implementation details for TAP-v2-Base. We utilize the AdamW [68] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a base learning rate of 1E−3/1E−4 in all stage1/stage2 experiments. A cosine learning rate schedule [69] is implemented. During pre-training on SemanticSA-1B/CaptionSA-400M, scale jitter [70] is applied with a range of [0.5, 2.0] for 180k/45k iterations, using a batch size of 256 across 256 GPUs. We fine-tune VG without augmentation for 60k/36k iterations, with a batch size of 64 across 64 GPUs. The training experiments were conducted on Cambrion MLU370 devices, with stage 1 and stage 2 taking approximately 300/80 A100 days. Additional hyper-parameters include a weight decay of 0.1, a drop path [71] rate of 0.1/0.2/0.2 for ViT-B/ViT-L/ViT-XL, and a dropout [72]

Table 1 Zero-shot instance classification on LVIS. The supervised training results are italic for reference. We evaluate all entries using GT boxes to fairly compare detection-based and detection-free methods. The model suffixes '-B', '-R50x4', '-L', '-BL', '-XL', and '-E' correspond to vision backbone with approximately 0.1, 0.1, 0.3, 0.4, 0.4, and 5 B parameters, respectively. The best results are in bold.

Model	Params (B)	Training data	AP	AP^R	AP^C	AP^F
Supervised detector:						
ViTDet-B	0.1	LVIS	*61.9*	*40.8*	*58.5*	*74.9*
ViTDet-L	0.3	LVIS	*68.8*	*51.5*	*65.6*	*79.9*
Image-level CLIP:						
CLIP-L	0.3	WIT-400M	48.8	52.8	50.0	45.6
EVA-CLIP-E	5	LAION-2B	64.3	72.4	65.3	59.7
Region-level CLIP:						
RegionCLIP-R50x4	0.1	CC-3M	50.7	50.1	50.1	51.7
RegionSpot-BL	0.4	O365, OI, V3D	56.6	50.6	50.2	68.8
Promptable tokenizer with concept prediction:						
TAPv2-B	0.1	SemanticSA-1B	57.4	58.6	56.8	57.5
TAPv2-L	0.3	SemanticSA-1B	59.1	61.7	58.9	58.3
Promptable tokenizer with region description:						
TAPv2-B	0.1	CaptionSA-400M	61.5	60.6	61.2	62.3
TAPv2-L	0.3	CaptionSA-400M	65.2	66.1	65.3	64.7
TAPv2-XL	0.4	CaptionSA-400M	67.2	69.2	**67.4**	66.1

rate of 0.1/0.4 for the image/text decoder. The image encoder is initialized from MAE [73] pre-trained weights, while all other layers are from scratch. For all experiments, we adopt up to 64 prompts per GPU at each sampling stage.

Implementation details for TAP-v2-Chat. We utilize the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$) with a base learning rate of 1E−3/2E−4 in all stage1/stage2 experiments. A cosine learning rate schedule is implemented. We pre-train and fine-tune the multimodal data for 2/1 epochs, with a batch size of 128 across 8 A100 GPUs over the course of 2 days. No weight decay or dropout is applied. The image projector's residual blocks are initialized as identities, and the LLM is initialized with pre-trained weights. We fine-tune the last 4 blocks of the image encoder to reach a satisfactory level of training efficiency.

4.2 Main results

Zero-shot instance classification. We prompt our model with ground-truth (GT) boxes to evaluate the recognition capability on LVIS. With GT boxes as visual prompts, our model substantially surpasses RegionCLIP [60] and RegionSpot [40], which are trained on limited image regions. These promising results suggest that employing concept prediction on exhaustive image regions can effectively empower SAM with semantic awareness. As shown in Table 1, the highly performant EVA-CLIP [3] achieves an impressive AP on rare and common objects. Nonetheless, deploying a standalone CLIP (5 B) model is impractical for real-time vision systems. We demonstrate that the knowledge of large CLIP models can be integrated into a compact tokenizer (0.1 B) with acceptable performance. Furthermore, by simultaneously learning the open-world knowledge from VG captions, TAPv2 achieves significant and consistent improvement on all object categories, showing a light to resolve the long-tailed classification problem elegantly.

Region-level captioning. We assess our model on Visual Genome [19] and RefCOCOg [64]. Initially, we utilize GT boxes to prompt the image decoder, and subsequently, we employ the resulting semantic tokens to prompt the text decoder. The evaluation results are presented in Table 2 [19,64,74]. Surprisingly, our model achieves a CIDEr score of 154.7 on Visual Genome, even with a frozen image encoder-decoder that is pre-trained on SA-1B and has not seen VG images before ('Stage1-PT'). By adopting a two-stage multimodal pre-training strategy ('Stage2-PT'), we achieve a new record with a CIDEr score of 163.9, only using a lightweight text decoder, while previous approaches [75,76] rely on LLMs such as [46,77,78]. It is noteworthy that the concurrent work ASM [21] is trained on a multi-modal dataset, including a vast repository of synthetic region-text pairs. In contrast, the semantic knowledge of our model is learnt from a CLIP model. Another concurrent work, SCA [79], additionally trains a 12-layer image decoder to learn caption tokens for text prompting. These results suggest that our semantic token effectively encodes

Table 2 Region captioning on Visual Genome [19] and RefCOCOg [64]. Following [74], we evaluate the generated captions using GT boxes as the crowded caption regions are out-of-domain for the detection models. The model suffixes '-B', '-L', '-XL', '-H', and '-g' correspond to the vision encoder with approximately 0.1, 0.3, 0.4, 0.6, and 1 B parameters, respectively. The best results are in bold.

Method	VisionEncoder	TextDecoder	TextPrompt	Visual Genome		RefCOCOg	
				METEOR	CIDEr	METEOR	CIDEr
GRiT	ViT-B	Small-43M	BoxFeature	17.1	142.0	15.2	71.6
GPT4ROI	CLIP-H	Vicuna-7B	BoxFeature	17.4	145.2	–	–
ASM	ViT-g	Husky-7B	BoxFeature	18.0	145.1	20.8	103.0
AlphaCLIP	ViT-L	Vicuna-7B	BoxFeature	18.9	160.3	16.7	109.2
SCA	SAM-H	Llama-3B	CaptionToken	17.4	149.8	15.6	74.0
TAPv2 (Stage1-PT)	ViT-B	Small-38M	SemanticToken	17.7	152.3	19.2	93.5
TAPv2 (Stage1-PT)	ViT-L	Small-38M	SemanticToken	17.9	154.7	19.5	95.6
TAPv2 (Stage2-PT)	ViT-B	Small-38M	SemanticToken	18.3	160.2	19.7	97.8
TAPv2 (Stage2-PT)	ViT-L	Small-38M	SemanticToken	18.6	162.9	19.9	99.3
TAPv2 (Stage2-PT)	ViT-XL	Small-38M	SemanticToken	18.6	**163.9**	20.0	100.2

Table 3 (Color online) Zero-shot instance segmentation on COCO and LVIS. The supervised training results are displayed in bronze. We evaluate the segmentation masks using the detection boxes from the ViTDet-H model, as outlined in [9]. The suffixes '-B', '-L', '-XL', and '-H' indicate backbones of approximately 0.1, 0.3, 0.4, and 0.6 B parameters, respectively. The best results are in bold.

Model	COCO				LVIS						
	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L	AP^R	AP^C	AP^F
ViTDet-H	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3	35.9	46.8	51.1
SAM-B	41.1	28.3	45.6	53.7	40.8	30.1	53.0	58.5	32.6	41.9	43.3
SAM-L	45.5	30.2	50.1	60.4	43.8	31.9	56.7	64.2	34.3	44.7	46.9
SAM-H	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5	34.6	45.5	47.8
SAM-B (our impl.)	45.1	28.1	50.1	61.4	42.1	29.3	54.9	64.2	33.2	43.2	44.7
SAM-L (our impl.)	46.0	29.0	50.7	62.2	43.1	30.2	56.0	65.3	33.4	44.2	46.1
TAPv2-B (Stage1-PT)	45.1	28.7	50.1	60.6	42.2	29.5	55.0	64.0	33.6	43.1	45.0
TAPv2-L (Stage1-PT)	46.0	29.1	50.9	62.2	43.0	30.2	55.9	65.1	33.7	44.1	46.0
TAPv2-B (Stage2-PT)	45.9	28.7	50.9	62.2	42.1	28.9	55.1	64.7	33.4	43.2	44.8
TAPv2-L (Stage2-PT)	46.6	29.1	51.9	63.5	42.9	29.4	56.0	65.8	34.1	44.0	45.7
TAPv2-XL (Stage2-PT)	46.9	29.0	**52.0**	**64.1**	42.7	29.0	55.8	**65.9**	33.6	43.7	45.5

sufficient region-level information during pre-training for captioning, supporting our earlier claim that TAPv2 can function as a location-aware image tokenizer.

Zero-shot instance segmentation. We evaluate our model in zero-shot instance segmentation, a task at which the original SAM excels. Following a common practice [9, 40], we first obtain detection boxes from a ViTDet-H model [48]. And subsequently, we utilize these boxes to prompt the image decoder and compare the segmentation performance (i.e., using the box category) on COCO and LVIS. For a fair comparison, we report results from both SAM and our reproduction (denoted as our impl.). As depicted in Table 3, our model achieves comparable segmentation results with SAM across different model scales. This demonstrates that additional concept prediction tasks do not compromise SAM's original capability. Moreover, when the region captioning task is simultaneously introduced (i.e., Stage2-PT), the mask AP has an increase on the large instances. But in general, segmentation, being an elementary and geometric task, may not fully exploit the semantic representation in vision foundation models.

Multimodal understanding. We evaluate model's multimodal perception and reasoning capabilities across three visual question answering (VQA) benchmarks: VQA-v2 [65], GQA [30], and TextVQA [66], as well as recently introduced benchmarks MMBench [67] and SEED-Bench [31], which are tailored for large multimodal models. As shown in Table 4, our pre-trained image encoder can achieve comparable performance as CLIP and its variants on benchmarks such as VQA-v2 and GQA. However, when applied to text-rich images in TextVQA and MMBench, our model exhibits inferior performance. We hypothesize that CLIP's vision encoder has effectively aligned extensive textual content with images, thereby circumventing the suboptimal data composition of `LLaVA-v1.5-mix` [15]. When benchmarked on high-resolution visual understanding tasks, our model outperforms CLIP and its variants on SEED-Bench,

Table 4 Multimodal understanding performance on popular benchmarks. The model suffixes '-L', '-XL', '-SO', '-H', and '-g' correspond to vision encoders with approximately 0.3, 0.4, 0.4, 0.6, and 1 B parameters, respectively.

Pre-training	VisionEncoder	LLM	$\mathrm{VQA}^{\mathrm{v2}}$	GQA	$\mathrm{VQA}^{\mathrm{T}}$	MMB	SEED
LLaVA-v1.5	SAM-H	Vicuna-7B	57.7	57.0	43.9	–	–
LLaVA-v1.5	DINOv2-g	Vicuna-7B	76.2	61.9	47.2	–	–
LLaVA-v1.5 (LoRA)	CLIP-L	Vicuna-7B	79.1	63.0	58.2	66.1	60.1
Bunny (LoRA)	CLIP-L	Phi-2	77.0	60.7	52.6	67.2	61.3
Bunny (LoRA)	EVACLIP-L	Phi-2	78.9	62.3	49.4	67.4	62.2
Bunny (LoRA)	SigLIP-SO	Phi-2	79.8	62.5	–	68.6	62.5
TAPv2 (LoRA)	CLIP-L	Phi-2	79.0	62.0	52.4	65.5	61.2
TAPv2 (LoRA)	TAP-L	Phi-2	78.5	62.5	47.6	59.8	63.3
TAPv2 (LoRA)	TAP-XL	Phi-2	79.0	62.8	49.0	60.4	63.7

Table 5 Ablation study on pre-training loss and text prompt. Default settings are italic. The best results are in bold.

Model	Pre-train	TextPrompt	VG Caption				Segmentation	
			BLEU@4	METEOR	ROUGE	CIDEr	$\mathrm{AP}_{\mathrm{COCO}}$	$\mathrm{AP}_{\mathrm{LVIS}}$
Model A	$\mathcal{L}_{\mathrm{seg}}$	MaskToken	8.8	13.2	29.0	105.2	**46.1**	**43.2**
Model B	$\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{concept}}$	MaskToken	11.1	16.6	34.0	138.9	46.0	43.0
Model C	$\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{feat}}$	SemanticToken	11.4	16.9	34.7	143.1	46.0	43.1
Model D	$\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{concept}}$	SemanticToken	***12.4***	***17.9***	***36.2***	***154.7***	*46.0*	*43.0*

Table 6 (Color online) Ablations on pre-training tasks for zero-shot classification. Default tasks are italic. The best results are in bold. The magnitude of the increase relative to the baseline is shown in green.

VisionEncoder	Pre-train	COCO				LVIS			
		AP	AP^{S}	AP^{M}	AP^{L}	AP	AP^{R}	AP^{C}	AP^{F}
ViT-L	$\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{feat}}$	62.0	44.5	69.9	75.4	39.1	35.5	37.4	42.7
ViT-L	$\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{concept}}$	***77.0*** (*+15.0*)	***60.0***	***83.7***	***90.0***	***59.1*** (*+20.0*)	***61.7***	***58.9***	***58.3***

highlighting the image encoder's effectiveness in capturing fine-grained visual details.

4.3 Ablation study

Pre-training loss. Ablation studies on pre-training loss are shown in Tables 5 and 6, where $\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{feat}}$ and $\mathcal{L}_{\mathrm{concept}}$ represent the pre-training with segmentation, feature prediction, and concept prediction, respectively (Subsection 3.1). As presented in Table 5, caption metrics are remarkably low when pre-trained with $\mathcal{L}_{\mathrm{seg}}$ alone (Model A). When combined with semantic prediction (Model B/C), caption performance sees a significant improvement. Despite showing semantic awareness, feature prediction is inferior to concept prediction in both classification and captioning tasks. These findings indicate that a concept space is crucial for acquiring CLIP priors. We consider that the concept space has efficiently facilitated the model's learning of negative text embeddings (i.e., T_C) from CLIP. In addition, the segmentation results presented in Table 5 indicate that pre-training with additional semantic prediction neither enhances nor compromises the mask AP on COCO and LVIS. This observation also suggests that the SAM's decoder architecture could incorporate more task supervision beyond the segmentation mask.

Semantic token. To verify the effectiveness of semantic tokens, we conduct four experiments. First, we pre-train our model using the loss listed in the 'Pre-train' column. Subsequently, we fine-tune the text decoder using the items outlined in 'TextPrompt', generated from the frozen pre-trained model. Model A serves as our baseline, pre-trained with only $\mathcal{L}_{\mathrm{seg}}$. Here, the mask token is directly used for the region-level captioning task, akin to using the original SAM's output to train the text decoder. Model D is our default model, jointly optimized with promptable segmentation and concept prediction. The semantic token is used to prompt the text decoder. As demonstrated in Table 5, semantic tokens consistently outperform mask tokens in the captioning task while achieving comparable AP in the segmentation task. Eventually, the semantic token proves to be the most effective. This suggests that semantic tokenization significantly unlocks the potential of the foundation model, facilitating more perception tasks.

Scaling text decoder. We scale up the text decoder along the depth and embedding dimension to ablate the caption bottleneck. As shown in Table 7, there is no substantial improvement with an increased

Table 7 Ablation on the model scale of the text decoder. Default configurations are italic. The best results are in bold.

VisionEncoder	TextDecoder			VG caption			
	Params	Depth	Dim	BLEU@4	METEOR	ROUGE	CIDEr
ViT-L	20M	6	512	12.0	17.6	35.9	153.2
ViT-L	25M	8	512	12.2	17.7	36.0	153.9
ViT-L	*38M*	*12*	*512*	***12.4***	***17.9***	***36.2***	***154.7***
ViT-L	43M	6	768	12.3	17.8	36.0	154.2

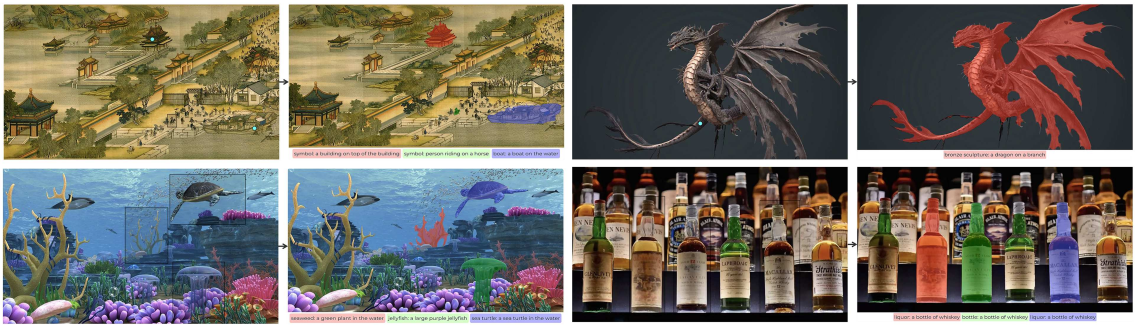

symbol: a building on top of the building
symbol: person riding on a horse
boat: a boat on the water
bronze sculpture: a dragon on a branch
seaweed: a green plant in the water
jellyfish: a large purple jellyfish
sea turtle: a sea turtle in the water
liquor: a bottle of whiskey
bottle: a bottle of whiskey
liquor: a bottle of whiskey

Figure 7 (Color online) Visualization of understanding open-world knowledge.

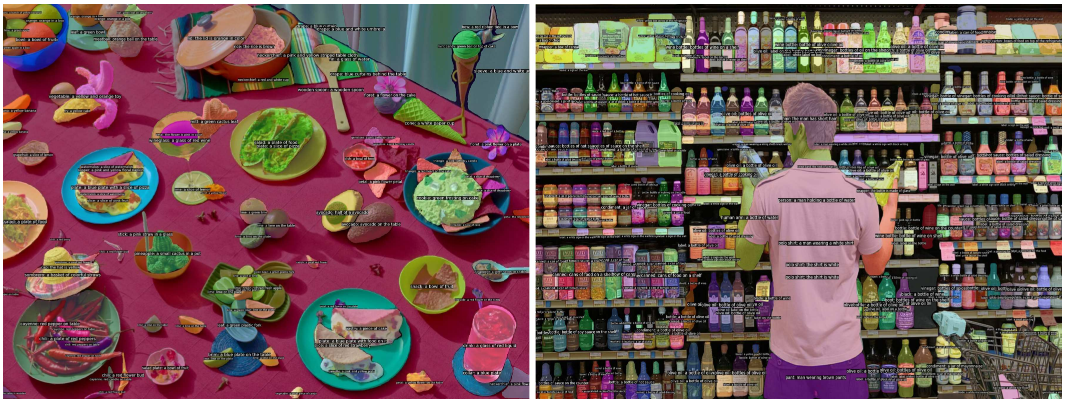


Figure 8 (Color online) Visualization of crowd understanding. Best viewed in color with zoom.

model scale on the VG dataset. This suggests that employing larger decoders for region captioning may not be necessary unless the text length and quantity can be further increased.

4.4 Qualitative results

We qualitatively evaluate TAPv2 using point-based prompts. By simply clicking or prompting with a grid of points, our model can simultaneously generate the segmentation, category, and description.

Open-world knowledge. Figure 7 showcases the instances that pose challenges in open-world scenarios. Due to the subjective nature of vocabulary design, concepts such as 'pepsi', 'cocacola', 'dragon', 'spider-man' and 'whisky' could hardly be selected via classification. However, our model demonstrates proficiency in handling these instances, indicating its capability to deal with open-world knowledge.

Crowd understanding. Figure 8 visualizes the crowd regions. Our model can accurately identify and segment various elements within crowded or bustling environments. The segmentation masks precisely outline the distinct regions occupied by people, food as well as various uncommon commodities and stationeries. Furthermore, the accompanying caption provides an overall summary.

5 Conclusion

In this study, we propose two models: TAPv2-Base and TAPv2-Chat. The first one is a promptable model capable of segmenting, recognizing, and captioning objects within arbitrary regions. To build

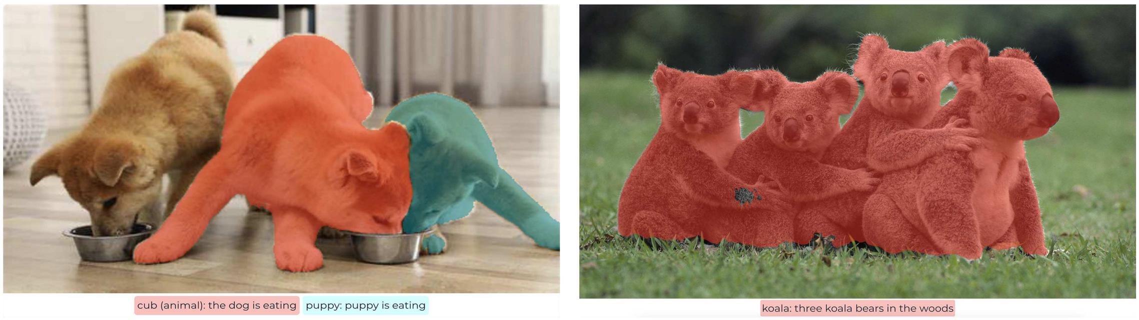

cub (animal): the dog is eating
puppy: puppy is eating
koala: three koala bears in the woods

Figure 9 (Color online) Visualization of two failure cases. The first instance demonstrates inconsistent ordering of analogous concepts, whereas the second highlights a failure in accurate object enumeration.

such a model, we explore a systematic solution that includes (1) two new datasets: SemanticSA-1B and CaptionSA-400M, (2) a novel framework: promptable tokenization, and (3) an effective learning method: concept prediction. We then successfully extend it to a descriptive model for general visual perception and reasoning tasks, and investigate its integration with a large language model to leverage synergies. As a unified framework, TAPv2 aims to advance segmenting anything towards perceiving anything via visual prompting and language instruction. We hope this work could inspire the community to develop more compact and significant vision-language foundation models.

Limitations. Despite its advancements, TAPv2 has two main constraints. It is trained using a human-curated label space, which still falls short of an open-world assumption. This constraint also leads to an unstable ranking of similar concepts during inference (Figure 9 left). Additionally, the text decoder, fine-tuned on a constrained set of region caption data, may limit the model's scalability and breadth of vision-language understanding. For example, the object counting cannot be well solved (Figure 9 right). Expanding the caption data is anticipated to instruct models for complex understandings.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2022ZD0116302). We would like to thank Hanxiao QU, Yan TIAN, and Xigang CAO for their help on Cambricon MLU resources for training, as well as other colleagues at BAAI for their support to this project.

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021
2 Zhai X, Mustafa B, Kolesnikov A, et al. Sigmoid loss for language image pre-training. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023
3 Sun Q, Fang Y, Wu L, et al. EVA-CLIP: improved training techniques for clip at scale. 2023. ArXiv:2303.15389
4 Chen Z, Wu J, Wang W, et al. Internvl: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024
5 Sun Q, Yu Q, Cui Y, et al. Emu: generative pretraining in multimodality. In: Proceedings of International Conference on Learning Representations, 2024
6 Sun Q, Cui Y, Zhang X, et al. Generative multimodal models are in-context learners. 2023. ArXiv:2312.13286
7 Chen Z, Wang W Y, Tian H, et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. Sci China Inf Sci, 2024, 67: 220101
8 Oquab M, Darcet T, Moutakanni T, et al. DINOv2: learning robust visual features without supervision. 2024. ArXiv:2304.07193
9 Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023
10 Wei H, Kong L, Chen J, et al. Vary: scaling up the vision vocabulary for large vision-language model. In: Proceedings of European Conference on Computer Vision, 2024
11 Tong S, Liu Z, Zhai Y, et al. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024
12 Lu H, Liu W, Zhang B, et al. DeepSeek-vl: towards real-world vision-language understanding. 2024. ArXiv:2403.05525
13 Jiang D, Liu Y, Liu S, et al. From clip to dino: visual encoders shout in multi-modal large language models. 2023. ArXiv:2310.08825
14 Karamcheti S, Nair S, Balakrishna A, et al. Prismatic VLMs: investigating the design space of visually-conditioned language models. In: Proceedings of International Conference on Machine Learning, 2024
15 Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024
16 Shi B, Wu Z, Mao M, et al. When do we not need larger vision models? In: Proceedings of European Conference on Computer Vision, 2024
17 Dong X, Zhang P, Zang Y, et al. InternLM-XComposer2-4KHD: a pioneering large vision-language model handling resolutions from 336 pixels to 4k HD. In: Proceedings of Conference on Neural Information Processing Systems, 2024
18 Schuhmann C, Beaumont R, Vencu R, et al. Laion-5b: an open large-scale dataset for training next generation image-text models. 2022. ArXiv:2210.08402

19 Krishna R, Zhu Y, Groth O, et al. Visual Genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis, 2017, 123: 32–73

20 Zhang Y, Huang X, Ma J, et al. Recognize anything: a strong image tagging model. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024

21 Wang W, Shi M, Li Q, et al. The all-seeing project: towards panoptic visual recognition and understanding of the open world. In: Proceedings of International Conference on Learning Representations, 2024

22 Liu H, Li C, Wu Q, et al. Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 36

23 Chen L, Li J, Dong X, et al. ShareGPT4V: improving large multi-modal models with better captions. In: Proceedings of European Conference on Computer Vision, 2024

24 He M, Liu Y, Wu B, et al. Efficient multimodal learning from data-centric perspective. 2024. ArXiv:2402.11530

25 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of European Conference on Computer Vision, 2014

26 Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. 2023. ArXiv:2303.08774

27 OpenAI. GPT-4v (ision) system card. 2023. https://openai.com/index/gpt-4v-system-card

28 Zhao B, Wu B, Huang T. SVIT: scaling up visual instruction tuning. 2023. ArXiv:2307.04087

29 Gupta A, Dollar P, Girshick R. LVIS: a dataset for large vocabulary instance segmentation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019

30 Hudson D A, Manning C D. GQA: a new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019

31 Li B, Wang R, Wang G, et al. SEED-Bench: benchmarking multimodal LLMs with generative comprehension. 2023. ArXiv:2307.16125

32 Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of International Conference on Machine Learning, 2021

33 Li Y, Fan H, Hu R, et al. Scaling language-image pre-training via masking. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023

34 Wang X, Wang W, Cao Y, et al. Images speak in images: a generalist painter for in-context visual learning. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023

35 Wang X, Zhang X, Cao Y, et al. SegGPT: towards segmenting everything in context. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023

36 Zou X, Yang J, Zhang H, et al. Segment everything everywhere all at once. In: Proceedings of Conference on Neural Information Processing Systems, 2023

37 Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. In: Proceedings of Conference on Neural Information Processing Systems, 2022

38 Lu J, Clark C, Zellers R, et al. Unified-io: a unified model for vision, language, and multi-modal tasks. In: Proceedings of International Conference on Learning Representations, 2023

39 Wang H, Vasu P K A, Faghri F, et al. SAM-CLIP: merging vision foundation models towards semantic and spatial understanding. 2023. ArXiv:2310.15308

40 Yang H, Ma C, Wen B, et al. Recognize any regions. In: Proceedings of Conference on Neural Information Processing Systems, 2024

41 Li F, Zhang H, Sun P, et al. Segment and recognize anything at any granularity. In: Proceedings of European Conference on Computer Vision, 2024

42 Zou X, Dou Z Y, Yang J, et al. Generalized decoding for pixel, image, and language. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023

43 Li J, Li D, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, 2023

44 Zhu D, Chen J, Shen X, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models. In: Proceedings of International Conference on Learning Representations, 2024

45 Microsoft. Phi-2: the surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of- small-language-models, 2023

46 Zheng L, Chiang W L, Sheng Y, et al. Judging LLM-as-a-judge with mt-bench and chatbot arena. In: Proceedings of Conference on Neural Information Processing Systems, 2023

47 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2020

48 Li Y, Mao H, Girshick R, et al. Exploring plain vision transformer backbones for object detection. In: Proceedings of European Conference on Computer Vision, 2022

49 Li Y, Wu C Y, Fan H, et al. Mvitv2: improved multiscale vision transformers for classification and detection. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 4804–4814

50 Minderer M, Gritsenko A, Houlsby N. Scaling open-vocabulary object detection. In: Proceedings of Conference on Neural Information Processing Systems, 2023

51 Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2017

52 Milletari F, Navab N, Ahmadi S, et al. Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of International Conference on 3D Vision, 2016

53 Su J, Ahmed M, Lu Y, et al. RoFormer: enhanced transformer with Rotary Position Embedding. Neurocomputing, 2024, 568: 127063

54 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, 2016

55 Cha J, Kang W, Mun J, et al. Honeybee: locality-enhanced projector for multimodal LLM. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024

56 McKinzie B, Gan Z, Fauconnier J P, et al. Mm1: methods, analysis & insights from multimodal llm pre-training. 2024. ArXiv:2403.09611

57 Hu E J, Shen Y, Wallis P, et al. Lora: low-rank adaptation of large language models. In: Proceedings of International Conference on Learning Representations, 2022

58 Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts. Neural Comput, 1991, 3: 79–87

59 Yao L, Han J, Wen Y, et al. Detclip: dictionary-enriched visual-concept paralleled pre-training for open-world detection. In: Proceedings of Conference on Neural Information Processing Systems, 2022

60 Zhong Y, Yang J, Zhang P, et al. Regionclip: region-based language-image pretraining. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022

61 Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017. 633–641

62 Shao S, Li Z, Zhang T, et al. Objects365: a large-scale, high-quality dataset for object detection. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019

63 Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset V4. Int J Comput Vis, 2020, 128: 1956–1981

64 Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016

65 Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017

66 Singh A, Natarajan V, Shah M, et al. Towards VQA models that can read. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019

67 Liu Y, Duan H, Zhang Y, et al. MMBench: is your multi-modal model an all-around player? In: Proceedings of European Conference on Computer Vision, 2024

68 Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of International Conference on Learning Representations, 2019

69 Loshchilov I, Hutter F. Sgdr: stochastic gradient descent with warm restarts. In: Proceedings of International Conference on Learning Representations, 2017

70 Ghiasi G, Cui Y, Srinivas A, et al. Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021

71 Huang G, Sun Y, Liu Z, et al. Deep networks with stochastic depth. In: Proceedings of European Conference on Computer Vision, 2016

72 Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res, 2014, 15: 1929–1958

73 He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022

74 Wu J, Wang J, Yang Z, et al. GRiT: a generative region-to-text transformer for object understanding. In: Proceedings of European Conference on Computer Vision, 2024

75 Zhang S, Sun P, Chen S, et al. GPT4ROI: instruction tuning large language model on region-of-interest. 2023. ArXiv:2307.03601

76 Sun Z, Fang Y, Wu T, et al. Alpha-CLIP: a clip model focusing on wherever you want. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024

77 Liu Z, He Y, Wang W, et al. Internchat: solving vision-centric tasks by interacting with chatbots beyond language. 2023. ArXiv:2305.05662

78 Geng X, Liu H. OpenLLaMA: an open reproduction of LLaMA. 2023. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models

79 Huang X, Wang J, Tang Y, et al. Segment and caption anything. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024