UNI-LADIR: LATENT DIFFUSION UNIFIES MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models increasingly think with different modalities such as images, 3D point clouds, and robot states, not just text. Yet each modality is still encoded into its own representation space, so successive reasoning steps can span different modality-specific representations. In this paper, we introduce Uni-LaDiR (**Uni**fied **La**tent **Di**ffusion **R**easoner), a framework that unifies different modalities into a shared latent space for multimodal reasoning. A unified encoder maps teacher reasoning steps from different modalities into latent thought tokens in a shared space, trained to extract the information needed for later reasoning steps and the final output. A diffusion reasoner, trained jointly with the encoder, generates these tokens at inference without teacher reasoning steps. Across eleven vision-language model (VLM) benchmarks and two vision-language-action (VLA) suites, Uni-LaDiR achieves relative gains over the strongest baselines of **7.3%** on four mathematical and logical VLM benchmarks and **6.1%** on RLBench manipulation tasks. Controlled comparisons show increasing gains as more teacher modalities are unified. These results suggest that unification improves multimodal reasoning by weaving it into a single thread, where the model predicts successive thoughts in a common representation space.

1 INTRODUCTION

Multimodal reasoning often requires combining information from different modalities. An image captures visual appearance, language describes events and relations, 3D point clouds provide geometric structure, and robot states describe interaction with the environment. These modalities provide complementary information, and many tasks require using several of them together rather than relying on any single modality (Lu et al., 2022b; 2024; Hao et al., 2025). A capable multimodal model should therefore be able to integrate different modalities throughout its reasoning process (Gao et al., 2024; Gu et al., 2025).

Recent work has explored several ways to incorporate multiple modalities into Chain-of-Thought (CoT) reasoning (Wei et al., 2022). In vision-language models (VLMs), reasoning chains can interleave text with visual steps, represented as either generated images (Li et al., 2025c; Gu et al., 2025) or latent visual tokens (Li et al., 2025b; Yang et al., 2025). Vision-language-action models (VLAs) similarly combine language, visual observations, and actions (Brohan et al., 2023; Kim et al., 2024; Zhao et al., 2025), with some methods additionally incorporating 3D information and robot states (Zhen et al., 2024; Qu et al., 2025; Liu et al., 2026).

However, placing different modalities in the same reasoning sequence does not make their representations shared. Most existing methods still encode each modality in its own representation space and then interleave the resulting tokens in the LLM sequence (Gao et al., 2024; Li et al., 2025b; Yang et al., 2025; Liu et al., 2026). An image reasoning step may therefore be represented by visual features, while a 3D step or robot-state step uses a different representation. When reasoning moves from one modality to another, the model predicts the next step across different modality-specific representation spaces.

This motivates a simple idea: represent all intermediate reasoning steps in a shared latent space. The reasoning content may change across modalities, while its latent representation remains in the same space. A shared latent interface allows the reasoning model to predict successive steps in a common

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

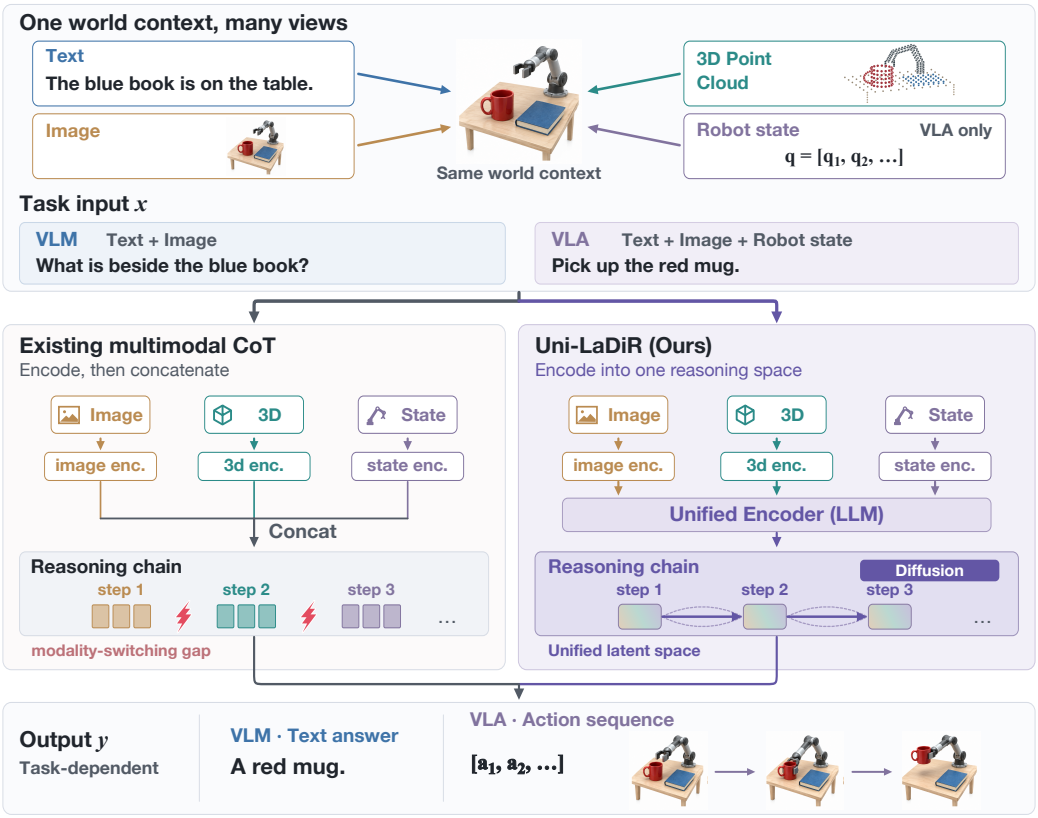

Figure 1: **Multimodal reasoning in a shared latent space.** Modality-specific reasoning (left) switches representations between steps. Uni-LaDiR (right) uses a shared latent space and diffusion to generate thought tokens for answers or actions.

representation space. Later steps can use earlier latent representations through this shared interface, regardless of their source modality.

The main challenge is how to learn such a shared reasoning space. Adopting the representation of any single modality would retain its modality-specific structure. Instead, we learn the representation according to what later reasoning steps need. Based on this idea, we introduce Uni-LaDiR (**Uni**fied **La**tent **Di**ffusion **R**easoner), a framework that maps reasoning steps from different modalities into a shared latent space. A unified encoder converts each teacher reasoning step into a fixed-sized block of latent thought tokens. Rather than reconstructing the original modality, these tokens are trained to predict subsequent reasoning steps and the final output. The learned representation is thus shaped by the information needed for future reasoning, rather than by the format of the original modality.

At inference time, the teacher reasoning steps are unavailable, so the model must generate these thought tokens from the task input and its previous reasoning. We use a diffusion reasoner to model this generation process (Ho et al., 2020; Rombach et al., 2022; Chi et al., 2023; Kang et al., 2025b; 2026b). Diffusion is a natural choice here, since the same context may admit multiple valid next thoughts. We jointly train the unified encoder and diffusion reasoner so that the latent thoughts satisfy two requirements: they retain information useful for subsequent reasoning and the final output, while remaining predictable from the context available at inference time.

We evaluate Uni-LaDiR on eleven VLM reasoning benchmarks and two VLA manipulation suites. Compared with the strongest modality-specific latent reasoning baseline in each setting, Uni-LaDiR improves mean accuracy by **7.3%** relative across four mathematical and logical VLM benchmarks and mean success by **6.1%** relative across ten RLBench manipulation tasks. Controlled experiments further isolate the effect of unification. Bringing more teacher modalities into the shared latent space consistently improves performance. These results support the central hypothesis of Uni-LaDiR:

multimodal reasoning benefits from keeping intermediate reasoning steps in a shared representation space, even when those steps originate from different modalities.

2 PRELIMINARIES

Existing multimodal reasoning methods often incorporate different modalities into their reasoning chains to combine complementary evidence for answering questions or planning actions (Gao et al., 2024; Li et al., 2025c; Gu et al., 2025; Zhao et al., 2025). In latent multimodal CoT, intermediate content is represented by modality-specific latent blocks (Li et al., 2025b; Yang et al., 2025; Liu et al., 2026). For a chain of K steps, let m_i denote the modality of step i and $z_i^{(m_i)} \in \mathcal{Z}_{m_i}$ its latent block, where $\mathcal{Z}_m$ is the latent space of modality m. The blocks are concatenated into a single reasoning sequence:

$$[x;\, z_1^{(m_1)};\, \cdots\, ;\, z_K^{(m_K)}] \;\longrightarrow\; y. \tag{1}$$

Here x is the task input, y is the task output, and $[\,;\,]$ denotes sequence concatenation. During training, modality-specific encoders provide teacher features s_i to supervise the latent blocks; at inference, the model generates these blocks from x and preceding blocks. The same formulation covers VLMs and VLAs (Figure 1): for VLMs, x contains an image and a question, and y is a text answer; for VLAs, x contains visual observations (including point clouds when available), a language instruction, and the current robot state, and y is an action sequence. Because each block $z_i^{(m_i)}$ lies in the space of its own modality, predicting the next step can span different latent spaces, $\mathcal{Z}_{m_i} \to \mathcal{Z}_{m_{i+1}}$.

3 UNIFIED LATENT DIFFUSION REASONING

Uni-LaDiR encodes reasoning steps from all modalities into a shared latent space, allowing the model to predict successive steps through a common latent interface. In this section, we first formalize the requirements of this shared latent space (Section 3.1), then describe how teacher steps are mapped to shared thought tokens (Section 3.2), how continuation prediction trains these tokens for subsequent reasoning and task prediction (Section 3.3), and how latent diffusion learns to generate them from the context available at inference (Section 3.4). Figure 2 provides an overview.

3.1 MOTIVATION

Recent work concatenates modality-specific reasoning steps into a single sequence. When adjacent steps use different modalities, their latent spaces differ, $\mathcal{Z}_{m_i} \neq \mathcal{Z}_{m_{i+1}}$, so the model predicts successive reasoning steps across modality-specific representation spaces (Eq. 2, top). We use a unified encoder E_θ, with parameters θ, to map teacher features s_i to latent thought tokens $z_i^\star = E_\theta(s_i)$ in a shared latent space $\mathcal{Z}$:

$$\begin{array}{lcc} & \text{reasoning} & \text{representation spaces} \\ \text{modality-specific concat:} & z_i^{(m_i)} \longrightarrow z_{i+1}^{(m_{i+1})} & \underbrace{\mathcal{Z}_{m_i} \longrightarrow \mathcal{Z}_{m_{i+1}}}_{\text{switch}} \\ \text{Uni-LaDiR:} & z_i^\star \longrightarrow z_{i+1}^\star & \underbrace{\mathcal{Z} \longrightarrow \mathcal{Z}}_{\text{no switch}} \end{array} \tag{2}$$

Now both $z_i^\star$ and $z_{i+1}^\star$ lie in $\mathcal{Z}$ (Eq. 2, bottom): the thought changes, but its format stays the same. The model predicts successive steps within this shared space, using earlier latent tokens as context. Rather than adopting the features of any one modality, we learn this space from the reasoning itself. We train these thought tokens through *continuation prediction* to retain information needed for later reasoning and the final output (*reasoning sufficiency*), and jointly train a diffusion reasoner to generate them from the available context (*context predictability*).

3.2 ENCODING REASONING STEPS INTO A SHARED LATENT SPACE

We train a unified encoder to map each teacher reasoning step to the corresponding thought tokens $z_i^\star$ in the shared latent space (Figure 2a). After modality-specific encoding, we append a fixed-length block of learnable embeddings q to each teacher feature s_i. Then our unified encoder maps the last hidden states at the q positions to the clean thought tokens $z_i^\star$.

Every modality uses the same q, output projection, and number of thought tokens per block. During training, we pack multiple teacher steps and their corresponding copies of q into one forward pass.

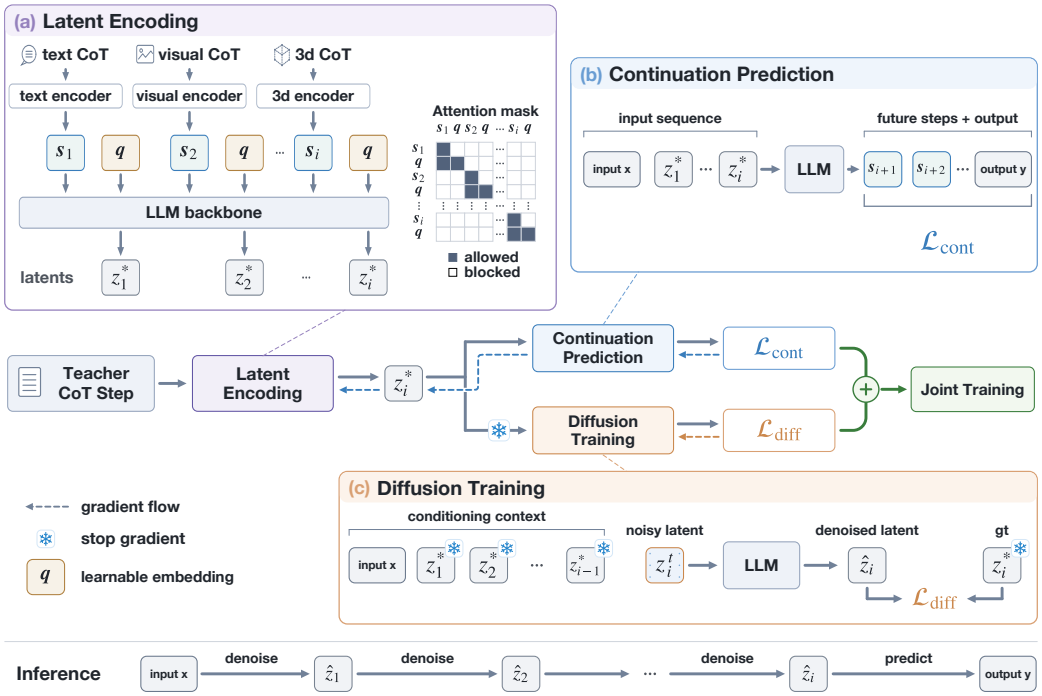

Figure 2: **Uni-LaDiR overview.** During training, a shared backbone **(a)** encodes teacher steps into thought tokens, **(b)** uses them to predict later teacher steps and the final output, and **(c)** learns their generation by diffusion. At inference, it generates thought blocks sequentially by diffusion, conditioned on the input and preceding blocks, then predicts the answer or action.

The latent-encoding mask isolates the steps: each copy of q attends only to itself and its associated teacher features, preventing information exchange across reasoning steps (Figure 2a). Thus, all teacher modalities are now encoded into a shared latent space, where later reasoning steps, rather than the source format, determine which information is retained (Section 3.3).

3.3 Learning thought tokens through continuation prediction

Continuation prediction trains each thought block to preserve the information later reasoning will need. Given x and the prefix $z^\star_{<i} = [z^\star_1; \ldots; z^\star_{i-1}]$, the backbone predicts the next teacher target s^{GT}_i; given the complete sequence $z^\star_{1:K} = [z^\star_1; \ldots; z^\star_K]$, it predicts the final output y (Figure 2b). Here s_i denotes the teacher features mapped to a thought block, whereas s^{GT}_i denotes the fixed supervision target for that step: text tokens or continuous teacher features. A continuation-prediction attention mask computes these predictions in one forward pass while exposing each prediction only to its permitted thought-token prefix and hiding later blocks and raw teacher steps (Appendix A.1).

Let $\hat{s}_i$ and $\hat{y}$ denote the corresponding predictions. We optimize

$$\mathcal{L}_{\mathrm{cont}} = \underbrace{\sum_{i=2}^{K} \ell(\hat{s}_i, s^{\mathrm{GT}}_i)}_{\text{reasoning steps}} + \lambda_y \underbrace{\mathcal{L}_{\mathrm{out}}}_{\text{final output}}, \qquad \ell(\hat{u}, u) = \begin{cases} \mathrm{CE}(\hat{u}, u), & u \text{ is text,} \\ \|\hat{u} - u\|_2^2, & u \text{ is continuous.} \end{cases} \tag{3}$$

Here u is a reasoning-step target, $\hat{u}$ its prediction, and CE the cross-entropy loss. The weight λ_y scales the final-output loss $\mathcal{L}_{\mathrm{out}}$. The continuous reasoning-step targets are image and 3D point-cloud features and robot-state vectors, supervised by the L2 term above. For VLMs, the final output y is a text answer, so we use cross-entropy loss, $\mathcal{L}_{\mathrm{out}} = \mathrm{CE}(\hat{y}, y)$. For VLAs, y is a continuous action sequence, which we model with a flow-matching action head; $\mathcal{L}_{\mathrm{out}}$ is its flow-matching loss. The targets s^{GT}_i and y are fixed, and any encoder used to construct target features is frozen. Gradients from $\mathcal{L}_{\mathrm{cont}}$ pass through $z^\star_i$ into the unified encoder and shared backbone, so the latent targets retain information useful for the continuation.

3.4 GENERATING THOUGHT TOKENS WITH JOINT DIFFUSION TRAINING

The diffusion objective trains the backbone to generate each thought block from the context available at inference (Figure 2c). Diffusion models a conditional distribution over the next thought tokens, allowing different valid thoughts under the same input and preceding thought tokens (Yu et al., 2025; Kang et al., 2025b; 2026b). During training, we independently sample Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I)$ for each block and time $t \sim \mathcal{U}[0, 1]$, where I is the identity covariance. Let $\mathrm{sg}(\cdot)$ denote stop-gradient, shown by snowflake markers in Figure 2c. We form the noisy block $z_i^t = (1-t)\epsilon_i + t\,\mathrm{sg}(z_i^\star)$, with target velocity $v_i^\star = z_i^\star - \epsilon_i$. Conditioned on the noisy block z_i^t, x, the detached prefix $\mathrm{sg}(z_{<i}^\star)$, and t, the shared backbone predicts the velocity $\hat{v}_i$.

We again process all reasoning steps in one forward pass. The diffusion mask restricts each noisy block to its own tokens, x, and the preceding clean thought blocks, preventing access to future thought blocks (Appendix A.1). We optimize the flow-matching objective (Lipman et al., 2023)

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon}\left[\sum_{i=1}^{K} \|\hat{v}_i - \mathrm{sg}(v_i^\star)\|_2^2\right], \tag{4}$$

The expectation is over t and the block noises $\epsilon = (\epsilon_1, \ldots, \epsilon_K)$. Integrating the predicted velocity from $t = 0$ to 1 generates a thought block, denoted z_i. Thus, t indexes denoising within a block, while i indexes reasoning steps.

Joint training. A useful thought token needs to retain task-relevant information and be predictable from the context available at inference. To meet both requirements, we share the LLM weights between the unified encoder and diffusion reasoner (Duggal et al., 2026) and optimize

$$\mathcal{L} = \mathcal{L}_{\text{cont}} + \lambda \mathcal{L}_{\text{diff}}, \tag{5}$$

where λ balances the two losses. The continuation loss trains the encoder to retain information needed for later reasoning and the final output. The diffusion loss trains the shared backbone to predict these representations from context, so its updates also change how the encoder represents teacher steps. Although we stop diffusion gradients through the clean targets and conditioning prefix, both losses still update the shared LLM weights. Diffusion training therefore also updates the unified encoder, allowing task utility and context predictability to jointly shape the latent representations.

Inference. At inference, the teacher reasoning steps and unified encoding path are not required. Starting from x, the model generates $z_1, \ldots, z_K$ sequentially by denoising one latent block at a time, with each block conditioned on the previously generated blocks; K is not fixed, and generation stops once the model stops predicting <BOT> (Appendix B.2) (Figure 2, bottom). The complete latent sequence then conditions prediction of the final answer or action y.

4 EXPERIMENTS

Our experiments test whether a shared latent space improves both visual reasoning and robotic manipulation. Section 4.1 introduces the benchmarks and baselines; Section 4.2 reports task performance and manipulation inference efficiency; Section 4.3 isolates the effect of unifying teacher modalities; and Section 4.4 ablates joint training, the encoder, and training objectives. See Appendix B.4 for additional experimental results.

4.1 EVALUATION PROTOCOL

Visual reasoning. We train on Zebra-CoT (Li et al., 2025a), which interleaves text and visual reasoning steps, and use Qwen2.5-VL-7B (Bai et al., 2025b) for the main experiment. We report accuracy on seven vision-centric benchmarks: VisualPuzzles (Song et al., 2025), ChartQA (Masry et al., 2022), V* (Wu & Xie, 2024), BLINK-Jigsaw (Fu et al., 2024), MMVP (Tong et al., 2024b), SAT (Ray et al., 2024), and CV-Bench (Tong et al., 2024a). We also evaluate four mathematical or logical benchmarks: MathVista (Lu et al., 2024), MathVision (Wang et al., 2024b), VisuLogic (Xu et al., 2025b), and EMMA (Hao et al., 2025).

VLA manipulation. We evaluate LIBERO (Liu et al., 2023) and RLBench (James et al., 2020). In this setting, future images, 3D point clouds (RLBench only), and robot states serve as training-only teacher CoTs; continuous actions remain task outputs. Following Liu et al. (2026), our VLA backbone is a mixture-of-transformers (MoT) that couples two Janus-Pro-1.5B experts (3.3B parameters total), and actions are produced by a flow-matching action head trained with the same action objective. We report success rate and, on RLBench, control frequency without action chunking.

Table 1: **Comparison on VLM reasoning benchmarks.** Accuracy (%) of Uni-LaDiR and baselines with Qwen2.5-VL-7B on **(a)** vision-centric and **(b)** mathematical and logical benchmarks. Best results are highlighted in bold.

(a) Vision-Centric Perception, Search, and Spatial Reasoning.

Method	VisPuz.	ChartQA	V*	BLINK-J	MMVP	SAT	CV-B	Avg.
Explicit Text and Visual Reasoning								
Qwen2.5-VL (base)	34.8	78.1	76.4	59.3	77.3	51.3	75.2	64.6
DeepEyes	48.0	82.5	85.3	72.0	78.0	64.0	84.0	73.4
PixelReasoner	44.0	81.5	81.7	70.0	76.0	60.0	82.0	70.7
Latent Reasoning								
LVR	46.0	81.0	81.7	52.0	71.7	60.0	84.0	68.1
Mirage	44.0	80.5	83.8	68.0	74.0	72.0	83.0	72.2
CoVT	41.0	79.5	78.0	68.0	58.7	61.3	80.0	66.7
Monet	35.0	79.0	83.3	65.3	50.0	55.3	71.1	62.7
VaLR-M	45.0	82.0	86.9	72.0	60.3	65.3	87.6	71.3
SLVR	34.2	77.2	82.2	68.0	74.0	62.0	83.0	68.7
Mull-Tokens	48.0	81.0	83.8	74.7	76.0	**77.0**	84.0	74.9
Uni-LaDiR (Ours)	**57.4**	**84.2**	**89.5**	**75.3**	**82.0**	72.0	**88.8**	**78.5**

(b) Visual Mathematics and Logic.

Method	MathVista	MathVision	VisuLogic	EMMA	Avg.
Explicit Text and Visual Reasoning					
Qwen2.5-VL (base)	68.2	25.4	26.0	26.0	36.4
Vision-R1	**73.5**	41.0	15.2	24.3	38.5
Open-Vision-Reasoner	69.5	51.8	27.2	30.0	44.6
DeepEyes	70.1	26.6	27.8	29.0	38.4
PixelReasoner	68.9	34.2	23.4	23.5	37.5
Latent Reasoning					
Mirage	68.6	40.5	26.6	27.0	40.7
Monet	62.5	33.8	10.6	22.1	32.3
ILVR	71.1	49.0	29.3	33.3	45.7
Uni-LaDiR (Ours)	73.2	**54.3**	**33.0**	**35.6**	**49.0**

Table 2: **Comparison of Uni-LaDiR and baselines on RLBench.** We report success rates (%) and inference speed (Hz). Best results are highlighted in bold.

Method	Size	Box	Laptop	Toilet	Sweep	Fridge	Phone	Umbrella	Frame	Wine	Plants	Mean	Std.	Hz
Direct Policies														
OpenVLA	7B	60.0	35.0	75.0	55.0	85.0	20.0	30.0	15.0	20.0	5.0	40.0	2.0	6.3
SpatialVLA	–	80.0	70.0	85.0	20.0	80.0	15.0	25.0	40.0	15.0	30.0	46.0	3.0	7.9
CogACT	7B	90.0	80.0	95.0	50.0	85.0	50.0	55.0	45.0	30.0	25.0	61.0	4.0	9.8
$\pi_{0.5}$	3B	90.0	95.0	85.0	75.0	**100.0**	5.0	10.0	**80.0**	75.0	35.0	65.0	4.0	13.8
HybridVLA	7B	85.0	95.0	**100.0**	**90.0**	**100.0**	50.0	50.0	70.0	50.0	50.0	74.0	4.0	6.1
Explicit Reasoning														
ManipLLM	7B	50.0	80.0	40.0	20.0	80.0	35.0	10.0	25.0	15.0	20.0	38.0	4.0	2.2
CoT-VLA	–	95.0	75.0	**100.0**	80.0	65.0	50.0	40.0	50.0	55.0	50.0	66.0	3.0	1.1
Latent Reasoning														
FiS-VLA	7B	**100.0**	**100.0**	95.0	55.0	90.0	50.0	50.0	70.0	55.0	20.0	69.0	3.0	**21.9**
LaST_0	3.3B	95.0	95.0	**100.0**	80.0	85.0	75.0	75.0	70.0	85.0	60.0	82.0	3.0	15.4
Uni-LaDiR (Ours)	3.3B	**100.0**	**100.0**	**100.0**	85.0	90.0	**80.0**	**80.0**	75.0	**90.0**	**70.0**	**87.0**	3.0	16.5

Baselines. For VLMs, we compare with the Qwen2.5-VL base model and explicit reasoners DeepEyes (Zheng et al., 2025), PixelReasoner (Wang et al., 2025), Vision-R1 (Huang et al., 2025b), and Open-Vision-Reasoner (Wei et al., 2025). Latent reasoning baselines include LVR (Li et al., 2025b), Mirage (Yang et al., 2025), CoVT (Qin et al., 2025b), Monet (Wang et al., 2026), VaLR-M (Jeon et al., 2026), SLVR (Xu et al., 2026a), Mull-Tokens (Ray et al., 2025), and ILVR (Dong et al., 2026). For VLAs, we consider three groups. *Action policies* include OpenVLA (Kim et al., 2024), SpatialVLA (Qu et al., 2025), CogACT (Li et al., 2024a), $\pi_{0.5}$ (Physical Intelligence et al., 2025), OpenVLA-OFT (Kim et al., 2025), HybridVLA (Liu et al., 2025), and UniVLA (Bu et al., 2025). *Explicit reasoning models* include ManipLLM (Li et al., 2023), WorldVLA (Cen et al., 2025), CoT-VLA (Zhao et al., 2025), and FlowVLA (Zhong et al., 2025). *Latent reasoning models* include FiS-VLA (Chen et al., 2025a), FUTURE-VLA (Xu et al., 2026b), ConsisVLA-4D (Li et al., 2026c), LaST_0 (Liu et al., 2026), AVA-VLA (Lei & Shan, 2026), and PearlVLA with and without PRL (Yang & Shan, 2026). Full settings appear in Appendix B.

4.2 Main Results

VLM visual reasoning. On the seven vision-centric benchmarks, Uni-LaDiR improves average accuracy over Qwen2.5-VL-7B by 21.4%; on the four mathematical and logical benchmarks, the relative gain is 34.7% (Table 1). Reasoning through shared latent thoughts thus strengthens both perception-heavy and multi-step mathematical reasoning. Uni-LaDiR also outperforms

Table 3: **Comparison of Uni-LaDiR and baselines on LIBERO.** We report success rates (%) across four suites. Best results are highlighted in bold.

Method	Size	Spatial	Object	Goal	Long	Mean
Direct Policies						
UniVLA	–	92.6	93.8	86.6	63.0	84.0
OpenVLA	7B	84.7	88.4	79.2	53.7	76.5
SpatialVLA	–	88.2	89.9	78.6	55.5	78.1
CogACT	7B	97.2	98.0	90.2	88.8	93.6
$\pi_{0.5}$	3B	98.8	98.2	98.0	92.4	96.9
OpenVLA-OFT	7B	97.6	98.4	97.9	94.5	97.1
Explicit Reasoning						
WorldVLA	–	85.6	89.0	82.6	59.0	79.1
CoT-VLA	–	87.5	91.6	87.6	69.0	83.9
FlowVLA	8.5B	93.2	95.0	91.6	72.6	88.1
Latent Reasoning						
FUTURE-VLA	4B	89.6	99.0	95.2	81.2	91.3
ConsisVLA-4D	7B	98.8	**99.8**	98.0	95.6	98.1
LaST_0	3.3B	99.2	99.6	98.0	95.6	98.1
AVA-VLA	7B	97.8	99.4	97.8	**98.1**	98.3
PearlVLA	7B	99.2	99.6	98.2	96.8	98.5
PearlVLA + PRL	7B	99.4	**99.8**	98.4	97.2	98.7
Uni-LaDiR (Ours)	3.3B	**100.0**	99.0	**100.0**	97.6	**99.2**

the strongest latent baseline in each group, by 4.7% over Mull-Tokens and 7.3% over ILVR on average. Since these baselines tie each latent step to modality-specific targets, the gains suggest that predicting successive reasoning steps in a shared latent space is more effective than reasoning across modality-specific representations. The gains over modality-specific baselines reach 30.3% on VisualPuzzles over Mirage, and 10.8% on MathVision and 12.6% on VisuLogic over ILVR. These tasks require combining visual evidence with multi-step abstract reasoning, where a shared latent space lets later steps build on earlier visual findings regardless of their source modality.

VLA manipulation. Uni-LaDiR improves average success over the strongest prior latent policies by 0.45 percentage points over PearlVLA + PRL on LIBERO and 6.1% relative to LaST_0 on RLBench (Tables 3 and 2). LaST_0 shares our backbone, demonstrations, and evaluation protocol but keeps a separate latent space for each modality (Appendix B.1), so the RLBench gain mainly reflects the benefit of a shared latent space. On RLBench, Uni-LaDiR improves nine of ten tasks and matches the remaining one; the largest gain is 16.7% on Plants. The benefit is therefore not tied to a few tasks, but holds broadly when reasoning spans images, 3D point clouds, and robot states. Without action chunking, Uni-LaDiR runs at **16.5 Hz**, a 7.1% higher inference rate than LaST_0 and **15.0**$\times$ the rate of explicit CoT-VLA. Reasoning in a compact shared latent space thus improves success without sacrificing control frequency, unlike explicit CoT that decodes intermediate images.

4.3 Analysis

Unifying more modalities. We first test whether all modalities need to be unified, or whether a subset suffices. The Separate baseline is the continuation-trained separate-encoding variant in Figure 5. We progressively route teacher modalities through the unified encoder: partial variants unify selected modalities while the other modalities keep their own LLM encoders, and the full variant unifies all modalities. Teacher modalities and continuation supervision remain fixed. On the VLM benchmarks, unifying text and image improves mean accuracy by 11.3% relative to separate encoders (Figure 3a). On RLBench, unifying any pair of the visual, 3D, and state modalities yields 1.8–3.0% relative gains, whereas unifying all three yields 6.1% over separate encoders and 3.0% over the best pair (Figure 3b). These results show that gains grow as more modalities are unified, and that leaving any one modality out limits the benefit.

4.4 Ablation Study

Joint training. To test whether the encoder and reasoner benefit from adapting together, we compare separate-stage and joint training under the same unified architecture and continuation supervision. Separate-stage training learns the encoder first and then fixes it while training the reasoner; joint training updates both together. Joint training improves the six-benchmark average by 12.7% relative to separate stages (Figure 4), including 8.8% on EMMA and 15.0% on RLBench. These results show that learning the representation and its predictor together outperforms fixing the representation before training the reasoner.

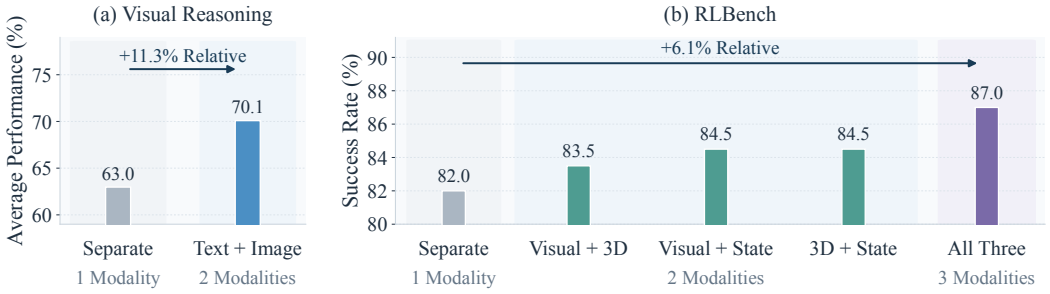

Figure 3: **Results of unifying more modalities.** **(a)** Qwen2.5-VL-7B mean accuracy on V*, MMVP, MathVista, and EMMA. **(b)** RLBench success with different modality groupings.

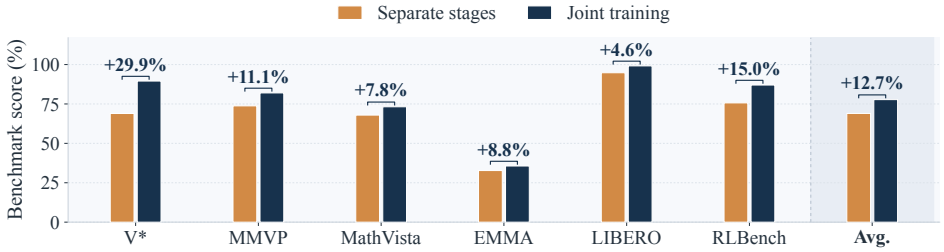

Figure 4: Separate-stage vs. joint training of the unified encoder and diffusion reasoner.

Unified encoding. To test whether the unified encoder improves reasoning, we compare it with separate encoding under both reconstruction and continuation supervision. In separate encoding, each modality has its own LLM encoder with the same architecture, whereas unified encoding shares one LLM encoder across all modalities; all other settings are identical. Unified encoding improves the six-benchmark average by 3.5 percentage points under reconstruction and 6.1 under continuation (Figure 5). These results show that unified encoding improves the average under both losses.

Latent target learning. To test whether learning the latent targets helps reasoning, we compare frozen teacher targets with targets learned by reconstruction or continuation. Reconstruction preserves source features, while continuation predicts later reasoning steps; both learned variants keep the same final-output supervision. With separate encoding, reconstruction and continuation improve the six-benchmark average over frozen targets by 3.7 and 10.2 percentage points (Figure 5), and continuation outperforms reconstruction by 6.6 points with separate encoding and 9.2 with unified encoding. These results show that learning the latent targets helps, particularly by predicting later reasoning steps rather than reconstructing source features.

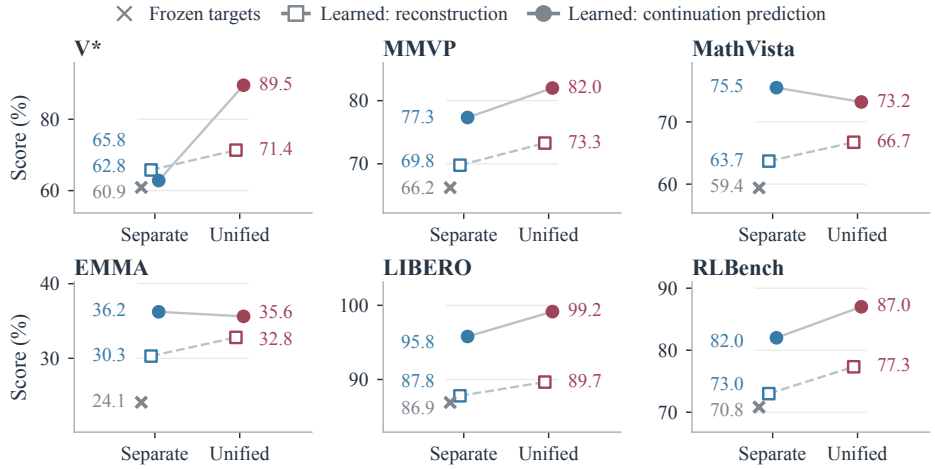

Figure 5: Comparison of separate and unified encoding with frozen targets or targets learned by reconstruction or continuation.

Diffusion objective. Finally, we test whether diffusion improves thought-token prediction, keeping joint training of the encoder and reasoner fixed across objectives. We compare flow match-

ing with direct thought-token prediction trained using squared L2 or cosine similarity loss (Appendix B.3). Flow matching improves the six-benchmark average by 16.3% relative to L2 and 21.0% relative to cosine similarity loss (Figure 6). These results show that diffusion-based thought-token prediction is an effective alternative to direct regression for downstream reasoning and manipulation.

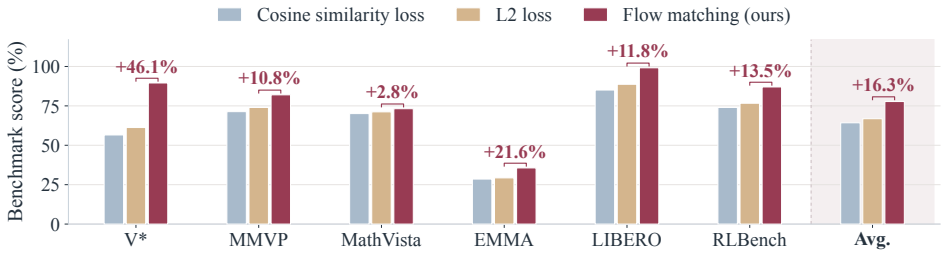

Figure 6: Comparison between different latent generation objectives.

5 Related Work

Multimodal reasoning. Multimodal chain-of-thought methods supervise or elicit textual rationales from visual evidence (Kang et al., 2025a; Lu et al., 2022b; Zhang et al., 2024; Rose et al., 2023; Xu et al., 2024; Dong et al., 2024; Guo et al., 2024; Yao et al., 2024; Huang et al., 2025b), or externalize computation through programs, tools, image operations, spatial abstractions, and visual scratchpads (Gupta & Kembhavi, 2022; Suris et al., 2023; Yang et al., 2023; Wu et al., 2023a; Hu et al., 2024b; Chen et al., 2024; Hu et al., 2024a; Zhou et al., 2024b). Recent work interleaves language with visual thoughts (Gao et al., 2024; Li et al., 2025a; Gu et al., 2025; Li et al., 2026b), and embodied methods guide control with language plans, visual predictions, and state or action trajectories (Ahn et al., 2022; Huang et al., 2022; Driess et al., 2023; Brohan et al., 2023; Jiang et al., 2022; Li et al., 2024b; Kim et al., 2024; Li et al., 2024a; Zhao et al., 2025; Zhong et al., 2026; Bai et al., 2026; Liu et al., 2026).

Latent reasoning. Latent methods internalize rationales or allocate extra computation without decoding every step (Zelikman et al., 2022; Goyal et al., 2024; Deng et al., 2023; 2024; Pfau et al., 2024; Zelikman et al., 2024; Yu et al., 2024; Wang et al., 2024a), compute through recurrence or continuous semantic states (Dehghani et al., 2019; Giannou et al., 2023; Yang et al., 2024; Kohli et al., 2026; Jeddi et al., 2026; Yang et al., 2026; Tu et al., 2026), or replace language rationales with continuous thoughts (Hao et al., 2024; Shen et al., 2025; Xu et al., 2025c;d; Deng et al., 2025b). Multimodal variants build latent thoughts from visual features, perceptual teachers, helper images, or task loss (Kang et al., 2026a; Pham & Ngo, 2025; Li et al., 2025b; Yang et al., 2025; Qin et al., 2025b; Wang et al., 2026; Jeon et al., 2026; Viveiros et al., 2026a; Li et al., 2026a; Xu et al., 2026a; Ray et al., 2025; Hu et al., 2026; Chen et al., 2026a; Shao et al., 2026; Chen et al., 2026b), and diffusion-based approaches model multimodal or multi-step latent trajectories (Kang et al., 2025b; Tu et al., 2026; He et al., 2023; Huang et al., 2025a; Bai et al., 2026; Liu et al., 2026; Viveiros et al., 2026b; Wu et al., 2026; Fan et al., 2026; Kang et al., 2026b). Mull-Tokens (Ray et al., 2025) is the closest to our approach: although its latent tokens are shared across modalities, each token is still supervised toward the target of its corresponding step—either text tokens or frozen image-encoder features. LaST_0 (Liu et al., 2026) is the closest VLA counterpart but uses a separate latent format for each modality. Uni-LaDiR instead learns a shared latent space through continuation prediction, rather than matching each step to text tokens or frozen modality-specific features. See more related work in Appendix C.

6 Conclusion

We introduced Uni-LaDiR, a unified framework for multimodal reasoning that represents all intermediate reasoning steps in a shared latent space. Its shared latent interface allows the reasoning model to predict successive steps in a common representation space. Uni-LaDiR outperforms the strongest prior methods by 7.3% on four mathematical and logical VLM benchmarks and 6.1% on RLBench. Notably, the advantage becomes larger as more modalities are incorporated into the shared space. These findings highlight the value of latent-space unification for enabling more effective multimodal reasoning.

AI USE STATEMENT

For the writing of this paper, we used AI tools only to edit and polish the language, including grammar, wording, and readability. The authors reviewed the edits and take responsibility for the final manuscript.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 2022.

Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, et al. 4M-21: An any-to-any vision model for tens of tasks and modalities. *arXiv preprint arXiv:2406.09406*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Shuanghao Bai, Jing Lyu, Wanqi Zhou, Zhe Li, Dakai Wang, Lei Xing, Xiaoguang Zhao, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Badong Chen, and Shanghang Zhang. Latent reasoning VLA: Latent thinking and prediction for vision-language-action models. *arXiv preprint arXiv:2602.01166*, 2026.

Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Jun Cen et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Boyuan Chen et al. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024.

Chao Chen, Zhixin Ma, Yongqi Li, Yupeng Hu, Yinwei Wei, Wenjie Li, and Liqiang Nie. Reasoning in the dark: Interleaved vision-text reasoning in latent space. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026a.

Hao Chen, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Renrui Zhang, Xiaoqi Li, Xiao He, Yandong Guo, Chi-Wing Fu, Shanghang Zhang, and Pheng-Ann Heng. Fast-in-slow: A dual-system vla model unifying fast manipulation within slow reasoning. In *Advances in Neural Information Processing Systems*, 2025a.

Hao Chen, Jiaming Liu, Zhonghao Yan, Nuowei Han, Renrui Zhang, Chenyang Gu, Jialin Gao, Ziyu Guo, Siyuan Qian, Yinxi Wang, Peng Jia, Shanghang Zhang, and Pheng-Ann Heng. LaST-R1: Reinforcing robotic manipulation via adaptive physical latent reasoning. *arXiv preprint arXiv:2604.28192*, 2026b.

Xi Chen, Xiao Wang, Soravit Changpinyo, et al. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Xi Chen et al. PaLI-X: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL `https://www.roboticsproceedings.org/rss19/p026.html`.

Yingqian Cui, Zhenwei Dai, Bing He, Zhan Shi, Hui Liu, Rui Sun, Zhiji Liu, Yue Xing, Jiliang Tang, and Benoit Dumoulin. How do latent reasoning methods perform under weak and strong supervision? *arXiv preprint arXiv:2602.22441*, 2026.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, et al. Universal transformers. In *International Conference on Learning Representations*, 2019.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025a.

Jingcheng Deng, Liang Pang, Zihao Wei, Shicheng Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. LLM latent reasoning as chain of superposition. *arXiv preprint arXiv:2510.15522*, 2025b.

Yuntian Deng et al. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Yuntian Deng et al. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved latent visual reasoning with selective perceptual modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29316–29335, 2026.

Yuhao Dong et al. Insight-V: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, et al. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023.

Shivam Duggal, Xingjian Bai, Zongze Wu, Richard Zhang, Eli Shechtman, Antonio Torralba, Phillip Isola, and William T. Freeman. End-to-end training for unified tokenization and latent denoising. *arXiv preprint arXiv:2603.22283*, 2026. URL `https://arxiv.org/abs/2603.22283`.

Ying Fan et al. Bridging the gap between latent and explicit reasoning with looped transformers. *arXiv preprint arXiv:2606.31779*, 2026.

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*, 2024.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. *arXiv preprint arXiv:2411.19488*, 2024.

Yuying Ge, Yixiao Ge, Ziyun Zeng, et al. Planting a SEED of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

Angeliki Giannou, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, et al. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, 2024.

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.

Jarvis Guo et al. MAmmoTH-VL: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *arXiv preprint arXiv:2211.11559*, 2022.

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.

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. In *International Conference on Machine Learning*, 2025.

Liqi He et al. Multi-modal latent space learning for chain-of-thought reasoning in language models. *arXiv preprint arXiv:2312.08762*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.11239.

Lianyu Hu, Shengqian Qin, Zeqin Liao, Qing Guo, Liang Wan, Wei Feng, and Yang Liu. CoLT: Teaching multi-modal models to think with chain of latent thoughts. *arXiv preprint arXiv:2606.31986*, 2026.

Yushi Hu, Weijia Shi, Xingyu Fu, et al. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, 2024a.

Yushi Hu et al. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Chi-Pin Huang et al. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. In *Advances in Neural Information Processing Systems*, 2025a.

Shaohan Huang, Li Dong, Wenhui Wang, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

Wenlong Huang, Fei Xia, Ted Xiao, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, 2022.

Wenxuan Huang et al. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Ahmadreza Jeddi et al. LoopFormer: Elastic-depth looped transformers for latent reasoning via shortcut modulation. *arXiv preprint arXiv:2602.11451*, 2026.

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *arXiv preprint arXiv:2602.04476*, 2026.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, et al. VIMA: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.

Yang Jin, Kun Xu, Kun Xu, et al. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *International Conference on Learning Representations*, 2024.

Haoqiang Kang, Enna Sachdeva, Piyush Gupta, Sangjae Bae, and Kwonjoon Lee. Gflowvlm: Enhancing multi-step reasoning in vision-language models with generative flow networks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3815–3825. IEEE, 2025a.

Haoqiang Kang, Yizhe Zhang, Nikki Lijing Kuang, Nicklas Majamaki, Navdeep Jaitly, Yi-An Ma, and Lianhui Qin. LaDiR: Latent diffusion enhances LLMs for text reasoning. *arXiv preprint arXiv:2510.04573*, 2025b.

Haoqiang Kang, Yinpeng Chen, Luyang Liu, Jesper Sparre Andersen, Abhijit Ogale, Baochen Sun, Lichan Hong, and Ed H Chi. Scaffolding minds: Optimizing latent visual target representations for multimodal reasoning. *arXiv preprint arXiv:2608.19669*, 2026a.

Haoqiang Kang, Yizhe Zhang, Nikki Lijing Kuang, Yi-An Ma, and Lianhui Qin. Ladi-rl: Latent diffusion reasoning prevents entropy collapse in reinforcement learning. *arXiv preprint arXiv:2602.01705*, 2026b.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Harsh Kohli et al. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers. *arXiv preprint arXiv:2604.07822*, 2026.

Dianqiao Lei and Lianlei Shan. Think less, act early: Reinforced latent reasoning with early exit in vision-language-action models. *arXiv preprint arXiv:2606.15099*, 2026.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025a.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025b.

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*, 2025c.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33457–33466, 2026a.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Zesheng Yang, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026b.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a.

Wei Li, Jizhihui Liu, Li Yixing, Junwen Tong, Rui Shao, and Liqiang Nie. Consisvla-4d: Advancing spatiotemporal consistency in efficient 3d-perception and 4d-reasoning for robotic manipulation. *arXiv preprint arXiv:2605.05126*, 2026c.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*, 2023.

Xinghang Li, Minghuan Liu, Hanbo Zhang, et al. Vision-language foundation models as effective robot imitators. In *International Conference on Learning Representations*, 2024b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

Zhuoyang Liu, Jiaming Liu, Hao Chen, Jiale Yu, Ziyu Guo, Chengkai Hou, Chenyang Gu, Xiangju Mi, Renrui Zhang, Kun Wu, Zhengping Che, Jian Tang, Pheng-Ann Heng, and Shanghang Zhang. LaST_0: Latent spatio-temporal chain-of-thought for robotic vision-language-action model. *arXiv preprint arXiv:2601.05248*, 2026.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022a.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 2022b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

David Mizrahi, Roman Bachmann, Oguzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. *arXiv preprint arXiv:2312.06647*, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, et al. Kosmos-2: Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, 2024.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-CoT: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025a.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. Chain-of-visual-thought: Teaching VLMs to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025b.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, Jiayuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A. Plummer, Kate Saenko, Ranjay Krishna, Leonidas Guibas, and Wen-Sheng Chu. Mull-tokens: Modality-agnostic latent thinking. *arXiv preprint arXiv:2512.10941*, 2025.

Scott Reed, Konrad Zolna, Emilio Parisotto, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.10752.

Daniel Rose et al. Visual chain of thought: Bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023.

Yifei Shao, Kun Zhou, Ziming Xu, Mohammad Atif Quamar, Shibo Hao, Zhen Wang, Zhiting Hu, and Biwei Huang. Learning modal-mixed chain-of-thought reasoning with latent embeddings. *arXiv preprint arXiv:2602.00574*, 2026.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

Quan Sun, Yufeng Cui, Xiaosong Zhang, et al. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

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 *International Conference on Learning Representations*, 2024b.

Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*, 2024c.

Didac Suris, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Guancheng Tu, Xiangjun Fu, Suhao Yu, Yao Tang, Haoqiang Kang, Lianhui Qin, Yizhe Zhang, and Jiatao Gu. Latent reasoning with normalizing flows. *arXiv preprint arXiv:2606.06447*, 2026.

André G. Viveiros, Nuno Gonçalves, Matthias Lindemann, and André Martins. LanteRn: Latent visual structured reasoning. *arXiv preprint arXiv:2603.25629*, 2026a.

André G. Viveiros et al. What's holding back latent visual reasoning? *arXiv preprint arXiv:2605.18445*, 2026b.

Boshi Wang et al. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. *arXiv preprint arXiv:2405.15071*, 2024a.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b.

Peng Wang et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 2022.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, Chunrui Han, Yuang Peng, Qi Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, and Vishal M. Patel. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *arXiv preprint arXiv:2507.05255*, 2025.

Chenfei Wu, Shengming Yin, Weizhen Qi, et al. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023b.

Yueh-Hua Wu et al. Continuous reasoning for vision-language-action. *arXiv preprint arXiv:2606.00229*, 2026.

Jinheng Xie, Weijia Mao, Zechen Bai, et al. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Guowei Xu et al. LLaVA-CoT: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Jin Xu et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Tianrun Xu, Yue Sun, Qixun Wang, Jingyi Lu, Yuan Wang, Tianren Zhang, Longteng Guo, Fengyun Rao, Jing Lyu, Feng Chen, and Jing Liu. Semantic-enriched latent visual reasoning. *arXiv preprint arXiv:2605.19342*, 2026a.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025b.

Xiaoxu Xu, Hao Li, Jinhui Ye, Yilun Chen, Jia Zeng, Xinyi Chen, Linning Xu, Dahua Lin, Weixin Li, and Jiangmiao Pang. FutureVLA: Joint visuomotor prediction for vision-language-action model. *arXiv preprint arXiv:2603.10712*, 2026b.

Yige Xu et al. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025c.

Yige Xu et al. Softcot++: Test-time scaling with soft chain-of-thought reasoning. *arXiv preprint arXiv:2505.11484*, 2025d.

Bochen Yang and Lianlei Shan. Pearlvla: Progressive embodied action-plan refinement in latent space. *arXiv preprint arXiv:2606.17924*, 2026.

Liu Yang et al. Looped transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024.

Xiao-Wen Yang et al. Stabilizing recurrent dynamics for test-time scalable latent reasoning in looped language models. *arXiv preprint arXiv:2605.26733*, 2026.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Huanjin Yao et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning: Training LLMs for divergent reasoning with minimal examples. In *Forty-second International Conference on Machine Learning*, 2025.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023.

Ping Yu et al. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 2022.

Eric Zelikman, Georges Harik, Yijia Shao, et al. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

Jun Zhan, Junqi Dai, Jiasheng Ye, et al. AnyGPT: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

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.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing thinking with images via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Tianran Zhang, Wenxuan Song, Jiayi Chen, Xinhu Zheng, Hesheng Wang, and Haoang Li. FlowVLA: Visual chain of thought-based motion reasoning for vision-language-action models. *arXiv preprint arXiv:2508.18269*, 2025.

Zhide Zhong, Junfeng Li, Junjie He, Haodong Yan, Xin Gong, Guanyi Zhao, Yingjie Cai, Jiantao Gao, Xu Yan, Bingbing Liu, Yingcong Chen, Liuqing Yang, and Haoang Li. DualCoT-VLA: Visual-linguistic chain of thought via parallel reasoning for vision-language-action models. *arXiv preprint arXiv:2603.22280*, 2026.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024a.

Qiji Zhou et al. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024b.

A Method Details

A.1 Attention masks for parallel block training

Figure 7 illustrates the three masks for K reasoning steps packed into each pass; ellipses indicate omitted intermediate blocks. The restrictions apply at every layer, preventing masked information from reaching predictions through other hidden states.

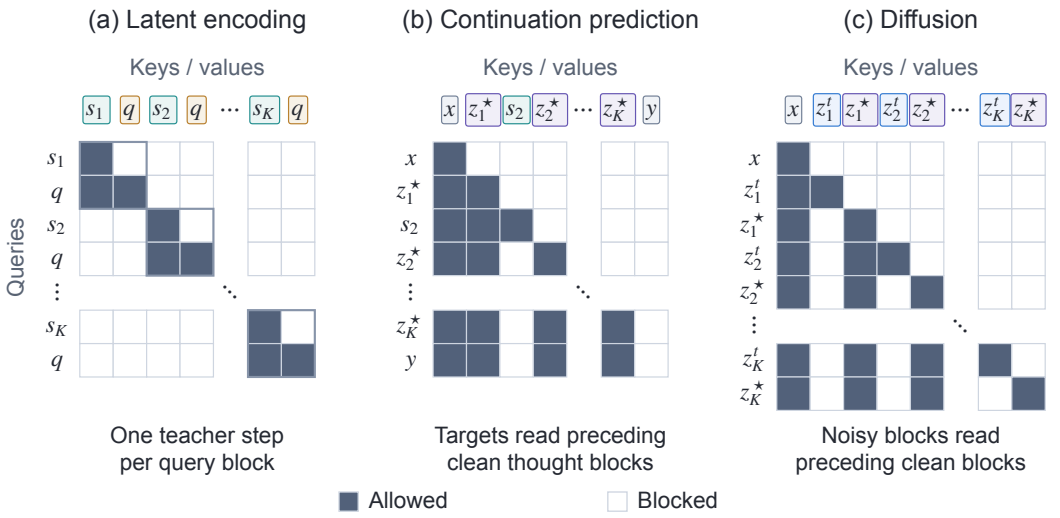

Figure 7: **Attention masks for parallel block training.** Rows attend to columns; each cell represents a token block. Continuation prediction interleaves clean thought blocks with target streams; diffusion interleaves noisy and clean thought blocks. Clean blocks never read target or noisy streams.

Latent encoding. The packed sequence contains pairs $[s_i; q]$ for all teacher steps. Attention is isolated within each pair: the learnable block reads its own teacher features and itself, but neither the task input nor other pairs.

Continuation prediction. The packed sequence is $[x; z_1^\star; s_2; z_2^\star; s_3; \ldots; z_K^\star; y]$, where s_i and y label target prediction streams. A prediction of s_i^{GT} can read x and $z_{<i}^\star$; the final-output prediction can read x and all $z_{1:K}^\star$. The conditioning stream preserves this prefix order and cannot read the target streams. Different target streams are isolated from one another. Within a text target, standard causal attention permits earlier tokens of that target; continuous predictions cannot read their ground-truth values.

Diffusion. The packed sequence is $[x; z_1^t; z_1^\star; z_2^t; z_2^\star; \ldots; z_K^t; z_K^\star]$. Each noisy block z_i^t can read x, the clean prefix $z_{<i}^\star$, and all tokens within itself. It cannot read its own clean target, later clean blocks, or other noisy blocks. The clean conditioning stream follows block-causal attention and cannot read noisy blocks.

B Experimental Details

B.1 Evaluation details

Benchmarks and scoring. All methods in Tables 1 and 6 use the same evaluation examples within each benchmark. Table 4 lists the evaluation splits, sample counts, and scoring units for the eleven visual benchmarks (Song et al., 2025; Masry et al., 2022; Wu & Xie, 2024; Fu et al., 2024; Tong et al., 2024b; Ray et al., 2024; Tong et al., 2024a; Lu et al., 2024; Wang et al., 2024b; Xu et al., 2025b; Hao et al., 2025). VisualPuzzles and VisuLogic use their released evaluation collections, even when the hosting service names the split `train`. VisuLogic evaluation uses the original 1,000-question benchmark (Xu et al., 2025b), rather than the current 1,003-row hosted collection.

Comparison protocol. On RLBench, we follow the LaST_0 protocol and take the baseline results from that paper (Liu et al., 2026). There, each baseline starts from its released checkpoint, is fine-tuned on the same 100 demonstrations per task, and is evaluated with the same rollouts; CoT-VLA is reimplemented on Janus-Pro. Uni-LaDiR uses the same backbone, demonstrations, keyframe sampling, training schedule, and evaluation as LaST_0, so the two differ mainly in how latent reasoning is designed. On LIBERO, all methods use the standard training demonstrations and evaluation protocol of the benchmark. On the VLM side, every method in Table 1 uses Qwen2.5-VL-7B and is scored on the same examples. Each baseline is trained with its own data and recipe, so Table 1 compares complete methods rather than isolating the latent design. The analyses in Sections 4.3 and 4.4 keep the backbone, training data, and teacher supervision fixed.

Table 4: **Shared visual evaluation sets.** Splits, example counts, and scoring units used for the benchmark comparisons.

Benchmark	Split	Size	Scoring unit
VisualPuzzles	Released collection	1,168	Question accuracy
ChartQA	Test	2,500	Relaxed answer accuracy
V^*	Test	191	Question accuracy
BLINK-Jigsaw	Validation, Jigsaw	150	Question accuracy
MMVP	Full	300	Question accuracy
SAT	Real-image test	150	Question accuracy
CV-Bench	Test	2,638	Mean of 2D and 3D scores
MathVista	Testmini	1,000	Answer accuracy
MathVision	Test	3,040	Answer accuracy
VisuLogic	Original benchmark	1,000	Question accuracy
EMMA	Test, four subjects	2,788	Question accuracy

B.2 Training and inference details

VLM training. We train on Zebra-CoT (Li et al., 2025a), using its interleaved text and image steps as teacher supervision. We use AdamW with a peak learning rate of 10^{-5}, cosine decay, a global batch size of 64, and one epoch of training, informed by prior multimodal fine-tuning recipes (Li et al., 2025a; Hu et al., 2026). We use bf16, a maximum sequence length of 8,192, 3% warmup, weight decay of 0.01, and gradient clipping at 1.0. Each backbone retains its native image processor. The unified encoder and shared reasoning backbone are optimized jointly; teacher traces are omitted at evaluation.

Thought tokens and latent sampling. We use four 512-dimensional thought tokens per block, with learned projections to and from the backbone width, taking the token budget from a latent-diffusion reference configuration (Kang et al., 2025b). The loss weights are $\lambda_y = 2$ and $\lambda = 5$. Training retains the annotated reasoning-step sequence, and the number of thought blocks is not fixed. At inference, the model decides when to stop: after each block, it predicts the next token, and if this is the special token <BOT>, it generates another block; otherwise, the latent chain ends. For VLMs, the chain is followed by greedy answer decoding with a 512-token limit.

For flow matching, we draw $\epsilon_i \sim \mathcal{N}(0, I)$ for each block and $t \sim \mathcal{U}[0, 1]$, set $z_i^t = (1-t)\epsilon_i + t\,\text{sg}(z_i^\star)$, and use $v_i^\star = z_i^\star - \epsilon_i$ as the velocity target (Lipman et al., 2023). At inference, we integrate from $t = 0$ to 1 per block with one sampled trajectory and no classifier-free guidance. Table 5 collects these settings; VLM and VLA training use the same loss weights.

RLBench. We evaluate ten tasks with a Franka Panda in CoppeliaSim, using 100 waypoint-generated training demonstrations per task and keyframe sampling (Liu et al., 2026). Inputs include a front-view RGB image resized to 384×384, a point cloud sampled to 1,024 points, the task instruction, and proprioception. We evaluate the final checkpoint with 20 rollouts per task and seed, giving 600 rollouts across three seeds.

LIBERO. We train a separate policy for each of the four suites, each containing ten tasks and 500 training demonstrations. Inputs use third-person and wrist RGB views resized to 384×384. We omit the point-cloud teacher because this modality is unavailable, and sample intermediate teacher

Table 5: **Main hyperparameters.** Settings shared by both domains span the two columns.

Setting	VLM	VLA
Backbone	Qwen2.5-VL-7B	MoT, 2× Janus-Pro-1.5B (3.3B)
Training data	Zebra-CoT	RLBench 100 demos/task; LIBERO 500 demos/suite
Teacher modalities	Text, image	Future image, point cloud (RLBench), robot state
Thought tokens per block	4, each 512-dimensional	
Loss weights	$\lambda_y = 2$, $\lambda = 5$	
Latent sampler	Flow matching; one trajectory per block; no CFG	
Output decoding	Greedy, $\leq$512 tokens	Flow-matching head, 10 steps, 7-D actions
Reasoning-to-action ratio	–	1:1, 1:2, 1:4 (train); 1:4 (eval)
Optimizer	AdamW	AdamW
Peak learning rate	10^{-5}, cosine decay	10^{-4}, cosine decay to zero
Warmup / weight decay	3% / 0.01	none / none
Global batch size	64	64
Training length	1 epoch	300 epochs
Precision / gradient clipping	bf16 / 1.0	
Max. sequence length	8,192	–

steps every eight frames (Liu et al., 2026). We evaluate the final checkpoint with 50 rollouts per task and seed, giving 6,000 rollouts across the four suites and three seeds. Both benchmarks use their native success conditions.

Seeds and aggregation. Our experiments use three random seeds. For question-level accuracy or rollout success on N examples per seed, the mean is $\bar{s} = \frac{100}{3N} \sum_{r=1}^{3} \sum_{j=1}^{N} c_{rj}$, where $c_{rj} \in \{0, 1\}$ indicates correctness or success. Grouped metrics retain the benchmark's aggregation rule: CV-Bench averages its 2D and 3D scores. For MMVP, we report question-level accuracy over 300 questions, matching the VLMEvalKit protocol used by the relevant comparison methods (Gu et al., 2025; Li et al., 2026b). Each answer is scored independently against its correct option; the reported result is the mean of the three seed-level accuracies. Benchmark averages give equal weight to each reported benchmark or suite. Relative gains are $100(s_{\text{ours}}/s_{\text{baseline}} - 1)$; rounding follows aggregation, and gains are computed before table entries are rounded to one decimal. RLBench Table 2 reports the standard deviation across seeds in percentage points, following the baseline paper (Liu et al., 2026). Reported gains describe differences in mean scores.

VLA training. We train for 300 epochs with AdamW, using a peak learning rate of 10^{-4}, cosine decay to zero, no weight decay or warmup, and gradient clipping at 1.0 (Liu et al., 2026). Training uses eight NVIDIA A800 GPUs with bf16 and DeepSpeed ZeRO-1. The batch size is eight per device and 64 globally, without gradient accumulation. Actions are seven-dimensional, with ten flow-matching steps in the action head. We mix reasoning-to-action ratios of 1:1, 1:2, and 1:4 during training and use 1:4 at evaluation. The latent diffusion objective is defined in Section 3.

Inference speed. All inference rates in Table 2, including those of the baselines, are measured on the same single NVIDIA RTX 4090, without action chunking, using one reasoning update per four action decisions (Liu et al., 2026). We use batch size one and bf16, with 100 warmup decisions followed by 1,000 timed decisions per seed. Wall-clock timing is synchronized with CUDA before and after the timed sequence and includes observation preprocessing, host-to-device transfer, reasoning updates, and action prediction; simulator stepping and rendering are excluded. Each method uses the same inputs and timing boundary. Throughput is the number of decisions divided by elapsed time, averaged across three seeds. The reported 16.5 Hz corresponds to 60.6 ms per decision.

B.3 Ablation protocols

Latent target supervision and encoding paths. The extended design in Figure 5 crosses two intermediate objectives with whether the LLM encoder is shared across modalities. In the *separate* path, each modality is encoded by its own LLM encoder, which has the same architecture, learnable queries, and output projection as the unified encoder. Recon and Cont train these encoders jointly with the reasoner. The Frozen (separate) control keeps each modality's encoder frozen, so the latent

targets stay fixed, and trains only the downstream reasoner. The Separate baseline in Figure 3 is this same continuation-trained separate variant; all shared scores use identical source values. The *unified* path uses the shared queries and LLM encoding pass described in Section 3.2. All five variants use the same number and width of thought tokens. The learned variants share data, training steps, diffusion loss, and $\lambda_y \mathcal{L}_{\text{out}}$. Reconstruction predicts source teacher features, whereas continuation predicts later teacher steps; text CE and continuous L2 use the same reductions and weights across paths. Target features remain fixed. Because the separate path uses one encoder per modality, it has at least as many encoder parameters as the unified path; the two paths differ only in whether the encoder is shared across modalities.

Aggregation and rounding. Figure 5 uses the same three-seed aggregation as the main tables: $N = 191$, 300, 1,000, and 2,788 questions per seed for V*, MMVP, MathVista, and EMMA; $N = 2{,}000$ and 200 rollouts per seed for LIBERO and RLBench. Scores are rounded to one decimal after aggregation; cross-benchmark differences use unrounded scores and weight benchmarks equally.

Thought-token prediction. In Figure 6, all three objectives are trained jointly with the unified encoder, using the same continuation loss and loss weights; only the thought-token prediction objective changes. L2 and cosine regression directly predict the next clean thought block. Flow matching predicts the velocity along the noisy-to-clean path, using the sampler specified in Appendix B.1. Each method conditions on the same input and preceding thought blocks and supplies its predicted tokens to the downstream reasoner. This comparison reports downstream performance for the respective prediction procedures.

B.4 Additional experiments

Reasoning across backbones. We evaluate Uni-LaDiR with BAGEL-7B, Qwen3-VL-8B, and Janus-Pro at 1.5B and 7B to test whether the gains extend beyond the main backbone (Table 6). On the seven vision-centric benchmarks, the average improves over the corresponding base models by 29.5%, 12.4%, 35.9%, and 51.5%, respectively. With BAGEL, Uni-LaDiR also improves the average over Visual-OPSD by 2.5%, leading on six of seven benchmarks; with Qwen3-VL, it improves over CoLT by 13.3%. On BAGEL's mathematical and logical benchmarks, the average gains are 27.9% over the base model and 20.0% over Uni-CoT (Qin et al., 2025a), including a 28.3% gain on EMMA over Uni-CoT. These improvements across model families and sizes suggest that shared latent reasoning is useful beyond a single backbone.

Latent interventions. Prior work shows that latent reasoning methods can bypass intermediate latents without losing accuracy (Cui et al., 2026; Viveiros et al., 2026b). We therefore test whether Uni-LaDiR uses its generated thought tokens. We zero the latent chain, shuffle its blocks, or replace it with a chain from another example, while keeping the input and sequence length unchanged:

$$\widetilde{z}_{1:K} = \begin{cases} z_{1:K}, & \text{intact,} \\ 0, & \text{zeroed,} \\ [z_{\pi(1)}; \ldots; z_{\pi(K)}], & \text{shuffled,} \\ z^{(j)}_{1:K}, & \text{cross-example.} \end{cases} \tag{6}$$

Here π permutes the latent blocks while preserving their internal token order, and $z^{(j)}_{1:K}$ comes from another example with the same latent shape. We recompute the output with fresh hidden states and pair decoding or rollout seeds across conditions. For VLA evaluation, we apply the intervention at every reasoning update under the same initial scene.

Zeroing the latent chain reduces performance by 32.0–33.7 percentage points, while shuffling its blocks causes a 22.0–23.7 point drop (Table 7). Replacing the chain with one from another example has the largest effect, decreasing performance by 36.0–38.7 points. Together with the supervision ablation in Figure 5, these results show that continuation prediction learns useful latent representations: the generated thought tokens retain input-specific, order-sensitive information used for prediction.

Qualitative manipulation comparison. Figure 8 compares a LaST_0 rollout with a successful Uni-LaDiR rollout from the same initial scene. LaST_0 approaches the phone but does not secure a grasp within ten decisions. Uni-LaDiR grasps the phone, transfers it to the base, and releases it,

Table 6: **Reasoning across backbones.** Accuracy (%) on **(a)** seven vision-centric benchmarks and **(b)** four mathematical and logical benchmarks. Bold marks the best result within each backbone.

(a) Vision-centric reasoning

Method	VisPuz.	ChartQA	V*	BLINK-J	MMVP	SAT	CV-B	Avg.
BAGEL-7B								
Base	35.0	61.8	55.5	67.3	70.3	44.7	76.0	58.7
Visual-OPSD	**86.0**	78.8	64.9	77.3	77.3	54.0	80.6	74.1
ThinkMorph	79.0	78.1	67.0	72.0	80.3	52.7	80.8	72.8
Uni-LaDiR (Ours)	75.1	**83.8**	**69.1**	**80.0**	**83.0**	**54.7**	**86.4**	**76.0**
Qwen3-VL-8B								
Base	37.0	82.6	84.3	68.7	77.7	54.0	85.6	70.0
CoLT	39.0	74.7	82.2	70.0	78.0	58.0	84.0	69.4
Uni-LaDiR (Ours)	**60.3**	**86.1**	**85.9**	**78.0**	**84.0**	**67.3**	**89.1**	**78.7**
Janus-Pro-1.5B								
Base	28.0	52.0	41.9	48.0	58.0	24.0	64.0	45.1
Uni-LaDiR (Ours)	**58.7**	**68.2**	**57.1**	**64.0**	**69.3**	**38.7**	**73.1**	**61.3**
Janus-Pro-7B								
Base	33.5	43.1	38.2	50.7	63.3	22.0	67.8	45.5
Uni-LaDiR (Ours)	**69.8**	**75.2**	**64.4**	**70.0**	**76.3**	**46.7**	**80.3**	**68.9**

(b) Mathematical and logical reasoning: BAGEL-7B

Method	MathVista	MathVision	VisuLogic	EMMA	Avg.
Base	72.5	36.0	28.9	28.7	41.5
Uni-CoT	73.3	48.5	25.0	30.2	44.2
Bagel-Zebra-CoT	72.1	43.5	0.0	20.6	34.0
ThinkMorph	67.8	41.5	6.5	22.4	34.6
Uni-LaDiR (Ours)	**79.8**	**57.9**	**35.9**	**38.7**	**53.1**

Table 7: **Latent-intervention results.** We modify the generated latent chain while keeping the input and sequence length fixed. Δ denotes the change from the intact chain in percentage points.

	V*		RLBench	
Latent chain	Accuracy (%)	Δ	Success (%)	Δ
Intact generated	**89.5**	—	**87.0**	—
Zeroed	55.8	−33.7	55.0	−32.0
Shuffled blocks	65.8	−23.7	65.0	−22.0
Cross-example	50.8	−38.7	51.0	−36.0

reaching success at 7.05 seconds of simulation time. This selected episode illustrates a difference at the grasping stage; aggregate success rates are reported in Table 2.

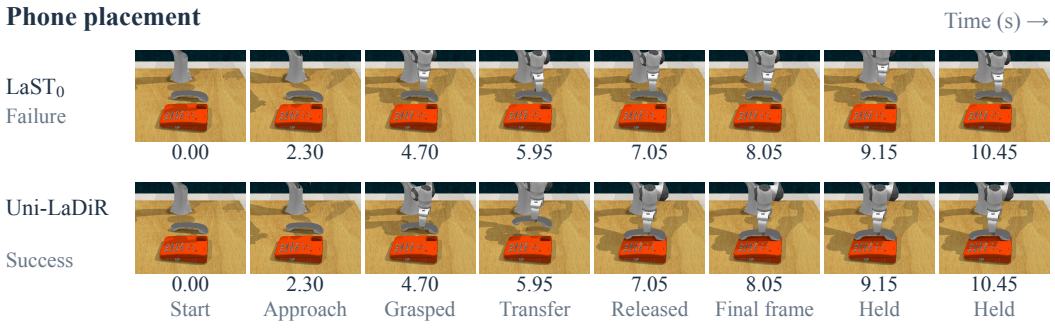

Figure 8: **Phone placement in RLBench.** LaST$_0$ does not grasp the phone within ten recorded decisions (top), while Uni-LaDiR completes placement (bottom). Columns share simulation timestamps, excluding inference latency; the final frame is held after each recording ends.

C Extended Related Work

Unified multimodal models. Early generalist models cast heterogeneous tasks as a shared sequence problem, spanning vision–language understanding, generation, and embodied control (Wang et al., 2022; Lu et al., 2022a; 2023; Reed et al., 2022; Mizrahi et al., 2023; Bachmann et al., 2024), and large-scale vision–language models connected pretrained language models to visual inputs through learned interfaces (Alayrac et al., 2022; Chen et al., 2022; 2023; Huang et al., 2023; Peng et al., 2024). Autoregressive any-to-any systems represent multiple modalities in a common token stream (Sun et al., 2024b;a; Yu et al., 2023; Ge et al., 2023; Jin et al., 2024; Wu et al., 2023b; Zhan et al., 2024; Chameleon Team, 2024; Wang et al., 2024c), while hybrid autoregressive–diffusion architectures unify discrete and continuous generation (Tang et al., 2023; Zhou et al., 2024a; Xie et al., 2024; Sun et al., 2024c; Wu et al., 2025; Chen et al., 2025b; Deng et al., 2025a; Xu et al., 2025a; Bai et al., 2025a). Whereas these models unify what is read and generated, Uni-LaDiR unifies the reasoning in between, representing and predicting the intermediate thoughts from every modality in a single latent space.

D Limitations

Our experiments focus on whether reasoning in a shared latent space helps, and we leave a closer study of the teacher CoTs themselves to future work. Beyond which modalities are shared (Section 4.3), we do not analyze how the teacher CoTs affect the learned reasoning, for example the granularity of teacher steps or the source and quality of the traces. In our setting, the VLA teachers are future images, point clouds, and robot states taken from the demonstrations, and the VLM teachers are the interleaved traces in Zebra-CoT; other sources, such as tactile signals or model-generated traces, may shape the latent space differently.