Chain-of-Visual-Thought: Teaching VLMs to See and Think Better with Continuous Visual Tokens

Yiming Qin[1], Bomin Wei[2], Jiaxin Ge[1], Konstantinos Kallidromitis[3], Stephanie Fu[1], Trevor Darrell[1], and XuDong Wang[1⋆]

[1] University of California, Berkeley, CA, USA
[2] University of California, Los Angeles, CA, USA
[3] AI Laboratory, Panasonic R&D Company of America, USA

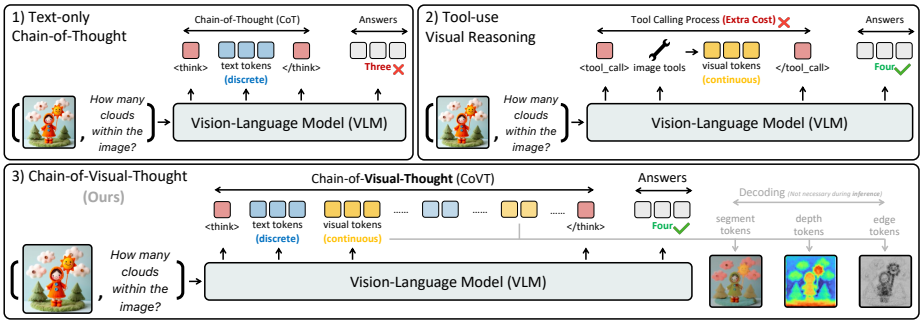

Fig. 1: Traditional VLM reasoning is restricted to the discrete language space (*Text-only CoT*) or relies on costly external tool calls to obtain visual signals (*Tool-use reasoning*). **In contrast, CoVT forms a visual thought chain that enables VLMs to reason in continuous visual space.** By introducing *continuous visual tokens* that encode perceptual cues (*e.g.*, segmentation, depth, instance, and edge structure), CoVT composes *chains of textual and visual thoughts* that link semantic reasoning with perceptual grounding. These visual "thought chains" bridge language and vision, enabling fine-grained understanding, spatial precision, and geometric awareness beyond the reach of text-based reasoning. During inference, these visual tokens can be decoded into different types of visual information when interpretation is desired.

Abstract. Vision–Language Models (VLMs) excel at reasoning in linguistic space but struggle with perceptual understanding that requires dense visual perception, *e.g.*, spatial reasoning and geometric awareness. This limitation stems from the fact that current VLMs have limited mechanisms to capture dense visual information across spatial dimensions. We introduce Chain-of-Visual-Thought (CoVT), a framework that enables VLMs to reason not only with discrete text tokens but also through continuous visual tokens—compact latent representations that encode rich perceptual cues. With a small budget of roughly 20 tokens, CoVT distills knowledge from lightweight vision experts that capture complementary properties such as 2D appearance, 3D geometry, spatial

⋆ Corresponding author.

layout, and edge structure. During training, the VLM with CoVT autoregressively predicts these visual tokens to reconstruct dense supervision signals (*e.g.*, depth, segmentation, edges, and DINO features). At inference, the model reasons directly in the continuous visual latent space, preserving efficiency while optionally decoding dense predictions for interpretability. Evaluated across more than ten diverse benchmarks, including CV-Bench, MME-RealWorld, MMVP, RealWorldQA, MMStar, WorldMedQA, and HRBench, integrating CoVT into strong VLMs such as Qwen2.5-VL and LLaVA consistently improves performance by 3% to 16% and demonstrates that compact continuous visual thinking enables more precise, grounded, and interpretable multimodal intelligence.

1 Introduction

Vision–Language Models (VLMs) [2,3,12,15,31,44,45,58,65,71] have become the cornerstone of modern multimodal intelligence, achieving remarkable progress in understanding and reasoning across text and vision. By projecting visual input into a language-centric token space, VLMs inherit the strong compositional and logical reasoning capabilities of large language models (LLMs), enabling unified multimodal interaction through natural language. Recent advances in text-based Chain-of-Thought (CoT) reasoning [54] further extend this paradigm, showing that structured intermediate reasoning steps can significantly enhance performance on tasks involving logic, mathematics, and knowledge grounding. However, despite these successes, such reasoning remains fundamentally *language-bound*.

When continuous visual information is projected into discrete text space, *rich perceptual cues,* e.g.*, boundaries, layout, depth, and geometry, are lost or poorly represented.* Yet these are precisely the fine-grained signals that humans rely on when reasoning about the visual world. Consequently, current VLMs often struggle with perception-intensive tasks such as counting, spatial correspondence, or relative depth estimation, even when equipped with powerful vision encoders [17, 39, 51], as shown in Fig. 1. Moreover, by forcing vision reasoning through a discrete text bottleneck, the model must verbalize continuous spatial and geometric relations. As a result, text-only CoT can misdirect and even *degrade* visual reasoning performance, as shown by Qwen3-VL-Thinking [3, 59], which performs over 5% worse than Qwen3-VL-Instruct with language CoT on spatial understanding benchmarks such as V* [56], HRBench8k [50], and VSI-Bench [60]. This exposes a fundamental limitation: ***visual information is inherently continuous and high-dimensional, yet existing models reason over it using symbolic language tokens that lack the fidelity of complex perceptual reasoning.***

A natural solution is to augment VLMs with external vision tools [21, 43], leveraging pre-built specialized models to recover fine-grained perception. While this approach can partially restore spatial and geometric information, it also introduces significant drawbacks: perception is delegated to external tools, and outcomes are bounded by them. It also introduces a higher GPU cost. Another

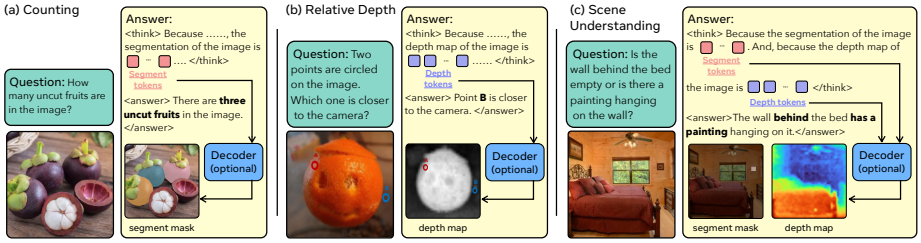

Fig. 2: Continuous visual thinking with CoVT. CoVT introduces compact, continuous visual tokens that encode fine-grained perceptual cues, such as object localization, spatial structure, and scene semantics, directly into VLM reasoning. These tokens ground multimodal reasoning in visual space, enabling the model to capture fine-grained relationships across vision-centric tasks (*e.g.*, counting, depth ordering, and scene understanding) without relying on external tools. They can also be decoded into dense predictions, offering human-interpretable visualizations of the model's reasoning process.

solution is generating or cropping images in the thinking process. However, these solutions still project the images into the text space, losing the dense visual information. These limitations motivate a central question: ***Can VLMs learn to reason the way humans do, by thinking visually rather than translating everything into words?*** More concretely, can we inject fine-grained visual signals directly into a VLM's reasoning process, allowing it to "see" and "think" simultaneously while remaining efficient and self-contained? Yes! We propose **C**hain-**o**f-**V**isual-**T**houghts **(CoVT)**.

CoVT enables reasoning over rich perceptual cues by grounding VLMs in continuous visual token space. Each group of visual tokens corresponds to a lightweight perceptual expert (*e.g.*, segmentation, depth, edge detection, or self-supervised representation learning) that encodes specific visual features. During training, the VLM is asked to *predict* these continuous visual tokens within its reasoning chain, compressing rich perceptual information into a compact latent space. These latent tokens are then decoded by task-specific lightweight decoders to reconstruct the corresponding expert targets (*e.g.*, segmentation masks, dense depth maps, edge maps, or DINO features). We backpropagate the reconstruction and distillation losses through the continuous tokens, aligning the model's internal latent representations with expert guidance. This process allows CoVT to internalize fine-grained perceptual knowledge directly into its token space, enabling grounded reasoning without explicit visual maps or external tool calls, as shown in Fig. 1.

More specifically, we highlight different aspects of fine-grained visual reasoning. We integrate both task-oriented experts (*e.g.*, SAM [25], DepthAnything v2 [61], PIDINet [42]) and representation-based experts (*e.g.*, DINO [6], contrastive encoders), with alignment strategies tailored to each: task-oriented signals are aligned at the prompt level, while representation-based signals are aligned in feature space. Training proceeds through four stages, *including com-*

prehension, generation, reasoning, and efficient reasoning, gradually teaching the model to reason effectively with visual thoughts.

At inference, the model forms *chains of visual thoughts*, reasoning across modalities to produce answers that are both semantically coherent and perceptually grounded. This self-contained, differentiable process enables VLMs to "think" directly in continuous visual space, thereby providing a more faithful bridge between internal reasoning and perceptual understanding. Moreover, this design supports interpretable multimodal intelligence, allowing users to visualize the model's visual thinking process when desired, as shown in Fig. 2. If visualization is not required, CoVT can operate solely on the continuous visual tokens without decoding them into dense predictions, thus maintaining efficiency.

Evaluated across diverse perception benchmarks, CoVT consistently improves fine-grained visual reasoning, outperforming strong VLM baselines on vision-centric tasks while maintaining competitive performance on general (non-vision-centric) benchmarks. For example, CoVT achieves a 5.5% overall gain on CV-Bench [46], delivering a substantial 14.0% improvement on its depth subtask, and 4.5% overall gain on HRBench [50]. In addition, CoVT offers flexible interpretability: the continuous visual tokens can be decoded into human-readable dense predictions, providing a window into the model's underlying visual reasoning process when desired. Together, these results demonstrate that compact continuous visual thinking enables more precise, grounded, and interpretable multimodal intelligence.

The main elements of our contribution are as follows:

- We propose Chain-of-Visual-Thought, a framework that equips VLMs with the ability to reason through *continuous visual tokens*, compact perceptual representations that serve as the building blocks for multimodal thinking.
- We develop tailored alignment strategies and a training pipeline (*comprehension, generation, reasoning, and efficient reasoning*) that enable VLMs to learn, interpret, and reason effectively within continuous visual space.
- We demonstrate consistent performance gains across diverse benchmarks, showing that continuous visual tokens enhance both perceptual grounding and interpretability.

2 Related Work

Tool-Augmented Reasoning Equipping VLMs with external tools enables them to use specialized vision models for targeted visual tasks [21, 34, 43, 55, 63]. While this improves performance, it also introduces computational overhead. Moreover, tool usage is inherently constrained, as the final performance is bounded by the ability of each tool instead of the reasoning process itself. In this work, we consider self-contained visual reasoning, which conducts reasoning flexibly and does not rely on external vision tools.

Text Space Reasoning Text space reasoning methods, such as Chain-of-Thought [26], have achieved big success in language reasoning [35, 52–54], solving problems like math, science, and logical reasoning. The strong performance of LLMs

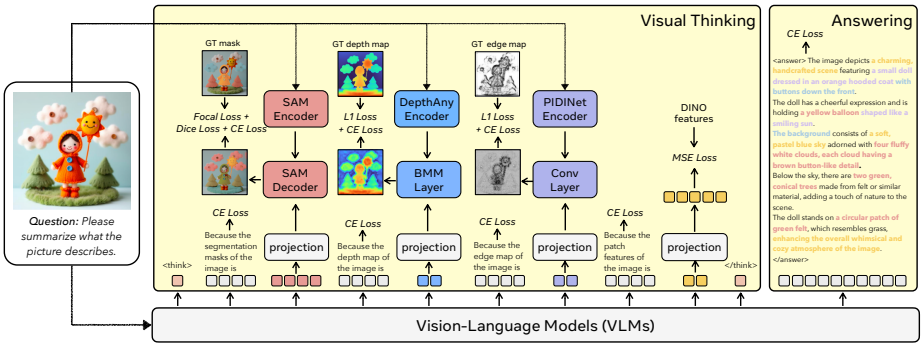

Fig. 3: The training pipeline of CoVT. CoVT first generates the thinking process, containing visual thinking tokens, and then leverages these visual thoughts to condition next token prediction and reason the final answer. To endow these tokens with perceptual meaning, we align them with lightweight vision experts (*e.g.*, SAM, DepthAnything, PIDINet, DINO) on their respective tasks during training. Specifically: SAM uses 8 visual tokens as mask prompts; DepthAnything uses 4 tokens to reconstruct depth; PIDINet uses 4 tokens to reconstruct edges; and DINO uses 4 tokens to match patch-level features. The VLM is finetuned with LoRA and all the projection layers are trainable. ***Note: During inference, dense predictions are decoded only when interpretability is desired; otherwise, reasoning occurs entirely in the latent visual space.***

with CoT capabilities has led to their broad adoption and success in models such as DeepSeek-R1 [11]. With the success of CoT, many works have extended reasoning to the visual modality. A straightforward approach is to generate dense captions and reason in language space [33], but this process is inherently lossy.

We compare CoVT with recent multimodal reasoning paradigms in Tab. 1. Visual CoT [41] relies on textual interpretations of images, limiting reasoning to the discrete text space. MCoT [10] enables continuous visual reasoning by editing or generating supplementary images, but requires substantial compute and lacks flexibility. VChain [24] interleaves images and text in the reasoning chain, yet still loses visual information by projecting images into text space. CoVT uniquely combines continuous visual reasoning, dense perceptual cues, and 3D-aware understanding within a single self-contained framework.

Latent Space Reasoning Concurrent work shows that reasoning in latent space can strengthen LLMs in complex, multi-step tasks [5, 7]. Coconut [22] finds that continuous latent embeddings are more efficient than explicit CoT, while CCoT [9] compresses CoT into continuous tokens for denser reasoning. Other studies explore specialized reasoning tokens [19] or use hidden states as implicit reasoning paths [13]. Latent reasoning has also been extended to VLMs. Aurora [4] employs VQ-VAE latents of depth and detection signals to enhance depth estimation and counting, whereas Mirage [62] uses latent imagination for visual reasoning tasks. LVR [28] achieves reasoning in shared semantic and language space by reconstructing key visual tokens. Monet [49] introduces a novel reinforcement learning algorithm to optimize latent visual reasoning. Different from these works (shown in Tab. 1), our work, CoVT, builds upon these previous

Desired Properties	VCoT	MCoT	VChain	Aurora	Mirage	LVR	Monet	Ours
Inferences without relying on external tools	✓	✗	✓	✓	✓	✓	✓	✓
Reasons in the continuous visual space	✗	✓	✓	✗	✓	✓	✓	✓
Leverages dense visual information for reasoning	✗	✓	✗	✓	✗	✗	✗	✓
Has any type of perception enhancement	✗	✗	✗	✗	✗	✗	✗	✓

Table 1: Comparison of key properties with prior multimodal reasoning methods. Unlike prior methods such as VCoT [41], MCoT [68], VChain [24], Aurora [4], Mirage [30], LVR [28], and Monet [49], CoVT uniquely satisfies all desired properties: it reasons in continuous visual space, leverages dense visual cues, incorporates any type of perception, and operates fully without external tools. Desired and undesired properties are shown in green and magenta, respectively.

contributions: we introduce a form of implicit tool-use directly embedded in dense continuous latent space, where the implicit 'tools' are ***any types of*** visual thinking tokens tied to their specific perceptual experts.

3 Chain-of-Visual-Thought (CoVT)

We first introduce the preamble in Sec. 3.1. We then show the overall pipeline of CoVT in Sec. 3.2. We also discuss how we select the visual token categories and explain how different visual tokens are aligned (Sec. 3.3). Finally, we present the model training pipeline, *e.g.*, the training loss formulation and the data framework design in (Sec. 3.4).

3.1 Preamble

Existing VLMs face two key limitations in fine-grained visual reasoning. **1)** ***Text-only CoT accumulates errors.*** Text-only CoT executes a long chain of thought, which may generate errors at the early stage. These mistakes will accumulate and ultimately lead to an incorrect final result. Therefore, we need a reasoning that is short and effective. **2)** ***Supervision is dominated by text responses,*** which provides little incentive for the model to capture *low-level perceptual cues* such as edges, depth, or regions. We need to equip VLMs themselves with the capability of extracting fine-grained visual information from the image, which can be further decoded by vision decoders.

CoVT intends to provide a foundation for the next generation of multimodal reasoning systems, capable of thinking fluidly across both language and vision in a self-contained, interpretable manner.

3.2 CoVT Overall Pipeline

We propose CoVT, a framework that augments VLMs with ***chains of visual thoughts***. Fig. 3 illustrates the overview of CoVT pipeline. Essentially, this

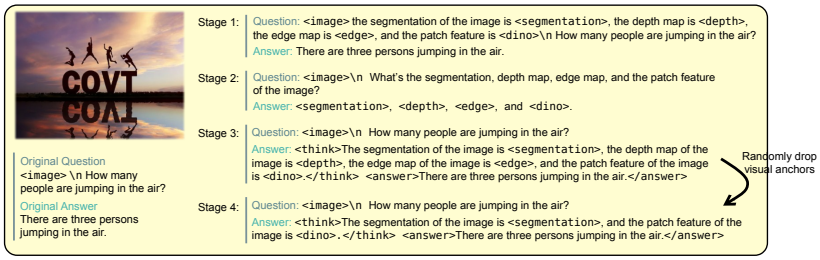

Fig. 4: Four-stage data formatting for CoVT. The first stage helps the model comprehend the visual tokens, and the second stage guides it to generate them. The third stage enables the VLM to integrate visual tokens into its reasoning process, while the final stage allows the model to efficiently select and utilize visual thinking tokens within visual thought chains.

framework equips VLMs with the capability of outputting fine-grained visual representations within a continuous visual token space, enabling them to reason directly over rich perceptual information and maintain spatial and geometric coherence throughout the reasoning process.

At its core, CoVT retains the standard next-token prediction paradigm. For standard VLMs, given visual features $\mathcal{V}$ extracted from a frozen vision encoder and text features $\mathcal{T}$ from a language encoder, the VLM estimates the probability of generating a sequence $Y = (y_1, y_2, \ldots, y_n)$ as:

$$P(Y \mid \mathcal{V}, \mathcal{T}; \theta) = \prod_{i=1}^{n} P\left(y_i \mid y_{<i}, \mathcal{V}, \mathcal{T}\right). \tag{1}$$

As shown in Fig. 3, CoVT extends this formulation by introducing *Chain-of-Visual-Thought tokens*, where each token y_i can represent either a visual token or a text token.

To effectively incorporate CoVT tokens into the VLM, we train the model to function as a dense visual encoder capable of generating multiple visual tokens that capture diverse fine-grained perceptual cues. The VLM is trained to generate CoVT tokens that, through task-specific decoders, reconstruct visual outputs under reconstruction supervision. Through this process, CoVT evolves to generate rich, fine-grained visual information across multiple perceptual dimensions within a thinking chain.

3.3 CoVT Tokens

Token selection based on core perception ability. As proposed in [70], the vision-centric perceptual ability of VLMs can be summarized as *(i) instance recognition*, *(ii) 2D and 3D spatial relationships*, *(iii) structure detection*, and *(iv) deep mining of semantic information*. Based on this categorization, we use four vision models to supervise CoVT tokens to learn each ability: *1) Segmentation tokens provide instance-level position and shape information*, which endow

VLMs with the instance recognition signals and 2D spatial perception. *2) Depth tokens provide pixel-level depth information*, equipping VLMs with the capability of figuring out 3D spatial relationships. *3) Edge tokens provide geometry-level details*, which assist models to detect structural cues and partially provide 2D spatial information. *4) DINO tokens provide the patch-level representation of the images*, delivering rich semantic information.

Tokens alignment based on granularity of visual models. Task-oriented models and representative models produce outputs at different levels of granularity. In general, task-oriented models tend to be more fine-grained, whereas representative models are usually less fine-grained. We adopt different strategies to align each type of token with the visual models based on different granularities. Essentially, we adopt two main alignment methods: For fine-grained task-oriented models, visual tokens are projected to the prompt space and then aligned at the prompt level with the decoders, while for representative models, alignment with the encoders is applied at the feature level after the projection. The projection layer consists of one multi-head attention layer and two full connected layers.

1) Segmentation tokens are supervised by SAM [25], which is a *task-oriented model* that contains dense visual features. Therefore, following LISA [27], we align Segmentation tokens with the SAM decoder. The 8 Segmentation tokens are aligned at the prompt level, and each token prompts one mask, formulated:

$$\hat{M}_i = \text{Decoder}(T_i^{\text{sam}}, f), \quad \hat{M}_i \in [0,1]^{H \times W}, \tag{2}$$

where $\hat{M}_i$ is the ith decoded mask, T_i^{sam} denotes the ith segmentation token predicted in CoVT, serving as the prompt fed into the SAM decoder, and f means the dense embedding from the SAM encoder. During the training process, the Hungarian matching algorithm is employed to match the predicted masks with the ground truths, while dice loss and focal loss are applied.

2) Depth tokens are supervised by DepthAnything v2 [61]*–a task-oriented model*. Since these tokens contain dense information, they are also aligned with the decoder at the prompt level. We use 4 Depth tokens to serve as 4 prompts to interact with the dense features extracted by DepthAnythingv2 through batch matrix multiplication (BMM) to reconstruct the depth map, formulated as:

$$\hat{D}_i = \text{softmax}\left(T_i^{\text{depth}} \cdot F_i^{\text{depth}\top}\right), \tag{3}$$

where $\hat{D}_i$ denotes the ith reconstructed depth map, T_i^{depth} represents the ith depth visual token, and F_i^{depth} is the ith middle layer feature from DepthAnything encoder. The final depth map is $\hat{D} = \frac{\sum_{i=0}^{3} \hat{D}_i}{4}$. The L1 reconstruction loss is employed for aligning Depth tokens.

3) Edge tokens are aligned with PIDINet [42]. Each of 4 Edge tokens functions as an 1×1 convolutional kernel applied to the dense features from PIDINet encoder to reconstruct the ith edge map $\hat{E}_i$. The final edge map is $\hat{E} = \frac{\sum_{i=0}^{3} \hat{E}_i}{4}$, and aligned via L1 loss function.

4) DINO tokens are supervised by DINOv2 [37], which is trained as the *representative model*, extracting patch-level features. Therefore, the 4 DINO tokens are mapped into the same shape as the DINO feature using the projection layer, and aligned under an MSE objective.

3.4 CoVT Training

Training Loss. During training, the *joint loss function* is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \gamma\big(\lambda_{\text{seg}} \cdot \mathcal{L}^{\text{seg}}_{\text{visual}} + \lambda_{\text{depth}} \cdot \mathcal{L}^{\text{depth}}_{\text{visual}} + \lambda_{\text{edge}} \cdot \mathcal{L}^{\text{edge}}_{\text{visual}} + \lambda_{\text{dino}} \cdot \mathcal{L}^{\text{dino}}_{\text{visual}}\big), \tag{4}$$

where $\mathcal{L}_{\text{ce}}$ is the typical cross-entropy loss for VLMs, γ is the coefficient of visual loss, and all of the λ coefficients are the weighting factors for the losses of the corresponding visual tasks. During the inference process, the visual thinking tokens are not decoded.

Additionally, our framework supports the flexible integration of new visual token types. Because the pipeline follows a clean next-token prediction paradigm, additional tokens can be incorporated with minimal modification.

Training Data. To enable VLMs to progressively learn the visual tokens while not losing ability in the text space, CoVT introduces four data formatting stages as shown in Fig. 4. This guides the VLMs to learn progressively through the sequence from understanding visual tokens *(comprehension stage)*, to generating visual tokens *(generation stage)*, to reasoning with chain of visual thoughts *(reasoning stage)*, and finally to dynamically using visual tokens in the thinking chain *(efficient reasoning stage)*.

In *1) comprehension stage*, we insert visual tokens after `<image>` to teach the VLMs to learn the basic semantics of the visual tokens. In *2) generation stage*, we modify the question and answer, as shown in Fig. 4, to guide the VLMs to generate the visual tokens precisely. *3) reasoning stage* introduces the chain-of-visual-thought format, where the visual tokens are used within the thinking process. This teaches the model to leverage the visual tokens to derive the final answers. *4) efficient reasoning stage* randomly drops out some sets (ranging from 0 to k, where k is the number of token types) of visual tokens. With a portion of visual token types, CoVT learns to utilize all features effectively rather than being constrained by a fixed output pattern.

Seg	Depth	DINO	Edge	CVBench	Count	Depth	Dist.
Closed-source Models							
Claude-4-Sonnet				76.3	62.2	77.7	80.5
GPT-4o				79.2	65.6	86.7	81.0
Qwen2.5-VL-7B				74.5	65.0	72.8	75.5
LVR†				77.2	67.0	86.2	71.2
Monet†				75.7	**68.8**	77.7	73.8
CoVT (1 Visual Token)							
✓				77.9	66.0	80.8	80.5
	✓			78.7	65.4	83.2	78.2
		✓		71.3	64.7	72.3	66.7
CoVT (3 Visual Tokens)							
✓	✓	✓		**80.0**	66.2	86.8	**82.5**
Δ (vs Baseline)				**+5.5**	**+1.2**	**+14.0**	**+7.0**
CoVT (4 Visual Tokens)							
✓	✓	✓	✓	79.8	66.1	**89.2**	80.5

Table 2: CoVT delivers consistent improvements across CV-Bench and reveals that each visual token type contributes most effectively to the tasks related to it. † indicates our reproduced results.

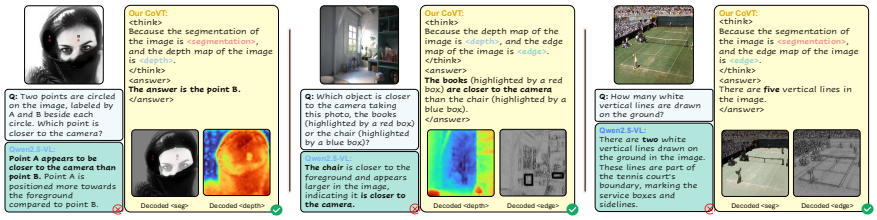

Fig. 5: Visualization of CoVT tokens. Different visual tokens contribute complementary cues that enable the model to solve complex perceptual reasoning tasks. **Left:** Segmentation tokens localize point B on the face, while the depth tokens capture the relative depth relationships. **Mid:** Depth visual tokens provide depth map information, and the edge tokens help highlight the positions of two boxes. **Right:** The Segmentation tokens identify the attended region, and the edge tokens delineate the fine-grained line structures.

The dataset used for training includes: (1) a vision-centric (and also real-world) subset of the LLaVA-OneVision dataset [29]. (2) spatial perception data, including TallyQA [1] and ADE20K-Depth [4, 69].

Seg	Depth	DINO	Edge	MME-RW	BLINK	RW-QA	MMT	MMStar-P	MMVP	V*	HR_{4K}	HR_{8K}
Closed-source Models												
Claude-4-Sonnet				-	39.6	63.7	-	58.8	48.7	15.2	32.3	22.7
GPT-4o				-	63.0	69.7	-	65.2	72.0	42.9	50.6	46.7
Qwen2.5-VL-7B				60.0	55.7	68.6	61.7	67.1	56.0	76.4	68.6	64.9
LVR†				-	55.9	69.7	-	67.7	50.7	81.7	69.6	65.6
Monet				-	52.9†	69.3†	-	67.1†	56.0†	**83.3**	71.0	68.0
CoVT (1 Visual Token)												
✓				62.1	**57.4**	71.1	62.1	68.5	**58.7**	79.1	71.9	69.0
	✓			62.0	56.4	71.5	**62.7**	**69.9**	**58.7**	79.1	71.9	69.4
		✓		61.1	55.8	71.5	62.5	67.9	57.3	77.5	71.0	68.6
CoVT (3 Visual Tokens)												
✓	✓	✓		**63.7**	56.0	71.6	62.1	69.2	**58.7**	78.0	**72.9**	69.4
Δ (vs Baseline)				**+3.7**	**+0.3**	**+3.0**	**+0.4**	**+2.1**	**+2.7**	**+1.6**	**+4.3**	**+4.5**
CoVT (4 Visual Tokens)												
✓	✓	✓	✓	63.3	56.2	**71.8**	61.9	68.4	56.7	78.5	72.5	**69.9**

Table 3: Comparison of CoVT with the baseline and closed-source models. CoVT delivers consistent improvements across all vision-centric benchmarks. † indicates our reproduced results based on the provided checkpoints.

4 Experiments

In the experiment section, we first describe the experimental settings of Chain-of-Visual-Thought (CoVT) in Sec. 4.1. Second, we introduce the benchmark results on both vision-centric and non-vision-centric datasets in Sec. 4.2. Third, we present the quantitative results demonstrating the advantages of CoVT in Sec. 4.3. Finally, we "visualize" the continuous visual tokens in Sec. 4.4 and conduct ablation studies in Sec. 4.5.

	CV-Bench			BLINK				
	Count	Depth	Dist.	Count	Obj. Loc.	Rel. Depth	Vis. Corr.	Vis. Sim.
LLaVA	59.3	61.8	50.2	56.7	54.9	52.4	29.7	51.1
Aurora† (*depth*)	54.9	67.7	**52.3**	53.3	55.7	62.9	26.2	47.4
CoVT (*w/ Depth*)	60.7	**71.0**	**52.3**	56.7	**59.8**	**75.8**	**31.4**	**53.3**
Δ (*vs Aurora*)	**+5.8**	**+3.3**	+0.0	**+3.4**	**+4.1**	**+12.9**	**+5.2**	**+5.9**
Aurora† (*count*)	56.0	62.2	47.8	31.7	26.2	24.2	26.7	21.5
CoVT (*w/ Seg*)	**61.9**	60.7	51.3	**58.3**	56.6	69.4	29.7	52.6
Δ (*vs Aurora*)	**+5.9**	**-1.5**	**+3.5**	**+26.6**	**+30.4**	**+45.2**	**+3.0**	**+31.1**

Table 4: Comparison between CoVT and Aurora based on LLaVA-v1.5-13B. † indicates our reproduced results based on the provided checkpoints.

4.1 Experiment Details

In our experiments, Qwen2.5-VL-7B [3] is selected as the main baseline, CoVT uses LoRA [23] tuning method, while the rank of LoRA is 16 and the LoRA alpha is 32. The learning rate of LoRA is set as 5×10^{-5}, and the projection layer learning rate is set to 1×10^{-5}. The first phase steps are 4000, the second and third phase steps are 3000, and the fourth phase steps are 5000. Batch size is set to 4. The experiments are carried out on 1×A100 or 4×A6000 GPUs. γ and all of the λ in Eq. 4 are set as 1.

4.2 Model Evaluation

All evaluations are performed using VLMEvalKit [14].

Vision-centric benchmarks. Our main focus is on CV-Bench. In particular, from CV-Bench, we highlight the sub-tasks *Count*, *Depth*, and *Distance*. These sub-tasks act as precise indicators to validate the effectiveness of our method. We further evaluate on other vision-centric benchmarks, including BLINK [18], RealWorldQA (RW-QA) [57], MMT-Bench (MMT) [64], MMStar [8], MMVP [47], MME-RealWorld (MME-RW) [67], V* Bench (V*) [56], and HRBench (HR_{4K} and HR_{8K}) [50]. Among them, for MMStar we specifically choose the *Coarse Perception*, *Fine-grained Perception*, and *Instance Reasoning* subsets (MMStar-P), as they are more aligned with real-world reasoning.

Non-vision-centric benchmarks. Besides, we also evaluate CoVT on some non-vision-centric visual benchmarks such as OCRBench [32], MME [16], MUIR-Bench [48], HallusionBench [20], A-OKVQA [40], TaskMeAnything [66], We-MATH [38], and WorldMedQA-V [36]. For MME, we select the text-centric sub-task *text translation* as the evaluation of the text-centric performance.

4.3 Quantitative Results

CoVT outperforms the baseline across the vision-centric benchmarks. As shown in Tab. 2 and Tab. 3, CoVT is capable of incorporating various kinds of visual tokens. We employ three visual tokens, Segmentation, Depth, and DINO, as our main results. Compared to the baseline, CoVT consistently achieve large gains across the main vision-centric benchmarks. CoVT improves by 5.5% on

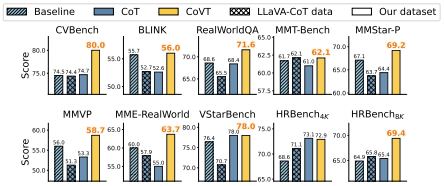

Fig. 6: Text-only CoT *vs* CoVT. Surprisingly, text-only CoT hurts performance on vision benchmarks; in contrast, CoVT substantially enhances VLMs' visual perception and understanding.

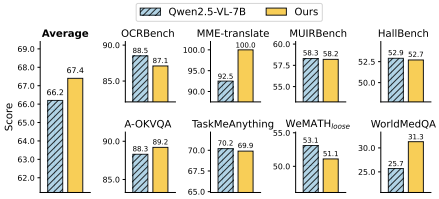

Fig. 7: Beyond the gains on vision-centric benchmarks, CoVT also achieves slight improvements on **non–vision-centric tasks**.

CV-Bench, 14.0% on the subtask *depth* in CV-Bench, 3.7% on MME-RealWorld, and 4.5% on HRBench8K. These results indicate CoVT with visual thinking chain improves across visual-centric and fine-grained perceptual tasks.

CoVT generalizes to the other baseline. In addition to the experiment based on Qwen, CoVT is also implemented based on LLaVA-v1.5-13B, in order to compare CoVT with Aurora. As shown in Fig. 4, for CoVT with depth tokens, it excels Aurora-*depth* by 12.9% on *relative-depth* in BLINK. For CoVT using segmentation CoVT tokens, outperforms Aurora-*count* by 26.6% on BLINK-*count* benchmark. These results indicate that CoVT generalizes on various baselines across vision-centric tasks.

4.4 Qualitative Results

To better understand why CoVT is effective, we select several examples and decode the CoVT tokens from the model outputs to visualize whether these tokens provide useful information for reasoning.

Fig. 5 illustrates that different CoVT tokens carry different rich fine-grained information, and the cues they provide are highly complementary. To be interpretable, CoVT tokens are decoded into fine-grained output (*e.g.* masks, depth maps, edge maps). ***For the left example***, the Segmentation token provides 2D perceptual cues by localizing "point B" on the face, while the Depth token supplies 3D information, indicating that the face region is closer to the camera than the surrounding areas. ***For the middle example***, the Depth token encodes depth perception, whereas the Edge token supplies fine-grained boundary cues for the two target objects. This example is from *Depth* sub-task in CV-Bench, thus the synergy explains the 2.4% improvement observed when CoVT uses four visual tokens instead of three on the CVBench-Depth task, as shown in Tab. 2. ***For the right example***, the Segmentation token localizes the target region and the Edge token emphasizes fine-grained boundaries, which is difficult for the Segmentation token, deriving the correct answer through chains of visual thoughts.

Quantity	CVBench	BLINK	RW-QA	MM*-P	MMVP	V*	HR_{4K}
0 token	76.6	55.5	70.7	68.0	55.3	77.5	68.6
16 empty	75.7	**56.0**	70.3	67.9	56.7	77.5	68.1
1 token	78.9	55.6	70.8	68.8	56.7	**78.5**	**73.0**
8 tokens	**80.0**	**56.0**	**71.6**	**69.2**	**58.7**	78.0	72.9
32 tokens	73.9	54.4	68.4	62.1	55.3	77.2	70.8

Table 5: CoVT Ablation on segmentation token numbers. The appropriate token number is essential for CoVT. **0 token** is the control group (direct fine-tuning), while **16 empty tokens** serve to isolate the role of the token embeddings themselves. 8 segmentation tokens perform the best.

Type	Align	CVBench	BLINK	RW-QA	MM*-P	MMVP	V*	HR_{4K}
Seg	Feature	76.8	55.2	70.6	67.7	56.0	78.0	69.8
	Ours	**77.9**	**57.4**	**71.1**	**68.5**	**58.7**	**79.1**	**71.9**
Depth	Feature	77.0	54.2	70.5	67.6	55.3	**78.0**	71.3
	Ours	**78.7**	**56.4**	**71.5**	**69.9**	**48.7**	77.5	**71.9**

Table 6: Our **alignment strategy** plays a crucial role in further improving CoVT.

4.5 Ablation Studies

Text-only Chain-of-Thought *vs*. Chain-of-Visual-Thought. To isolate the contribution of continuous visual tokens, we conduct an ablation comparing text-only CoT with our full CoVT framework. For the text-only setting, we follow the CoT formatting paradigm used in the LLaVA-CoT 100k dataset and apply the same formatting to our own training data, ensuring full consistency with CoVT except for the absence of visual tokens. Fig. 6 shows that text-only CoT not only fails to improve performance on vision-centric reasoning tasks, but often degrades it. In contrast, CoVT consistently enhances performance across vision-centric benchmarks, highlighting the necessity of continuous visual tokens for effective visual reasoning.

Token Numbers. We ablate various numbers of segmentation visual tokens, as shown in Tab. 5. The "0 token" setting corresponds to directly fine-tuning the base model on our dataset. The "16 empty" setting replaces our 16 visual thinking tokens with 16 ordinary tokens without any visual alignment, serving as a pure latent-reasoning baseline. Settings with 1, 8, and 32 Segmentation tokens vary the token count while keeping the Depth and DINO tokens fixed at 4 each; the 8-token setting corresponds to our full model.

We observe that using too few Segmentation tokens leads to performance degradation, though still better than the "0 token" baseline. However, increasing the token count to 32 harms performance, maybe due to the difficulty of aligning a large number of segmentation tokens. The poor performance of the "16 empty" variant further highlights the importance of visually aligned tokens. Overall, the results demonstrate that 8 Segmentation tokens, together with 4 Depth and 4 DINO tokens, form a balanced and effective configuration, and that visual alignment is essential for enhancing vision-centric perception in VLMs.

Segmentation and Depth Alignment Strategies. We ablate two alignment approaches. Our primary method aligns CoVT tokens through task decoders, enabling the tokens to capture richer and more fine-grained perceptual cues. In contrast, direct feature alignment applies an MSE loss between the visual tokens

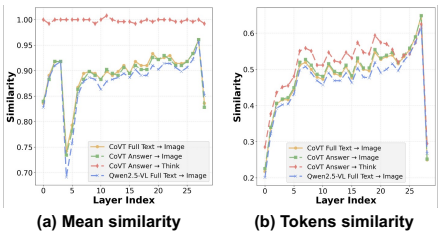

Fig. 8: CoVT exhibits higher similarity between text tokens and image features.

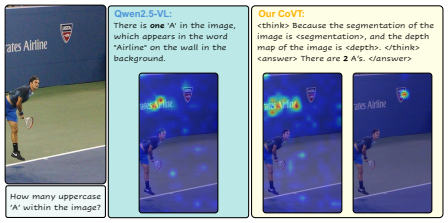

Fig. 9: In CoVT, latent tokens across layers consistently attend to relevant regions correctly.

and the encoder features of the corresponding visual model, which inevitably loses important perceptual details from the image.

As shown in Tab. 6, direct feature alignment consistently underperforms CoVT. These results highlight the importance of our tailored alignment strategies and demonstrate that aligning visual tokens with decoders yields more effective and perceptually grounded representations.

CoVT remains competitive on non-vision-centric benchmarks. Fig. 7 shows that CoVT achieves comparable performance, with a 1.2% gain across eight non-vision-centric benchmarks, indicating no notable degradation in generalization and a slight overall improvement.

4.6 CoVT Better Aligns Reasoning with Visual Evidence

We analyze CoVT through (i) representation similarity between language tokens and visual features, and (ii) cross-attention patterns over the image. In Fig. 8, we compute the layer-wise cosine similarity between image features and different groups of output tokens from CoVT, and compare against the baseline. Concretely, we measure two kinds of similarity: (a) *mean similarity*, the mean-pooled token similarity, and (b) *tokens similarity*, the average of the tokens similarity, between (i) full output text and image features, (ii) answer tokens and image features, and (iii) answer tokens and intermediate "thinking" tokens.[4]

Across intermediate layers, CoVT shows consistently higher similarity between answer tokens and image features than the baseline, indicating stronger visual grounding throughout the forward pass. Moreover, CoVT also has high similarity between answer tokens and thinking tokens, suggesting that its intermediate reasoning states are more predictive of the final answer and better aligned with the visual input.

We further inspect cross-attention from answer tokens to image tokens (Fig. 9). Across layers, the baseline attends to only part of the relevant regions, whereas CoVT more precisely and consistently covers the task-relevant regions, aligning with the baseline's failure and CoVT's correct prediction.

5 Conclusions

In this paper, we introduced **CoVT**, the chain of continuous visual thoughts that enables vision–language models to reason beyond discrete linguistic space by

[4] All similarities are computed on the hidden states of the corresponding layer.

leveraging compact, dense visual representations. CoVT consistently improves visual-centric reasoning across diverse perception benchmarks and reveals that different types of visual tokens contribute to complementary aspects of multimodal understanding. These findings suggest that CoVT provides a general framework for integrating perceptual reasoning into multimodal systems.

6 Acknowledgement

We sincerely thank Himanshu Dubey and Sowmay Jain for API credit support from Anannas AI. We greatly thank Baifeng Shi, Ji Xie, Shaofeng Yin for their insightful discussions and valuable feedback on our paper. We especially thank Mahtab Bigverdi for her assistance in reproducing the baseline results.

References

1. Acharya, M., Kafle, K., Kanan, C.: Tallyqa: Answering complex counting questions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8076–8084 (2019)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bigverdi, M., Luo, Z., Hsieh, C.Y., Shen, E., Chen, D., Shapiro, L.G., Krishna, R.: Perception tokens enhance visual reasoning in multimodal language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3836–3845 (2025)
5. Biran, E., Gottesman, D., Yang, S., Geva, M., Globerson, A.: Hopping too late: Exploring the limitations of large language models on multi-hop queries. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 14113–14130 (2024)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chen, H., Feng, Y., Liu, Z., Yao, W., Prabhakar, A., Heinecke, S., Ho, R., Mui, P., Savarese, S., Xiong, C., et al.: Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. arXiv preprint arXiv:2411.04282 (2024)
8. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
9. Cheng, J., Van Durme, B.: Compressed chain of thought: Efficient reasoning through dense representations. arXiv preprint arXiv:2412.13171 (2024)

10. Cheng, Z., Chen, Q., Xu, X., Wang, J., Wang, W., Fei, H., Wang, Y., Wang, A.J., Chen, Z., Che, W., Qin, L.: Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. ArXiv **abs/2505.15510** (2025)
11. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), arXiv preprint arXiv:2501.12948
12. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
13. Deng, Y., Prasad, K., Fernandez, R., Smolensky, P., Chaudhary, V., Shieber, S.: Implicit chain of thought reasoning via knowledge distillation. arXiv preprint arXiv:2311.01460 (2023)
14. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., wen Dong, X., Zang, Y., Zhang, P., Wang, J., Lin, D., Chen, K.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. Proceedings of the 32nd ACM International Conference on Multimedia (2024)
15. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
16. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. ArXiv **abs/2306.13394** (2023)
17. Fu, S., Bonnen, T., Guillory, D., Darrell, T.: Hidden in plain sight: Vlms overlook their visual representations (2025), arXiv preprint arXiv:2506.08008
18. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)

19. Goyal, S., Ji, Z., Rawat, A.S., Menon, A.K., Kumar, S., Nagarajan, V.: Think before you speak: Training language models with pause tokens. arXiv preprint arXiv:2310.02226 (2023)
20. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14375–14385 (2023)
21. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14953–14962 (2023)
22. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024)
23. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021)
24. Huang, Z., Yu, N., Chen, G., Qiu, H., Debevec, P., Liu, Z.: Vchain: Chain-of-visual-thought for reasoning in video generation. arXiv preprint arXiv:2510.05094 (2025)
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
26. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
27. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9579–9589 (2024)
28. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning. arXiv preprint arXiv:2509.24251 (2025)
29. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
30. Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought (2025), arXiv preprint arXiv:2501.07542
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
32. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X., lin Liu, C., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Sci. China Inf. Sci. **67** (2023)
33. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering (2022), arXiv preprint arXiv:2209.09513
34. Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.W., Wu, Y.N., Zhu, S.C., Gao, J.: Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems **36**, 43447–43478 (2023)
35. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann,

K., Welleck, S., Yazdanbakhsh, A., Clark, P.: Self-refine: Iterative refinement with self-feedback (2023), arXiv preprint arXiv:2303.17651
36. Matos, J., Chen, S., Placino, S., Li, Y., Pardo, J.C.C., Idan, D., Tohyama, T., Restrepo, D., Nakayama, L.F., Pascual-Leone, J.M.M., Savova, G.K., Aerts, H., Celi, L.A., Wong, A.K.I., Bitterman, D.S., Gallifant, J.: Worldmedqa-v: a multilingual, multimodal medical examination dataset for multimodal language models evaluation. ArXiv **abs/2410.12722** (2024)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
38. Qiao, R., Tan, Q., Dong, G., Wu, M., Sun, C., Song, X., Gongque, Z., Lei, S., Wei, Z., Zhang, M., Qiao, R., Zhang, Y., Zong, X., Xu, Y., Diao, M., Bao, Z., Li, C., Zhang, H.: We-math: Does your large multimodal model achieve human-like mathematical reasoning? ArXiv **abs/2407.01284** (2024)
39. Rahmanzadehgervi, P., Bolton, L., Taesiri, M.R., Nguyen, A.T.: Vision language models are blind: Failing to translate detailed visual features into words (2025), arXiv preprint arXiv:2407.06581
40. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision (2022)
41. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems 37 (2024)
42. Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., Liu, L.: Pixel difference networks for efficient edge detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5117–5127 (2021)
43. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11888–11898 (2023)
44. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
45. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
46. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
47. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
48. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., Yan, T., Mo, W.J., Liu, H.H., Lu, P., Li, C., Xiao, C., Chang, K.W., Roth, D., Zhang, S., Poon, H., Chen, M.: Muirbench: A comprehensive benchmark for robust multi-image understanding. ArXiv **abs/2406.09411** (2024)
49. Wang, Q., Shi, Y., Wang, Y., Zhang, Y., Wan, P., Gai, K., Ying, X., Wang, Y.: Monet: Reasoning in latent visual space beyond images and language. arXiv preprint arXiv:2511.21395 (2025)

50. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7907–7915 (2025)
51. Wang, X., Zhou, X., Fathi, A., Darrell, T., Schmid, C.: Visual lexicon: Rich image features in language space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19736–19747 (2025)
52. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models (2023), arXiv preprint arXiv:2203.11171
53. Wang, X., Zhou, D.: Chain-of-thought reasoning without prompting (2024), arXiv preprint arXiv:2402.10200
54. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
55. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
56. Wu, P., Xie, S.: V?: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024)
57. XAI: Grok-1.5 vision preview (2024), https://x.ai/news/grok-1.5v, accessed 27 June 2026
58. Xie, J., Darrell, T., Zettlemoyer, L., Wang, X.: Reconstruction alignment improves unified multimodal models. arXiv preprint arXiv:2509.07295 (2025)
59. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L.C., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S.Q., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.C., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. ArXiv **abs/2505.09388** (2025)
60. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
61. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
62. Yang, Z., Yu, X., Chen, D., Shen, M., Gan, C.: Machine mental imagery: Empower multimodal reasoning with latent visual tokens (2025), arXiv preprint arXiv:2506.17218
63. Yin, S., Lei, T., Liu, Y.: Toolvqa: A dataset for multi-step reasoning vqa with external tools. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4424–4433 (2025)
64. Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., et al.: Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. arXiv preprint arXiv:2404.16006 (2024)

65. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
66. Zhang, J., Huang, W., Ma, Z., Michel, O., He, D., Gupta, T., Ma, W.C., Farhadi, A., Kembhavi, A., Krishna, R.: Task me anything. ArXiv **abs/2406.11775** (2024)
67. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? arXiv preprint arXiv:2408.13257 (2024)
68. Zhang, Z., Zhang, A., Li, M., Karypis, G., Smola, A., et al.: Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research (2024)
69. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302 – 321 (2016)
70. Zhou, C., Wang, M., Ma, Y., Wu, C., Chen, W., Qian, Z., Liu, X., Zhang, Y., Wang, J., Xu, H., Luo, F., Chen, X., Hao, X., Li, H., Zhang, A., Wang, W., Li, L., Lu, Z., Lu, Y., wang Guo, Y.: From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. ArXiv **abs/2509.25373** (2025)
71. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)