Auto-Controlled Image Perception in MLLMs via Visual Perception Tokens
- *Supplementary Material* -

S1. Implement Details

S1.1. Training Details

Our training process consists of two phases: alignment and finetuning. The alignment stage aligns the additional vision features with the LLM embeddings. If the original vision encoder is used for re-encoding, the alignment stage is omitted. We use the same image-text pair data for the LLaVA 1.5 alignment, and only use the additional vision branch as the LLM's input. During training, all components except the projector are frozen. In this phase, we train the model for 1 epoch with a learning rate of 2e-3 and a batch size of 128. The second finetuning stage allows the model to learn to output the correct Region Selection Tokens and to transmit information through the Vision Re-Encoding Tokens. We finetune the model using our constructed dataset, as well as remaining samples from the LLaVA 1.5 finetuning dataset that were not included in our dataset. In this stage, all components except the original visual encoder and the additional vision encoder are unfrozen. In this phase, we train the model for 1 epoch with a learning rate of 2e-5 and a batch size of 256. For both the first and the second phase, we use AdamW optimizer. The experiments are deployed on 8 A100 GPU. The total training time is about 20 hours. For the 7B model, the rank of the LoRA is set to 512.

S1.2. Evaluation Prompt

Following established practices [2, 9], we used GPT-4o (2024-08-06) to evaluate the alignment between the model's responses and the ground truth for each question. We use the evaluation prompt in [6].

> Evaluation Prompt
>
> You are responsible for proofreading the answers, you need to give a score to the model's answer by referring to the standard answer, based on the given question. The full score is 1 point and the minimum score is 0 points. Please output the score in the form 'score: <score>'. The evaluation criteria require that the closer the model's answer is to the standard answer, the higher the score.
> Question: <question>
> Ground Truth: <ground truth>
> Answer: <answer>

S1.3. Template of the Training Data Examples

Here, we show the format of our training examples. The training example for the Region Selection Token is essentially the same as the samples used in [6], except that the method for representing regions has changed from bounding boxes to region tokens. The training example for the Vision Re-Encoding Token is almost identical to the data in the original LLava [3] finetuning dataset, with the only difference being the insertion of an additional round of dialogue between the original question and answer. This added dialogue includes the Vision Re-Encoding Token.

> Template of Training Example for Region Selection Token
>
> **User**: <image> <question> Please identify the region that can help you answer the question better, and then answer the question.
> **Assistant**: <Region_Selection_Start> <x_min> <y_min> <x_max> <y_max> <Region_Selection_End>.
> **User**: <image>
> **Assistant**: <ground truth>

	MME		MMB	
	Cognition	Perception	en	cn
Qwen2-VL-2B	1434	280	78.20	77.30
Qwen2-VL-7B	1664	335	78.70	**83.30**
2B+VPT (DINO)	1511	274	79.11	76.64
2B+VPT (CLIP)	1510	273	79.53	77.41
2B+VPT (SAM)	1475	270	80.22	76.99
7B+VPT (CLIP)	**1706**	**336**	**83.80**	**83.30**

Table S1. Performance comparison of MLLMs with and without Visual Perception Tokens on MME and MMBench Benchmarks.

Template of Training Example for Vision Re-Encoding Token

User: <image> <question> Please require additional perception features, and then answer the question.
Assistant: <Re-Encoding_Start> <Re-Encoding_Control> <Re-Encoding_End>.
User: <image>
Assistant: <ground truth>

The training for the free-choice experiment differs from other experiments only in the sample template. For the free-choice experiment, we removed the additional prompt from the questions. The training sample template is as follows.

Template of Training Example for Region Selection Token (Free Choice)

User: <image> <question>
Assistant: <Region_Selection_Start> <x_min> <y_min> <x_max> <y_max> <Region_Selection_End>.
User: <image>
Assistant: <ground truth>

Template of Training Example for Vision Re-Encoding Token (Free Choice)

User: <image> <question>
Assistant: <Re-Encoding_Start> <Re-Encoding_Control> <Re-Encoding_End>.
User: <image>
Assistant: <ground truth>

S2. Supplementary Experiments

We conducted experiments on the MME [1] and MM-Bench [4] benchmarks without using the Visual Perception Token, allowing the model to generate answers directly. This assessed the impact of our fine-tuning on general benchmarks. Results in Tab. S1 show that our model does not cause degeneration and even improves performance on these benchmarks.

To verify the advantage of the Region Selection Token over direct BBox prediction, we compared the predicted regions with ground truth using IoU and Intersection over Ground Truth (IoGT), defined as:

$$(\text{IoGT} = \frac{\text{Area of } (GT \cap \text{Pred})}{\text{Area of } GT}).$$

Results in Tab. S2 show that Region Selection Token significantly outperforms direct BBox prediction in accuracy.

S3. Further Examples

Here we present additional examples obtained using the visual perception token. Figs. S1 and S2 include the responses generated with the Vision Re-Encoding Token. Figs. S3 and S4 present the responses generated with the Region Selection Token, with the regions selected by the Region Selection Token highlighted in the images.

	Metric	DocVQA	TextVQA	TextCap
Directly Predicting BBox	IoU	0.15	0.26	0.25
	IoGT	0.20	0.28	0.27
Using Region Selection Token	IoU	0.26	0.56	0.50
	IoGT	0.38	0.71	0.66

Table S2. Performance comparison of MLLMs with and without Visual Perception Tokens on MME and MMBench Benchmarks.

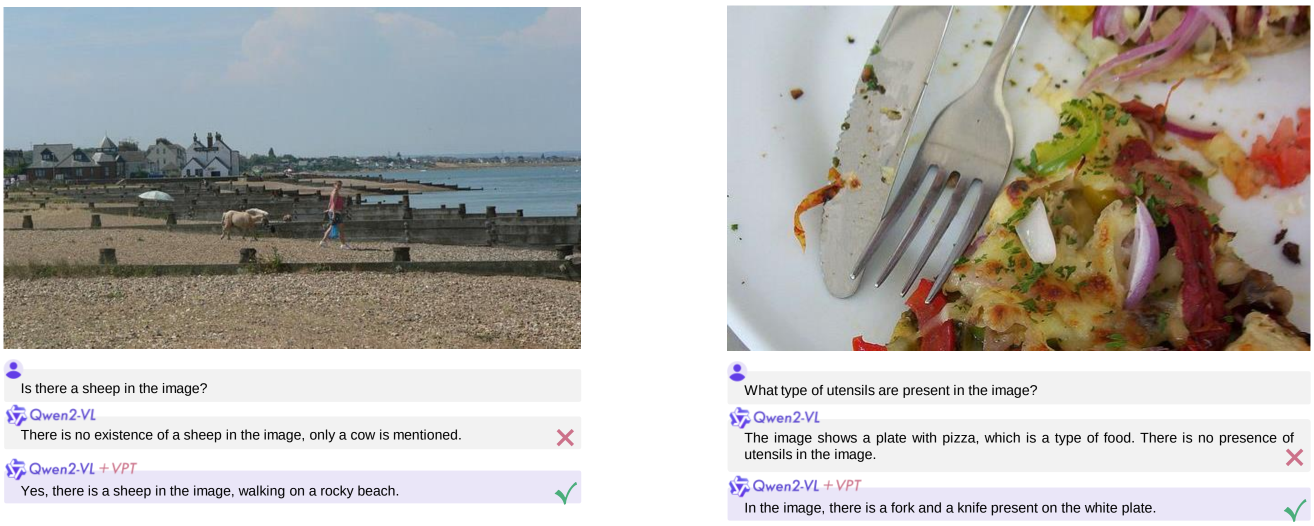


Figure S1. This set of images demonstrates how the DINO Feature Token assists MLLMs in identifying specific objects within images. These objects are often difficult for MLLMs to recognize directly due to their small size or interference from surrounding objects.

S4. Additional Related Works

S4.1. Reasoning Token

In Large Language Model (LLM), there are tokens, similar to Visual Perception Token, which are designed to control the generation process of LLM. These token are termed reasoning tokens or planning token and have recently been introduced in OpenAI's o1 model [5] and other LLMs. For example, to enhances models' reasoning capabilities, reasoning tokens were explicitly integrated into OpenAI's o1 models to segment prompts into smaller, manageable parts, exploring multiple response strategies before generating the final output [5]. Similar methods aim to incorporate CoT reasoning into language models through planning tokens or distillation techniques. For example, a hierarchical generation framework using planning tokens has been proposed, embedding high-level plans at each reasoning stage with minimal parameter increase [7]. Moreover, techniques like Rephrase and Respond have been distilled back into models, improving efficiency and accuracy in reasoning, as demonstrated in [8].

Our work focuses on MLLMs, where we design visual perception tokens to enhance the visual perception capabilities of MLLMs, not for LLM. Moreover, our exploration goes beyond LLM reasoning tokens. Unlike these tokens, which merely trigger specific actions and lack the ability to convey detailed instructions or rich information, we focus on designing tokens capable of transmitting nuanced control information for fine-grained visual perception.

S5. Discussion

Adaptability of Visual Perception Token. The design of the visual perception token depends on the specific visual perception method. In this paper, we use Crop and the addition of vision features as examples to introduce two types of visual perception tokens. However, our approach can be extended to other visual prompting techniques or visual encoder models, and even to LLM-agent or LLM-tool systems beyond vision.

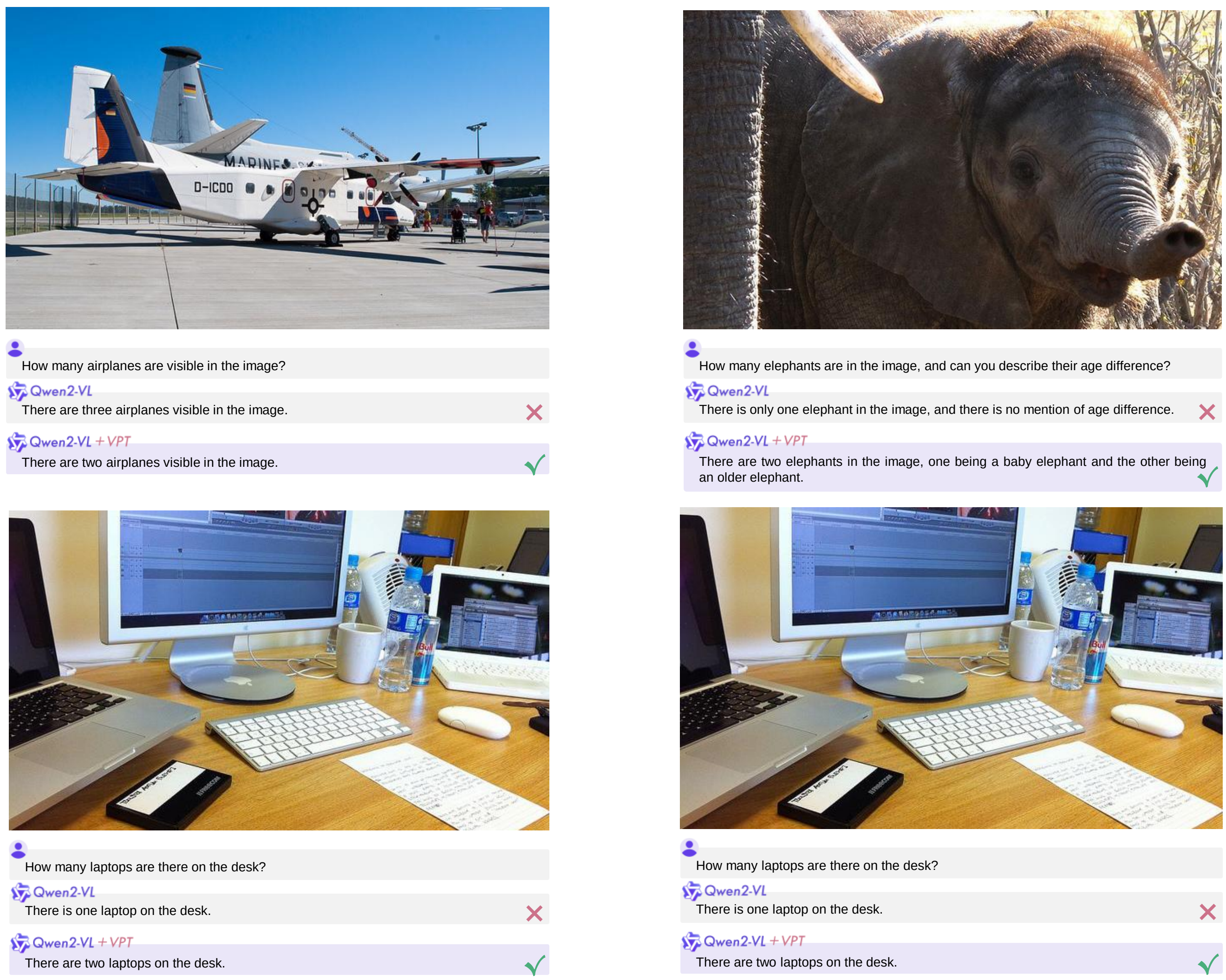
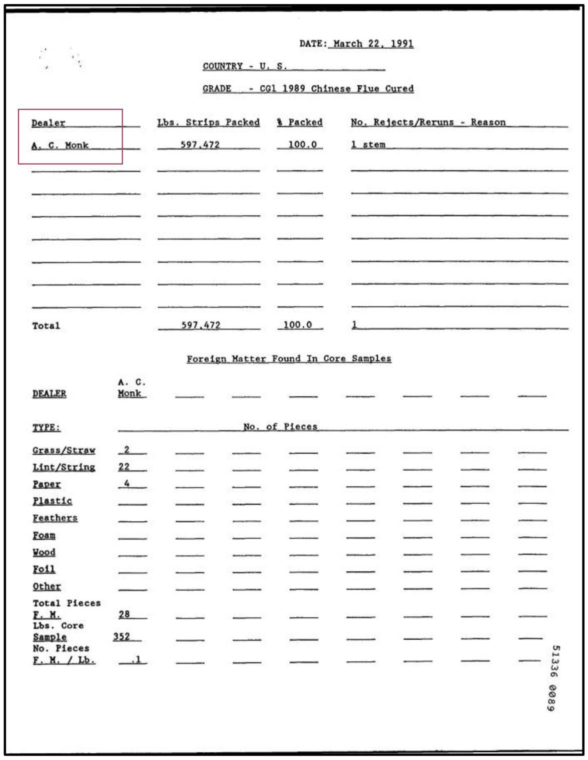


Figure S2. This set of images illustrates how the DINO Feature Token assists MLLMs in counting the number of objects in an image. Counting has long been a significant limitation for MLLMs. By leveraging the DINO Feature, the DINO Feature Token enables precise localization of individual objects within the image, thereby improving the counting capability of MLLMs.



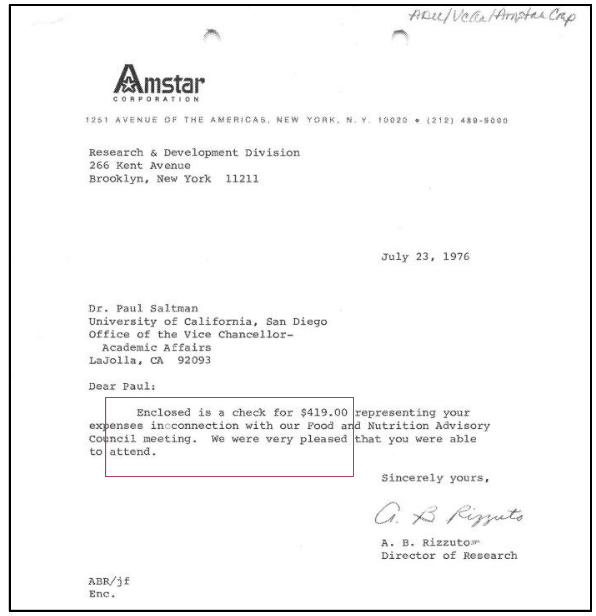
What is the name of the Dealer?

Qwen2-VL

The dealer is C. York

Qwen2-VL + VPT

The dealer is A. C. Monk.



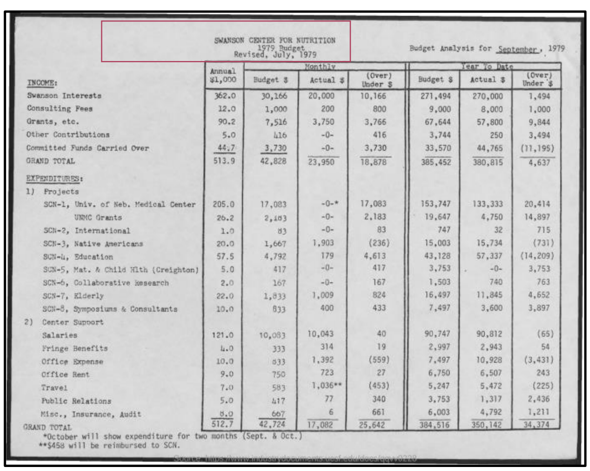
What is the amount of the check enclosed?

Qwen2-VL

The check enclosed is for $1,000.00.

Qwen2-VL + VPT

The amount of the check enclosed is $419.00.



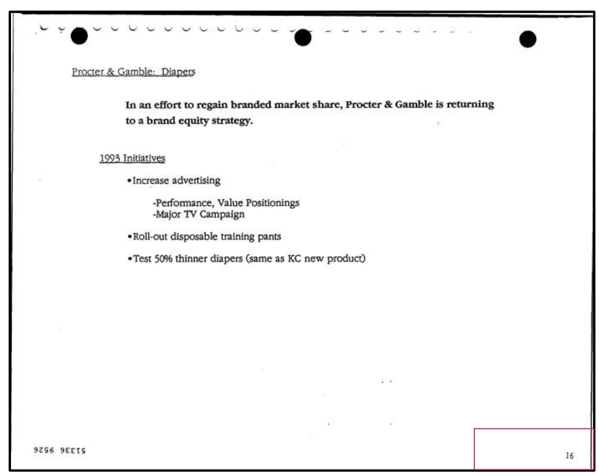
What is the name of the institute mentioned in the title?

Qwen2-VL

The name of the institute mentioned in the title is "Sanford Center for Nutrition."

Qwen2-VL + VPT

The name is “SWANSON CENTER FOR NUTRITION”.



What is the page no mentioned in this document?

Qwen2-VL

The page number mentioned in the document is 72.

Qwen2-VL + VPT

16.

Figure S3. This group of examples shows how the Region Selection Token aids MLLMs in understanding textual information within images by correctly identifying the corresponding regions. The image inputs primarily consist of large but structured documents, such as tables, forms, or letters.

What is the name of the bar?

Qwen2-VL

The name of the bar is "Sushi Bar." ✗

Qwen2-VL + VPT

The name of the bar is "moon bar." ✓

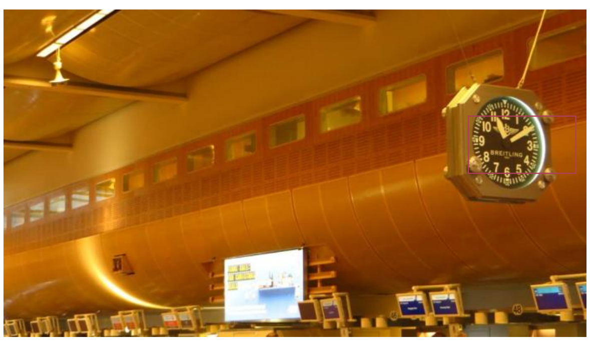
What is the number on the sign behind the player?

Qwen2-VL

The number on the sign behind the player is 24. ✗

Qwen2-VL + VPT

The number on the sign behind the player is 340. ✓



What brand is the clock?

Qwen2-VL

The clock in the image is a mechanical clock, and it does not appear to be branded. ✗

Qwen2-VL + VPT

The brand is Breitling. ✓

What kind of brick does the box mention?

Qwen2-VL

The box mentions "American Beauty" bricks. ✗

Qwen2-VL + VPT

The box mentions "arto" bricks. ✓

Figure S4. This set of images illustrates how the Region Selection Token enables MLLMs to comprehend textual information within real-world scenes by accurately identifying the corresponding regions. The image inputs consist of real-world scenarios, such as signboards and trademarks, where the text occupies only a small portion of the overall scene and is highly susceptible to interference from the surrounding context.

References

[1] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. 2

[2] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1

[4] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281, 2023. 2

[5] OpenAI. How reasoning works, 2024. Accessed: 2024-11-05. 3

[6] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024. 1

[7] Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. Guiding language model reasoning with planning tokens. In *First Conference on Language Modeling*, 2024. 3

[8] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1, 2024. 3

[9] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. 1