Mull-Tokens: Modality-Agnostic Latent Thinking

Arijit Ray[1,4] Ahmed Abdelkader[1] Chengzhi Mao[1] Bryan A. Plummer[4] Kate Saenko[4]
Ranjay Krishna[2] Leonidas Guibas[1,3] Wen-Sheng Chu[1]

[1]*Google* [2]*University of Washington* [3]*Stanford University* [4]*Boston University*

https://arijitray.com/multimodal_thinking/

Abstract

Reasoning goes beyond language; the real world requires reasoning about space, time, affordances, and much more that words alone cannot convey. Existing multimodal models exploring the potential of reasoning with images are brittle and do not scale. They rely on calling specialist tools, costly generation of images, or handcrafted datasets. We offer a simpler alternative – Mull-Tokens – latent tokens that can be pre-trained to hold intermediate information in either image or text modalities so as to think towards the correct answer. We investigate best practices to train Mull-Tokens inspired by latent reasoning frameworks. We first train Mull-Tokens using supervision from interleaved text-image traces, and then fine-tune without any supervision by only using the final answers. Across four challenging spatial reasoning benchmarks involving tasks such as solving puzzles and taking different perspectives, we demonstrate that Mull-Tokens improve upon several baselines utilizing text-only reasoning or interleaved image-text reasoning, achieving a $+3\%$ average improvement and up to $+16\%$ on a puzzle solving reasoning-heavy split compared to our strongest baseline. Adding to conversations around challenges in grounding textual and visual reasoning, Mull-Tokens offers a simple solution to implicitly think in multiple modalities.

Figure 1. Compared to existing approaches for reasoning in text or interleaving image and text, we offer a simpler alternative - modality-agnostic/multimodal thinking in the vision-language space using Mull-Tokens.

1. Introduction

Real-world visual tasks [54, 55] require a synchrony of perception and reasoning [25]. For example, when solving visual puzzles and IQ tests, people imagine patterns and manipulate visual maps to solve the problem [40]. Such reasoning—often bundled up as visual/spatial reasoning [45]— is difficult for humans and machines alike [62]. It requires reasoning in space [29], in 3D [4, 33], and often in time [5, 62]. While textual Chain-of-Thought (CoT) [57] has advanced verbal logical reasoning for language models, it falls short on visual reasoning: models still falter when problems demand more than verbal image descriptions [24, 26]. Mastering *vision-centric* language reasoning remains an open challenge.

To advance visual reasoning, researchers have explored "think-and-sketch" strategies [25, 42] using interleaved visual thoughts, but with conflicting results [33]. Tool-augmented designs rely on external visual modules [39] such as cropping tools [42] or specialized sketching models [25, 73] - which are incapable of complex visual manipulations and are often brittle [21]. Unified models [7, 12, 50] offer a more integrated alternative by generating intermediate images but are expensive to train. The latest visual reasoning models instead use image thoughts, either as explicit visual tokens [4] or dense continuous embeddings [65]. However, they require bespoke task-specific training data and hence, do not offer a general recipe for visual reasoning. For instance, we find that naively adding *modality-specific* visual thoughts can sometimes hurt: supervising a model to interleave textual thoughts and visual latents actually reduced performance on a visual puzzle task compared to reasoning in text. Hence, an effective strategy to use the growing availability of multimodal reasoning traces [6, 30] remains elusive.

To address this gap, we present Mull-Tokens:

modality-agnostic latent tokens that function as a multi-modal scratchpad for the model's internal reasoning, capable of representing *both visual or textual information* as needed. During inference, we append these special tokens to the input question, prompting the model to utilize those modality-agnostic slots for arbitrary intermediate computations (spatial mappings, depth predictions, verbal manipulations, etc.) towards the answer, without the need for explicit decoding into text or images.

Inspired by recent latent reasoning frameworks [20, 46, 65, 66], we develop a two-stage curriculum for training a model to use *Mull*-Tokens effectively. The first stage trains the model using multimodal reasoning traces where each *Mull*-Token is *anchored* to a relevant textual or visual concept [30]. For example, one token might learn to encode a visual object layout or a textual symbolic mapping if those are useful for the task at hand. The second stage relaxes the supervision on *Mull*-Tokens and fine-tunes the model using only the final answer. This stage gives the model an opportunity for optimizing its latent trajectories freely, so as to maximize end-task performance. *Mull*-Tokens augment a model's vocabulary, similar to special tokens (e.g. <plan> or <pause>) used in reasoning models [4, 20]. Finally, we follow by a third refinement stage leveraging the latest reinforcement learning (RL) techniques [22] to reward good latent chains of "thinking."

We fine-tune a Multimodal Language Model (MLM), Qwen2.5-VL (7B) [2], with *Mull*-Tokens as described above. We then evaluate the resulting model on a wide range of visual reasoning benchmarks covering both images and videos. In particular, we test on spatial reasoning problems such as visual puzzles and IQ-tests from BLINK [17], VSI-Bench [62] (video spatial reasoning), SAT [45] (action and motion-based spatial relationship reasoning), and MMSI-Bench [63] (multi-image reasoning), offering a thorough evaluation of *Mull*-Tokens' capabilities.

Across all benchmarks, our *Mull*-Tokens significantly improves the model's reasoning accuracy compared to standard answer-only fine-tuning, textual reasoning, and a recent approach that interleaves explicit text reasoning with image latents [65]. On average, *Mull*-Tokens achieves a +3% absolute accuracy gain over our strongest baseline, which is surprisingly obtained by fine-tuning the base model directly on answers without reasoning. The benefits of *Mull*-Tokens are even larger on the most reasoning-intensive tasks, improving +16% on a visual puzzle split.

These results echo the gains reported by prior works when introducing learned visual tokens [4], but we emphasize that our approach requires far fewer and modality-agnostic tokens. Only $10-40$ *Mull*-Tokens are sufficient to yield the above improvements. This is a drastic reduction in inference overhead compared to generating hundreds of word tokens in a typical text CoT rationale, or hundreds of visual tokens to represent an image, reminiscent of findings that shorter reasoning traces can be just as effective [49].

In summary, our main contribution is an effective reasoning paradigm utilizing causal multimodal *Mull*-Tokens. Our approach improves upon the latest existing approach by 9% with up to 16% on reasoning-heavy splits. *Mull*-Tokens is also faster than producing verbose thoughts [16] or explicit images as thoughts [4] (20 <*Mull*> vs several hundreds), establishing a Pareto-dominance of our modality-agnostic approach. Adjacent to conversations around causal world-model representations for reasoning, *Mull*-Tokens offers a simple approach to reason using visual and textual representations in a unified manner.

2. Related Work

The primary theme of our work concerns the incorporation of multi-modal latents to enhance the visual and spatial reasoning capabilities [9, 45, 62] of MLMs [2, 71].

Visual Reasoning. Building upon powerful perception in vision-language models [1, 36], there has been great interest in improving reasoning over both modalities [27, 72]. Inspired by the success of textual CoT reasoning in language models [15, 56], recent works have looked into using similar techniques for visual questions [16, 38]. Due to limitations of textual reasoning for visual tasks [21, 38], many works now reason using visual thoughts as well-but rely on expensive unified models [37, 48, 50, 59–61, 68] to explicitly generate images [13, 21, 30, 70], or latents [4, 10, 31, 65, 65, 67, 69], or external tools that may be brittle [25, 42]. We aim to incorporate latents as simpler direct and less expensive alternative without needing bespoke data [65]. While fixing problems with textual reasoning using text alone is a valid avenue for exploration [16, 34, 38], we offer a simpler and faster alternative of appending a few modality agnostic thinking tokens. Within visual reasoning, there is recently a special emphasis on spatial reasoning [29, 35, 52, 53, 62, 71] with an increased variety of spatial annotations [11, 45], benchmarks [17, 51, 52, 62], and data curation [16, 44, 53, 64]. Our work contributes to training paradigms for improving spatial reasoning [4, 5, 11, 14, 41, 45, 64].

Latent Reasoning. Our thinking tokens are heavily inspired by latent reasoning techniques in language models [20, 23, 46]. A closely related theme is test-time scaling, allocating more compute at inference time before producing an answer [20, 28, 47]. While some approaches adopt recurrent models over continuous tokens [18, 23, 74], they break standard parallel processing and accumulate errors [23, 65]. Concurrent works MVoT [32] and MIRAGE [65] also explore visual thought tokens interleaved with text, similar to the interleaved image-text baseline we compare against. In contrast, our *Mull*-Tokens are modality-agnostic and do not require explicit switching between modalities. We adopt discrete think tokens with a rich internal recurrence achieving substantial gains.

3. Approach

Our primary objective is to improve the visual reasoning performance of multimodal language models [1, 36]. Formally, the model with parameters θ implements a conditional distribution $p_\theta(y|x^{img}, x^{txt})$, where x^{img} and x^{txt} denote the visual and textual inputs, and y is the target answer sequence.

Textual Chain-of-Thought (CoT) augments this input-output mapping with intermediate textual tokens $c_{1:T}^{\text{text}}$ effective for verbal reasoning, but often detrimental for visual tasks [62] due to problems such as drifting from visual inputs [38]. Recent approaches therefore supervise models to "think with images," *e.g.*, by generating intermediate images or visual sketches $c_{1:T}^{img}$ during reasoning [4, 21, 25]. However, these methods typically require (i) bespoke multimodal reasoning datasets $\mathcal{D}_{CoT}$ with aligned text–image chains, or (ii) expensive generation of subgoal images that are hard to generalize beyond the specific training domains.

We instead introduce a simpler, modality-agnostic mechanism, *Mull*-`Tokens`, implemented as a fixed-length sequence of special latent tokens $z_{1:K} = (\text{<Mull>}_1, \dots, \text{<Mull>}_K)$, which can carry both image-conditioned and text-conditioned information. During training and inference, the model can use these tokens purely as *internal compute* to improve visual reasoning, without being constrained to produce textual or image outputs at the intermediate steps.

At a high level, we replace explicit CoT supervision on intermediate text/images with supervision on the *function* of the latent chain: the model is trained so that using $z_{1:K}$ improves $p_\theta(y \mid x^{\text{img}}, x^{\text{txt}})$, while the internal semantics of $z_{1:K}$ are free to adapt. Below, we formalize our two-stage latent reasoning framework and a third RL refinement stage.

Modality-agnostic *Mull*-`Tokens`

Inspired by latent reasoning approaches in NLP [20], our training is decomposed into three stages (Figure 2): (1) *warm-up* the *Mull*-`Tokens` tokens to align with explicit multimodal reasoning steps; (2) *relax* supervision to only the final answer, allowing the model to optimize the latent chain internally; and (3) *refine* the causal usefulness of the latents via GRPO. Our base backbone is Qwen2.5-VL [2], which provides a text decoder and a frozen image encoder.

Stage 1: Warm-up the *Mull*-`Tokens` tokens

Let $\mathcal{D}_{\text{CoT}}$ be a dataset (see Figure 2 (left)) of multimodal CoT trajectories. Each training example consists of an input $(x^{\text{img}}, x^{\text{txt}})$, a reasoning trace

$$c_{1:T} = (c_1, \dots, c_T), \quad c_t \in \mathcal{V}^{txt} \cup \mathcal{V}^{img},$$

and a final answer sequence $y_{1:L}$. Here, $\mathcal{V}^{txt}$ is the text vocabulary and $\mathcal{V}^{\text{img}}$ denotes "pseudo-tokens" corresponding to subgoal images (*e.g.*, the jigsaw configuration after inserting a candidate piece).

We construct an interleaved training sequence

$$s = (q_{1:M}, z_1, \tilde{c}_1, z_2, \tilde{c}_2, \dots, z_T, \tilde{c}_T, y_{1:L}),$$

where $q_{1:M}$ encodes the question and context, $\tilde{c}_t$ is the textual substep or an image placeholder for c_t, and $z_t =$ <Mull>_t is the t-th latent token. The number of <*Mull*> here matches the length of the reasoning sequence in the dataset. The transformer is trained in an auto-regressive manner over s.

Let $h_t^{\text{Mull}} \in \mathbb{R}^d$ denote the hidden state at the position of z_t produced by the transformer. We apply *two* kinds of supervision:

- If $c_t \in \mathcal{V}^{txt}$ (next step is a word token), we warp h_t^{Mull} through the language model head to predict $p_\theta(c_t \mid s_{<t})$ and minimize the cross-entropy $\mathcal{L}_t^{text} = -\log p_\theta(c_t \mid s_{<t})$.
- If $c_t \in \mathcal{V}^{img}$ is (a subgoal image $I_t \in \mathbb{R}^{H \times W \times 3}$), we encode it using the frozen Qwen2.5-VL image encoder g_ϕ: $v_t = g_\phi(I_t) \in \mathbb{R}^d$. We then supervise h_t^{Mull} via cosine similarity: $\mathcal{L}_t^{\text{img}} = 1 - \cos(\hat{h}_t, \hat{v}_t)$

The full Stage-1 objective for a sequence s is

$$\mathcal{L}_{\text{stage1}} = \sum_{i \in \mathcal{I}_{\text{AR}}} \mathcal{L}_i^{\text{AR}} + \lambda_{\text{text}} \sum_{t \in \mathcal{T}_{\text{text}}} \mathcal{L}_t^{\text{text}} + \lambda_{\text{img}} \sum_{t \in \mathcal{T}_{\text{img}}} \mathcal{L}_t^{\text{img}},$$

where $\mathcal{I}_{AR}$ indexes standard tokens (question and answer) with conventional next-token cross-entropy $\mathcal{L}_i^{AR}$, and $\mathcal{T}^{text}$ / $\mathcal{T}^{img}$ index <*Mull*> positions followed by text or image steps, respectively.

This setup naturally extends to other modalities (e.g., 3D or trajectory representations) by changing the embedding function g_ϕ and similarity loss; we leave this extension to future work in the absence of such multimodal CoT data.

Stage 2: *Mull*-`Tokens` Relaxed training

After warm-up, we transition to a relaxed training stage where the latent tokens are no longer explicitly supervised on the next reasoning step. Let $s' = (q_{1:M}, z_{1:K}, y_{1:L})$ denote the sequence where $z_{1:K}$ is inserted after the question, but the intermediate CoT steps are no longer present.

We now optimize only the answer likelihood and drop all losses on $z_{1:K}$.

$$\mathcal{L}_{\text{stage2}} = -\sum_{\ell=1}^{L} \log p_\theta(y_\ell \mid s'_{<\ell}),$$

This relaxation is critical: the model is not forced to mimic the potentially suboptimal supervised CoT scaffolding from Stage 1. Instead, $z_{1:K}$ becomes an internal scratchpad

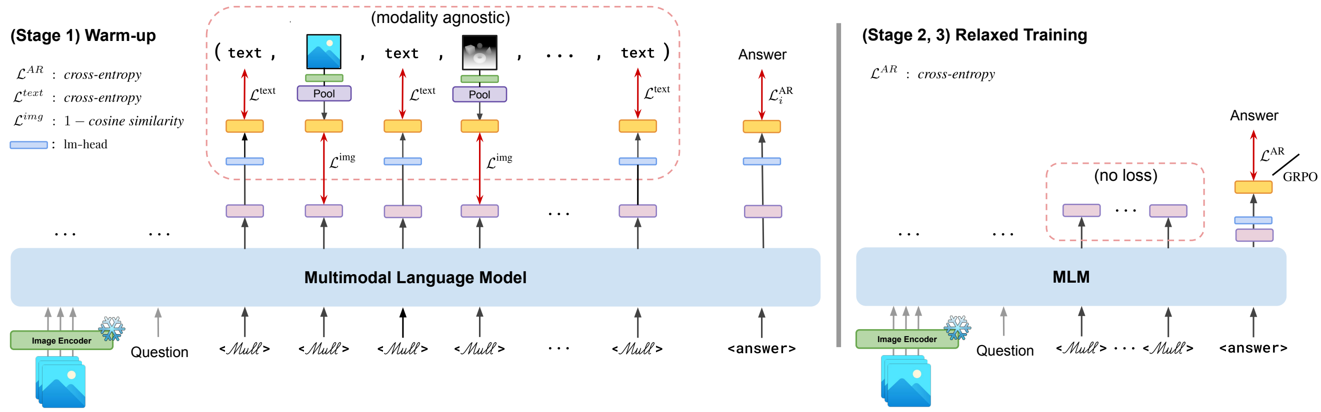

Figure 2. Our Mull-Tokens training involves two stages inspired by approaches in latent reasoning. We first pre-train/warm-up our Mull-Tokens to hold both image and text modalities depending on the context image/video and query. Next, the model free-form optimizes these Mull-Tokens to achieve the final correct answer. We see that pre-training the Mull-Tokens with to hold both image and text reasoning traces is key, as opposed to using them simply as extra compute without such pretraining or using text alone.

whose semantics are determined solely by how much they help minimize $\mathcal{L}_{\text{stage2}}$ on the final answer. Note that while K matches the reasoning trace length in Stage 1 (one <Mull> per CoT token), at Stage 2 we fix K to a small constant (*e.g.*, 20), effectively compressing the full reasoning trajectory into a compact latent representation. This compression may also mitigate the reasoning drift observed with verbose text CoT [38], since the model cannot linger on potentially misleading intermediate verbalizations.

A key design choice is the form of recurrence used for multimodal latents. Let $z_t^{\text{cont}} \in \mathbb{R}^d$ denote continuous recurrent latents as in [23, 65]. These approaches explicitly iterate over t: $z_{t+1}^{\text{cont}} = f_\theta(z_t^{\text{cont}}, x)$. requiring sequential updates both at training and inference time, and leading to: (i) *inefficiency*, since they break standard parallel token processing in transformers; and (ii) *instability*, since errors accumulate over long chains, as empirically noted in [23, 65].

Therefore, we adopt discrete token latents as in [4, 20]. We allocate a fixed number K of <Mull> tokens in the sequence. Let $H^{\text{Mull}} = (h_1^{\text{Mull}}, \dots, h_K^{\text{Mull}})$ denote their hidden states after the transformer layers; these states are jointly computed via standard self-attention, and then attended to by the answer tokens $y_{1:L}$. We ablate K as a hyperparameter in our experiments. This design is compatible with transformer parallelism, while still allowing rich internal recurrence through self-attention over $z_{1:K}$.

Stage 3: RL Refinement with GRPO

Although Stage 2 encourages the model to *use Mull-*Tokens to reduce answer loss, it does not enforce that the latent chain is *causally* responsible for the final prediction. In principle, the model could learn a shortcut mapping from $(x^{\text{img}}, x^{\text{txt}})$ directly to y and ignore $z_{1:K}$, e.g., by placing most of the computation in the final layers conditioned only weakly on the <Mull> states.

To explicitly reward latent chains that causally contribute to the correct answer, we introduce a third stage based using GRPO [22], following recent text-based reasoning frameworks [8, 16].

Let $\pi_\theta(y_{1:L}, z_{1:K} \mid x)$ denote the policy induced by the MLLM, where $x = (x^{\text{img}}, x^{\text{txt}})$, and $z_{1:K}$ are sampled internal latents (or deterministically generated but treated as part of the trajectory). For each training instance with ground truth answer y^*, we define a reward

$$r(y_{1:L}, y^*) = \begin{cases} 1, & \text{if } y_{1:L} \text{discrete answers} \\ \text{score}(y_{1:L}, y^*) & \text{if numeric / continuous,} \end{cases}$$

where score is a graded similarity, e.g., based on normalized absolute error; the exact form detailed in the supplementary.

In practice, we follow the standard GRPO update that effectively shapes the gradients to emphasize reward-improving trajectories. Because our Mull-Tokens are a fixed discrete prefix of length K, and the first answer token y_1 is sampled from the final Mull state (*i.e.*, its hidden representation h_K^{Mull} under self-attention), the gradient signal for r flows back primarily through h_K^{Mull} and, via self-attention, through the entire latent chain $h_{1:K}^{\text{Mull}}$. Note that while the policy gradient loss is applied to the sampled answer tokens, the latent tokens are optimized *indirectly*: the answer token logits depend on attending to $h_{1:K}^{\text{Mull}}$, so gradient updates reshape the latent representations.

Thus, GRPO encourages latent trajectories $(h_1^{\text{Mull}}, \dots, h_K^{\text{Mull}})$ that *causally* lead to correct answers, rather than merely co-occurring with them. Empirically, we find that this third stage further improves performance compared to SFT-only training.

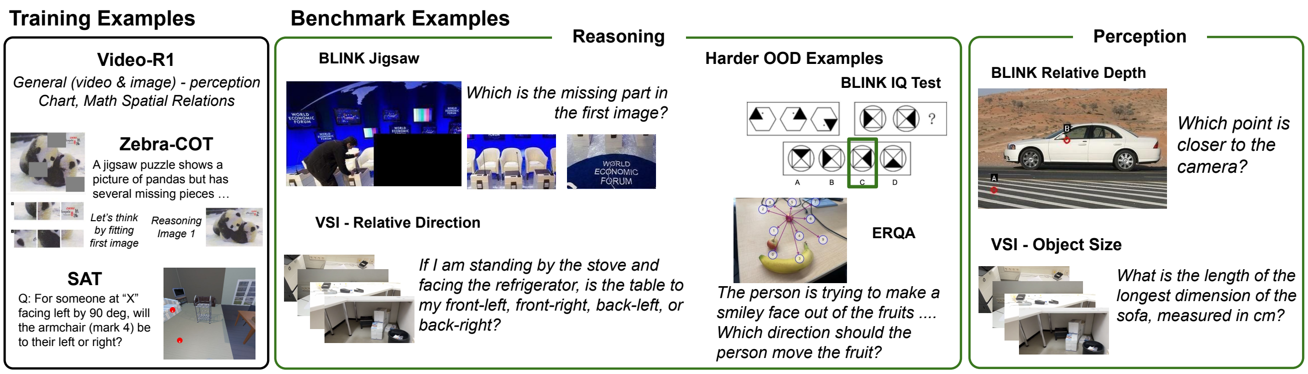

Figure 3. Examples of training and our benchmark data. We test on diverse reasoning benchmarks, which includes hard examples that do not have the reasoning outlined in the training sets. Our training consists of two types of data- general video QA's with text reasoning, data of interleaved image text reasoning, and data with only the final answer.

4. Experiments

We evaluate and ablate *Mull*-Tokens on four recent challenging visual reasoning benchmarks to assess: i) improvements compared to existing visual reasoning paradigms of reasoning in text or interleaving visual thoughts, ii) how important multimodal information in *Mull*-Tokens is for performance gains - showing *multimodal* thinking compute is crucial for performance compared to text-only or extra compute without such anchoring, and iii) scaling behavior with latent tokens and recurrence design choices when using *Mull*-Tokens - showing discrete tokens [20] deliver better performance and speed continuous embeddings used in recent work [23, 65].

4.1. Experimental setup

Base model. We choose a powerful open-source multimodal language model, Qwen2.5-VL (7B) model [2] as our base since it is also used in existing visual reasoning works [16, 38, 65]. All our ablations and baselines are trained using 8 H100 (80GB) GPUs using Deepspeed [43].

Evaluation benchmarks. To test performance, we evaluate the answer accuracy on spatial questions for images and videos across four recent benchmarks that test a variety of spatial reasoning capabilities. Below we highlight the specific benchmarks we use (also displayed in Figure 3):

BLINK [17] We focus on the spatial splits in BLINK. We group multi-view reasoning, jigsaw puzzles, and IQ tests as *reasoning* (*Reas* in the Tables) splits since they require comparing/manipulating visual elements.

SAT-Real (SAT-R) [45] This includes questions that require reasoning about actions based on one/two image(s).

VSI-Bench (VSI-B) [62] This is a video benchmark with questions about relationships and distances from other perspectives, and optimal routes. We group relative direction, relative distance, and route planning into the reasoning splits since they involve perspective-taking or planning. Absolute numbers may differ from some concurrent works due to differences in frame count, answer matching logic, and system prompts. We use consistent hyperparameters across all ablations.

We also include some broader datasets to test generalization, like ERQA [52] and MMSI-Bench [63].

Training datasets. We use three datasets: Video-R1, Zebra-CoT, and SAT as shown in Figure 3.

Video-R1 [16] is a general video reasoning dataset with reasoning traces containing 160K examples. It contains videos and images with questions spanning multiple topics like chart understanding, spatial reasoning, and perception tasks. When we use the dataset without the reasoning trace and supervise the answer directly, we denote that as Video-R1-Ans in the tables.

Zebra-CoT [30] is a recent image-text interleaved CoT dataset with reasoning traces that include both text and subgoal images as reasoning traces towards the answer for a range of visual and spatial tasks such as playing chess, solving puzzles, and counting.

SAT [45] contains 175K simulated examples of spatial reasoning questions and answers with both single-image and multi-image inputs.

Baselines. All baselines are trained using the training datasets listed above. We explore four baseline approaches.

+DirAns trains the base model directly on the final answer without any reasoning traces.

+TextCoT trains using the direct answers from SAT [45] and ZebraCoT [30] as well as the text-based reasoning traces from Video-R1 [16]. *+GRPO* further fine-tunes the above with GRPO [8, 16, 22] to refine the reasoning trace using only the final answer. We use a group size of 2 following recent work [19] and a batch size of 256 with gradient accumulation for stability.

+Interleave Im-Txt replicates a recent approach [65] that

Table 1. *Mull*-Tokens (row f, g) improve over direct answer finetuning (row b), and existing methods for visual reasoning tasks using text-only reasoning (row c, d) [8, 16], or text-image interleaving (row e) [65]. Data splits are described in Section 4.1.

Model	**BLINK**							**SAT-R**	**VSI**		**ERQA**	**Avg(All)**
	MV	**RelDep**	**SpRel**	**Jig**	**IQT**	**Avg**	**Reas**		**Avg**	**Reas**		
a. Qwen2.5-VL (7B)	39.00	61.29	**92.38**	58.66	25.33	55.33	41.00	59.00	23.96	22.96	**38.91**	44.30
b. + DirAns FT	57.14	**87.09**	74.12	58.66	30.00	61.40	48.60	71.66	32.40	30.65	38.00	50.87
c. + TextCoT FT	53.38	74.19	67.83	69.30	25.33	58.01	49.34	68.33	30.47	31.04	38.78	48.90
Δ *(vs. DirAns)*	*-3.76*	*-12.90*	*-6.29*	*+10.64*	*-4.67*	*-3.39*	*+0.74*	*-3.33*	*-1.93*	*+0.39*	*+0.78*	*-1.97*
d. + GRPO	54.88	71.77	69.23	72.00	25.33	58.64	50.74	69.00	30.34	30.15	36.00	48.50
Δ *(vs. DirAns)*	*-2.26*	*-15.32*	*-4.89*	*+13.34*	*-4.67*	*-2.76*	*+2.14*	*-2.66*	*-2.06*	*-0.50*	*-2.00*	*-2.37*
e. + Interleave Im-Txt	57.14	69.36	75.52	68.67	25.33	59.20	50.38	74.00	30.27	32.96	38.50	50.49
Δ *(vs. DirAns)*	*+0.00*	*-17.73*	*+1.40*	*+10.01*	*-4.67*	*-2.20*	*+1.78*	*+2.34*	*-2.13*	*+2.31*	*+0.50*	*-0.38*
f. + *Mull*-Tokens	63.15	83.06	81.81	74.00	**32.00**	66.80	56.38	**77.66**	32.98	32.85	38.25	53.92
Δ *(vs. DirAns)*	*+6.01*	*-4.03*	*+7.69*	*+15.34*	*+2.00*	*+5.40*	*+7.78*	*+6.00*	*+0.58*	*+2.20*	*+0.25*	*+3.05*
g. + GRPO	**64.66**	83.87	81.81	**74.67**	30.66	**67.13**	**56.66**	77.00	**33.51**	**33.49**	38.50	**54.04**
Δ *(vs. DirAns)*	*+7.52*	*-3.22*	*+7.69*	*+16.01*	*+0.66*	*+5.73*	*+8.06*	*+5.34*	*+1.11*	*+2.84*	*+0.50*	*+3.17*

Table 2. Warming up *Mull*-Tokens with multimodal image-text CoT (MM warm-up, row e) is crucial for performance, compared to no warm-up (row c), text-only warm up (row d). Data splits are described in Section 4.1.

Model	**BLINK**		**VSI**		**SAT-R**	**ERQA**
	Avg	Reas	Avg	Reas		
a. QwenVL	55.3	41.0	24.0	23.0	59.0	**38.9**
b. + DirAns FT	61.4	48.6	32.4	30.7	71.7	38.0
c. + No warm-up	59.2	45.2	29.8	27.4	67.3	37.8
d. + Txt warm-up	65.9	52.9	32.0	**33.5**	71.3	38.5
e. + MM warm-up	**66.8**	**56.4**	**33.0**	32.9	**77.7**	38.3

uses latents to encode visual information and *retains* explicit, text-based reasoning traces. We follow a similar two-stage approach to our method. Instead of warming up the <*Mull*> with both image and text, we interleave text with image-only supervision for <*Mull*>. During stage 2, we relax the loss on the image <*Mull*> tokens.

4.2. Main results

Text-based reasoning improves base model, but hurts compared to direct answer fine-tuning. We observe that supervising the model with text-based reasoning traces, while improves upon the base model performances as also found in concurrent work [16], *hurts* performance compared to directly tuning on the final answer of the same datasets. The results in Table 1 show that direct answer fine-tuning (row b) is a strong baseline, outperforming the text-

Table 3. We can add *Mull*-Tokens before predicting the explicit text reasoning and the final answer (row d). This improves answering performance compared to directly predicting the text reasoning (row b) or having interleaved visual thoughts with text reasoning (row c) before predicting the answer.

Model	**BLINK**	**SAT-R**	**VSI**	**ERQA**	**Avg**
a. Qwen 2.5 VL 7B	55.3	59.0	24.0	**38.9**	44.3
b. Text	58.0	68.3	30.5	38.8	48.9
c. Img + Text	59.2	**74.0**	30.3	38.5	50.5
d. *Mull*-Tokens + Text	**62.4**	72.0	**32.6**	37.3	**51.1**

based reasoning fine-tuning (row c) and also after GRPO optimization of the reasoning chains (row d). However, we do note that text-based reasoning tends to help on more reasoning-heavy splits compared to just fine-tuning with the final answer - BLINK Jigsaw and reasoning average (by $\sim 2.1\%$) and VSI-Bench reasoning ($\sim 0.4\%$).

Interleaving image thoughts fail to improve significantly. Recall, that the intuitive solution to "fix" text-based reasoning is to include visual thoughts (image latents) along with the explicit text thoughts [65]. The results for such an approach applied to our datasets (described in Section 4.1) are shown in Table 1 in row e. We see that it indeed performs better than text-based reasoning (row e vs d), with more gains seen on video reasoning splits (VSI reasoning) and on splits that require reasoning about action consequence and perspectives (SAT-Real and ERQA). However, it fails to improve overall from direct-answer finetuning across all benchmarks except SAT [45]. It also fails to improve on reasoning splits that are harder and out-of-domain such as

BLINK IQ Tests. Empirically analyzing the interleaved reasoning outputs reveal two weaknesses - i) the model *rarely* switches to image thoughts and resorts to mostly text reasoning and hence, suffers the same weaknesses, and ii) if we prompt the model to force it to switch to image thoughts (more details in the supplementary), it actually reduces performance - *e.g*. on BLINK, performance went down by 2% when forced to reason with image thoughts. We observe that the text still suffers from drifting away from visual inputs even after a visual thought (examples in supplementary). This suggests that while interleaving image thoughts may be helpful sometimes, the model still cannot effectively use it to think towards an accurate answer.

Using modality-agnostic *Mull*-Tokens performs best overall. As shown in Table 1 row f and g, we see that our *Mull*-Tokens, which are agnostic to text or image modality, perform best overall, including reasoning splits that harder out-of-domain tasks. Our *Mull*-Tokens (after stage 2) (row f) improves compared to direct-answer finetuning (row b), where various other reasoning approaches (row c, d, e, f) fail to. Notably, we have strong improvements in reasoning about camera movements, jigsaw puzzles and IQ tests (under BLINK column) that require reasoning with visual and symbolic modalities. Finally, we observe that latest advancements in RL, GRPO refinement, when applied to our *Mull*-Tokens (row g) can further improve the reasoning-heavy splits in BLINK and VSI-Bench. Improvements to interleaved image-text reasoning (row e) suggest that effective thinking compute requires *Mull*-Tokens to be *modality-agnostic* rather than restricted to vision. Notably, *Mull*-Tokens achieve these gains using only 20 tokens, compared to 200–500 tokens in a typical text CoT rationale, translating to significant inference speedup over verbose reasoning approaches. We also note that on ERQA, all fine-tuned variants perform close to the base model. We attribute this to ERQA being perception-heavy: many questions ask about directly visible object states or trajectories, limiting the headroom for reasoning-based improvements.

***Mull*-Tokens generalize to more complex multi-image benchmarks** To test broader generalization, we evaluate our on MMSI-Bench (multi-image spatial) and SiteBench (general spatial, 3k randomly sampled due to time constraints). Table 4 shows we improve multi-step reasoning (MSR, +1.2%), judging attributes from different perspectives (Appr. +8.0%), and also general spatial tasks (SiteBench +2.1%), suggesting broader generalization.

Adding *Mull*-Tokens can improve answering accuracy while also predicting the text rationale. While *Mull*-Tokens without text-based reasoning improves overall performance, predicting the text-based reasoning (or rationale) can often be of educational and interpretable value. Hence, we explore if *Mull*-Tokens can also be appended before predicting both the text rationale and the final answer. Intuitively, this allows the model to think in an abstract modality-agnostic space before verbalizing the rationale and predicting the answer. To perform this analysis, we include both the text rationale and the answers from the same training datasets in Stage 2 (Section 3) of our training after the Mull tokens. Since we do not have a reliable way to evaluate the rationales, we report the final answer accuracy in Table 3 and show some qualitative results in the supplementary. We observe that the answering accuracy after predicting the text rationale (row d) is higher than if we were to directly predict the rationale and answer (row b). This result is significant for interpretability: while *Mull*-Tokens are not directly decodable into human-readable output, they can be paired with explicit text rationales, retaining the accuracy benefits of latent reasoning *while also* producing interpretable reasoning traces.

Table 4. Performance on other multi-image or general-purpose spatial benchmarks

	MMSI-Bench			SITE-B
Model	**MSR**	**Attr.**	**Avg**	**Avg**
Qwen 7B	25.8	18.2/25.0	25.9	61.3
+DirAns	27.4	25.3/25.7	28.8	63.5
+*Mull*-Tokens	**28.7**	**33.3/28.6**	**29.0**	**65.6**

4.3. Ablations

We further ablate what causes the performance gains. We explore whether the *Mull*-Tokens utilize the multimodal information in its reasoning latents. The model could just be taking advantage of a wider computational pathway of meaningless tokens instead.

No Warm-up. If the model were to simply use *Mull*-Tokens as a wider computational pathway, we would see gains if we only fine-tuned according to Stage 2 (Section 3 with the new tokens (without any multimodal information imparted into them) as extra compute for the model to optimize to arrive at the final answer. Hence, we perform this ablation and perform only Stage 2 tuning using direct answers from our training datasets

Text-only Warm-up. To ablate whether multimodal - *i.e*. including images with the text - is necessary, or text alone is sufficient, we warm-up the *Mull*-Tokens using *only* the text-based CoT traces from our training data, discarding any traces that include interleaved images. We still use the final answers from the image COT trace dataset to not let domain differences in the datasets affect the performance.

Our final approach. Our final *Mull*-Tokens approach, warms up the tokens on all available interleaved image-and-text reasoning traces in stage 1 (denoted by **MM warm-up**) before performing stage 2 training without any constraints on our latent tokens and only using the final answers from all our training datasets.

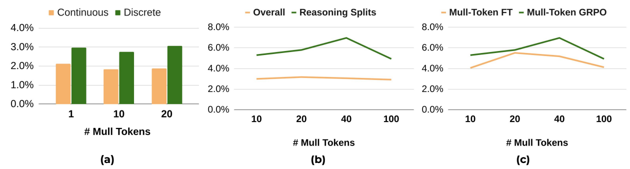

Figure 4. Hyperparameter choices in latent design. (a) Choice of continuous embeddings vs discrete tokens: We see that discrete performs better for each k (number of latent at test time) values. (b) The impact of number of latent tokens at test time: we see that too many latents can hurt performance. Reasoning splits tend to improve with more Mull-Tokens, but too many is still detrimental. (c) We see that performance scales with number of latent tokens more after GRPO since GRPO rewards the latent chain causally.

Continuous Embeddings vs Discrete Tokens Recall that the choices of using latent Mull-Tokens tokens in stage 2 of the training (Section 3 and during inference are to use continuous embeddings that are recurrently fed in for k steps [23], or use discrete <Mull> tokens [20]. For the continuous ablation, in stage 2, for each latent $\texttt{MULL}_k$, we feed in the previous $\texttt{MULL}_{k-1}$ embedding output before the language head of the Qwen model - both during inference and training.

4.4. Analysis

Mull-Tokens with both image and text anchoring is crucial to performance. While Mull-Tokens is effective, a question remains whether Mull-Tokens represents reasoning in a multimodal space or simply using extra computation [20] to improve performance. Hence, we ablate the choice of warming-up our Mull-Tokens as described in Section 4.3 and present the results in Table 2. We find that multimodal warm-up allowing the Mull-Tokens to be multimodal - *i.e*. can hold meanings in either text or images - to be beneficial for performance. Mull-Tokens used simply as extra compute (row c) improves upon the base model by 4.2%, however, undeperforms the direct answer baseline (row b) Mull-Tokens with text-only warm-up (row d) offers only a marginal 1.07% improvement over the direct answer fine-tuning baseline (row b), showing that a text-only warm-up, while somewhat beneficial, is still insufficient. Finally, warming up with both image and text achieves the highest gains ($+3.05\%$) over the direct answer fine-tuning. This suggests that for Mull-Tokens to be effective, they *must* be pre-trained to handle *both* visual and textual information simultaneously.

Discrete latent tokens over continuous embeddings. Both strategies improve upon the direct answer finetuning baseline as shown in Figure 4 - plot (a). However, discrete tokens perform better. Performance also starts to degrade with more continuous latents since errors accumulate with longer continuous embeddings, also seen by concurrent works [65]. Discrete also allows us to use the latest optimization tricks, such as token parallelism in standard frameworks for MLMs, resulting in significantly faster training and inference.

Scaling effect of latent tokens. We investigate the effect of scaling the number of latent tokens used during inference - both after the Mull-Tokens FT and GRPO stages. Our model was trained using 20 latents. The results in Figure 4 - (b) and (c) show the tradeoff. We first see that reasoning tasks tend to benefit more with a higher number of tokens as shown in plot (b), but performance degrades with too many tokens- reminiscent of degradation due to overthinking in LLMs [58]. We see that performance scales more positively after the GRPO phase, as shown in (c). We believe this might be because GRPO rewards the optimal <Mull> chains on-policy, unlike SFT.

5. Discussion

Limitations. Our experiments focus on the Qwen2.5-VL (7B) model [2]; while effective, the generalization of our findings to other backbones and model scales remains to be verified. Based on trends in latent reasoning for language models [20, 46], where gains grow with model capacity, we anticipate our approach to benefit similarly from larger base models. A more intrinsic limitation is the lack of direct interpretability for the latent tokens. As these tokens are modality-agnostic—representing neither pure text nor pure image features—they cannot be directly decoded into human-readable output. However, from a performance angle, a few Mull-Tokens improve performance over verbose, albeit interpretable, text thoughts. As shown in Table 3, Mull-Tokens can also be used *in conjunction* with interpretable text thoughts, offering both accuracy gains and interpretability.

Future work. We identify several key avenues for future work. First, extension of our approach to include diverse modalities such as 3D point clouds, audio, and other structured data, a direction currently constrained by the scarcity of suitable reasoning datasets. Finally, we see significant potential in integrating learning from world models [3] to discover robust causal reasoning chains, which is essential for bridging the remaining performance gap between current models and human capabilities.

Conclusion. In this work, we present a novel fine-tuning paradigm that significantly improves multimodal reasoning compared to existing text-only or image-text interleaved reasoning approaches. Our multimodal latent <Mull> tokens are Pareto-dominant: they improve performance ($+3\%$) while being computationally more efficient (20 tokens vs 100's). We hope this work paves the way for efficiently incorporating causal multimodal reasoning in autoregressive AI models.

Acknowledgments

We wish to thank Howard Zhou, Andre Araujo, Kevis-Kokitsi Maninis, and Ye Xia for helpful initial discussions on the directions. We also thank Mi Luo, Ellis Brown, and Saining Xie for their thoughts that helped polish the paper. We would like to thank the Google XCloud and Cloud Storage Teams for providing a solid infrastructure to train and evaluate open-source models on open-source datasets and benchmarks. Finally, we would also like to acknowledge the excellent codebases provided by Video-R1 and MIRAGE authors that we built upon. RK was partially supported by SRC Jump 2.0 CoCoSys.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 5, 8

[3] Philip J. Ball and Jakob Bauer et al. Genie 3: A new frontier for world models, 2025. 8

[4] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3836–3845, 2025. 1, 2, 3, 4

[5] Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, and Saining Xie. SIMS-V: Simulated instruction-tuning for spatial video understanding, 2025. 1, 2

[6] Zikui Cai, Andrew Wang, Anirudh Satheesh, Ankit Nakhawa, Hyunwoo Jae, Keenan Powell, Minghui Liu, Neel Jay, Sungbin Oh, Xiyao Wang, et al. Morse-500: A programmatically controllable video benchmark to stress-test multimodal reasoning. *arXiv preprint arXiv:2506.05523*, 2025. 1

[7] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 1

[8] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V`, 2025. Accessed: 2025-02-02. 4, 5, 6

[9] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models, 2025. 2

[10] Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025. 2

[11] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, pages 135062–135093. Curran Associates, Inc., 2024. 2

[12] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 1

[13] Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025. 2

[14] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. Mm-spatial: Exploring 3d spatial understanding in multimodal LLMs, 2025. 2

[15] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems*, pages 70757–70798. Curran Associates, Inc., 2023. 2

[16] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025. 2, 4, 5, 6

[17] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 2, 5

[18] Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025. 2

[19] Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. The delta learning hypothesis: Preference tuning on weak data can yield strong gains, 2025. 5

[20] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 4, 5, 8

[21] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025. 1, 2, 3

[22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma,

Xiao Bi, et al. Deepseek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 2, 4, 5

[23] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024. 2, 4, 5, 8

[24] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025. 1

[25] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024. 1, 2, 3

[26] Jingjing Jiang, Chao Ma, Xurui Song, Hanwang Zhang, and Jun Luo. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3034–3046, 2025. 1

[27] Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, et al. Explain before you answer: A survey on compositional visual reasoning. *arXiv preprint arXiv:2508.17298*, 2025. 2

[28] Michael Kleinman, Matthew Trager, Alessandro Achille, Wei Xia, and Stefano Soatto. e1: Learning adaptive control of reasoning effort. *arXiv preprint arXiv:2510.27042*, 2025. 2

[29] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025. 1, 2

[30] Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025. 1, 2, 5

[31] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 2

[32] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought, 2025. 2

[33] Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *arXiv preprint arXiv:2506.04633*, 2025. 1

[34] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025. 2

[35] Disheng Liu, Tuo Liang, Zhe Hu, Jierui Peng, Yiren Lu, Yi Xu, Yun Fu, and Yu Yin. Deconstructing spatial intelligence in vision-language models, 2025. 2

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 2, 3

[37] Yanzuo Lu, Xin Xia, Manlin Zhang, Huafeng Kuang, Jianbin Zheng, Yuxi Ren, and Xuefeng Xiao. Hyper-bagel: A unified acceleration framework for multimodal understanding and generation, 2025. 2

[38] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning, 2025. 2, 3, 4, 5

[39] Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, Ranjay Krishna, and Silvio Savarese. Taco: Learning multi-modal action models with synthetic chains-of-thought-and-action, 2024. 1

[40] Hanspeter A Mallot. *From geometry to behavior: An introduction to spatial cognition*. MIT Press, 2024. 1

[41] Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. Tips: Text-image pretraining with spatial awareness, 2025. 2

[42] OpenAI. Thinking with images. `https://openai.com/index/thinking-with-images/`, 2025. 1, 2

[43] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. 5

[44] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A. Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval, 2023. 2

[45] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 1, 2, 5, 6

[46] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing Chain-of-Thought into Continuous Space via Self-Distillation. *arXiv preprint arXiv:2502.21074*, 2025. 2, 8

[47] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 2

[48] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[49] Renliang Sun, Wei Cheng, Dawei Li, Haifeng Chen, and Wei Wang. Stop when enough: Adaptive early-stopping for chain-of-thought reasoning. *arXiv preprint arXiv:2510.10103*, 2025. 2

[50] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 2

[51] CVBench Team. Cvbench: A benchmark for cross-video multimodal reasoning, 2025. 2

[52] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. 2, 5

[53] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs, 2024. 2

[54] Barbara Tversky and Masaki Suwa. *Thinking with sketches*, pages 75–84. Oxford University Press, 2009. 1

[55] Marina Vasilyeva and Stella F Lourenco. Development of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3):349–362, 2012. 1

[56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837. Curran Associates, Inc., 2022. 2

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1

[58] Zihao Wei, Liang Pang, Jiahao Liu, Jingcheng Deng, Shicheng Xu, Zenghao Duan, Jingang Wang, Fei Sun, Xunliang Cai, Huawei Shen, and Xueqi Cheng. Stop spinning wheels: Mitigating LLM overthinking via mining patterns for early reasoning exit, 2025. 8

[59] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12966–12977, 2025. 2

[60] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[61] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 2

[62] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643, 2025. 1, 2, 3, 5

[63] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence, 2025. 2, 5

[64] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video, 2025. 2

[65] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens, 2025. 1, 2, 4, 5, 6, 8

[66] Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. Hybrid latent reasoning via reinforcement learning. *arXiv preprint arXiv:2505.18454*, 2025. 2

[67] Chi Zhang, Haibo Qiu, Qiming Zhang, Zhixiong Zeng, Lin Ma, and Jing Zhang. Deepsketcher: Internalizing visual manipulation for multimodal reasoning. *arXiv preprint arXiv:2509.25866*, 2025. 2

[68] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025. 2

[69] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. 2

[70] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713, 2025. 2

[71] Xu Zheng, Zihao Dongfang, Lutao Jiang, Boyuan Zheng, Yulong Guo, Zhenquan Zhang, Giuliano Albanese, Runyi Yang, Mengjiao Ma, Zixin Zhang, et al. Multimodal spatial reasoning in the large model era: A survey and benchmarks. *https://arxiv.org/abs/2510.25760*, 2025. 2

[72] Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, et al. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. *arXiv preprint arXiv:2509.25373*, 2025. 2

[73] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual

reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024. 1

[74] Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025. 2