

中国科学技术大学
University of Science and Technology of China

# UniVLR

## Unifying Text and Vision in Visual Latent Reasoning for Multimodal LLMs

Houcheng Jiang

2026/11/06

# CONTENT


1 **Background:** *From Latent to Latent Visual Reasoning*

2 **Motivation:** *Vision as a Unified Source*

3 **Method:** *UniVLR*

4 **Results & Discussion**

# Background: From Latent to Latent Visual Reasoning

## 1. Latent Reasoning: A Viable Route beyond Verbose CoT

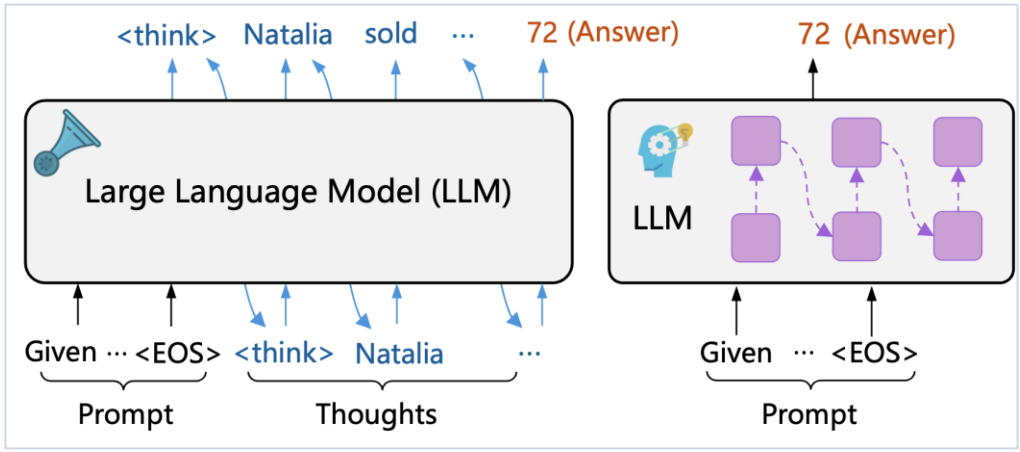
### Frontier signal: GPT-6 Astra (OpenAI, Sep 2026)

- Reportedly built on a **looped / recurrent-depth Transformer**, which adds computation in depth instead of in tokens
- **Fewer output tokens at matched accuracy** than GPT-5.6 (rumored, not officially confirmed)

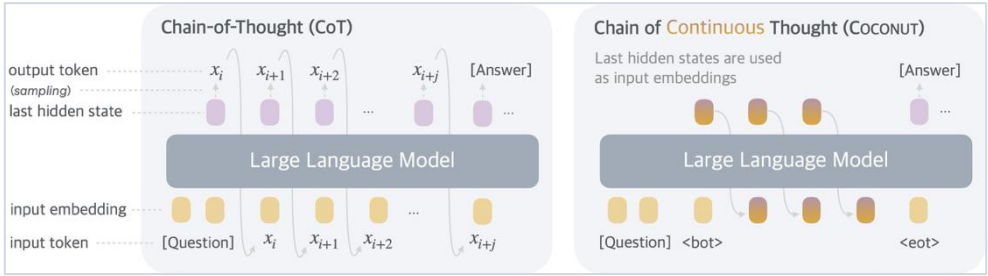
### Research evidence

- **Coconut**, **recurrent-depth LMs**, **Ouro / MoR / SMELT**: latent or looped compute boosts reasoning

**Outlook:** **latent reasoning** for the heavy lifting + **concise explicit text** for verifiable steps, a hybrid paradigm that cuts tokens without losing reasoning power.




Explicit CoT vs. latent CoT

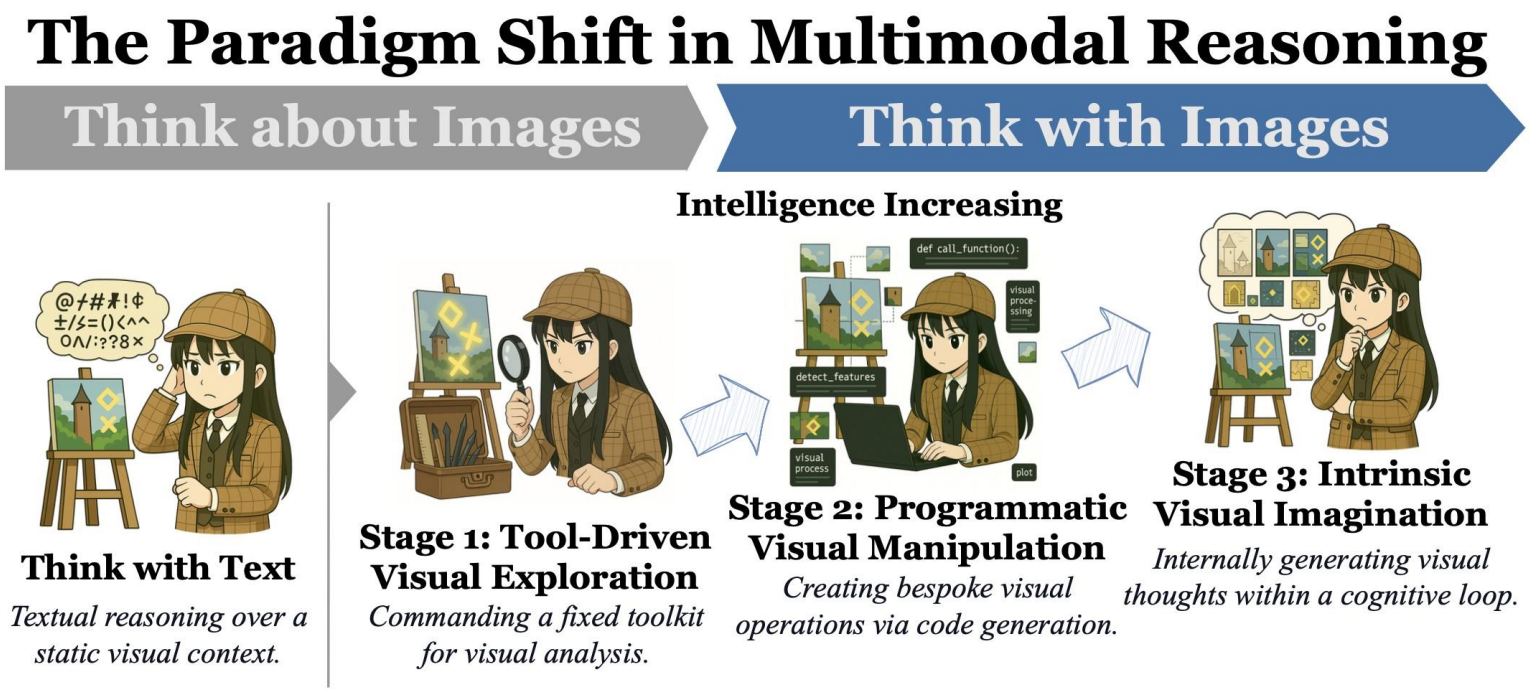



**Reasoning need not be spoken: think in continuous space, verbalize only what is needed.**

RASCHKA, GPT-6 ASTRA, LOOPED TRANSFORMERS, AND HIDDEN REASONING (SEP 2026); COCONUT (2024); GEIPING ET AL. (2025)

# Background: From Latent to Latent Visual Reasoning

## 2. Multimodal Frontier: Thinking with Images Goes Latent




**1 Thinking with images**

Perception-heavy tasks need visual evidence revisited during reasoning, not encoded once

**2 Tools are rigid**

Crop / zoom / sketch tools are predefined, slow (multi-turn) and unstable

**3 Go latent**

Generate **visual thoughts inside the model**: latent visual reasoning

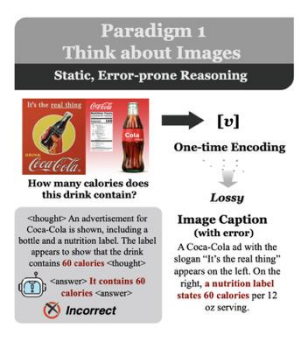
**Latent reasoning × thinking with images = latent visual reasoning**

SU ET AL., THINKING WITH IMAGES FOR MULTIMODAL REASONING: FOUNDATIONS, METHODS, AND FUTURE FRONTIERS, 2025

# Background: From Latent to Latent Visual Reasoning

## 3. Prior Paradigms at a Glance

### Textual CoT



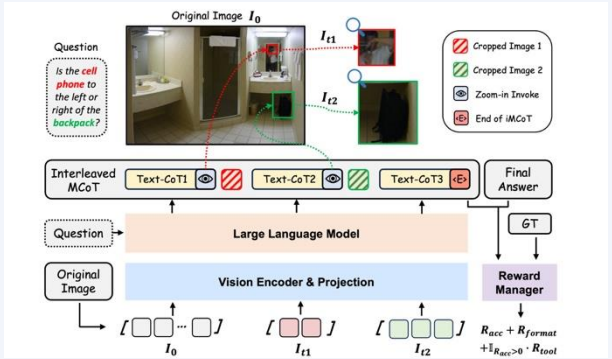

*Reason in language over a one-shot image encoding*

**e.g.** GPT-4o, Qwen2.5-VL, R1-style MLLMs

✓ Simple, readable rationale

✗ Lossy static perception; ~260 text tokens

### Tool-based Visual Reasoning



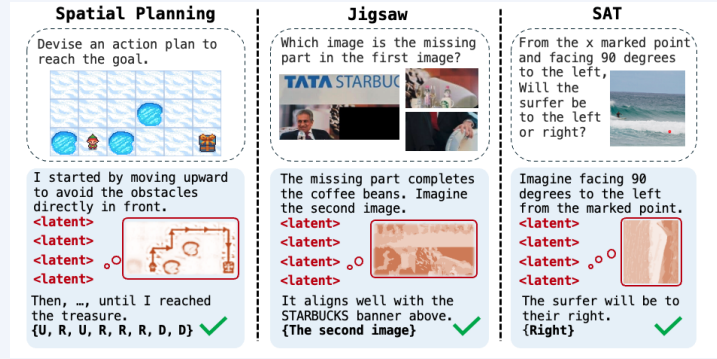

*Interleave text CoT with crops / zooms / sketches from tools*

**e.g.** DeepEyes, Pixel Reasoner, Visual Sketchpad

✓ Revisits fine-grained evidence

✗ Rigid tool set, multi-turn latency, unstable auxiliary images

### Visual Latent Reasoning




*Generate continuous visual latents aligned to auxiliary-image embeddings*

**e.g.** Mirage, LVR, Monet, SkiLa, CoVT

✓ Tool-free internal imagination

✗ Text CoT and visual latents stay two separate channels

**Common gap: text and vision remain two separate sources, with no unified latent space.**

GPT-4O; QWEN2.5-VL; DEEPEYES; PIXEL REASONER; VISUAL SKETCHPAD; MIRAGE; LVR; MONET; SKILA; CoVT (2024–2025)

# CONTENT


1 **Background:** *From Latent to Latent Visual Reasoning*

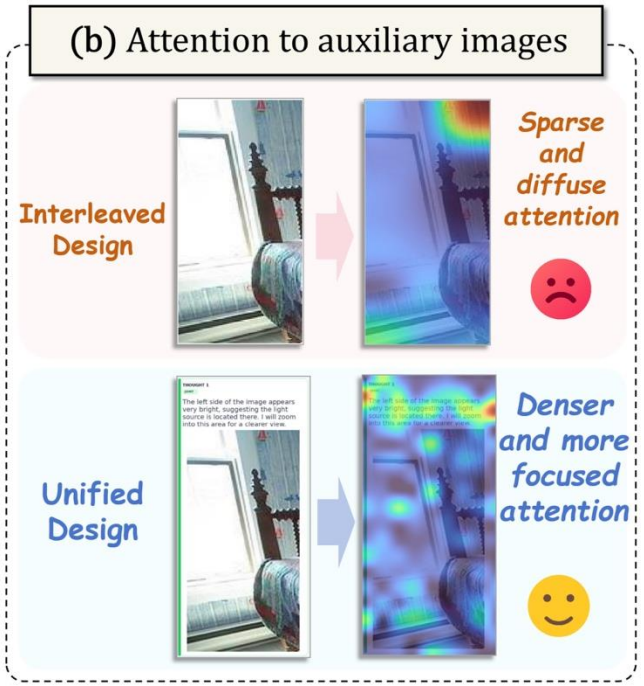
2 **Motivation:** *Vision as a Unified Source*

3 **Method:** *UniVLR*

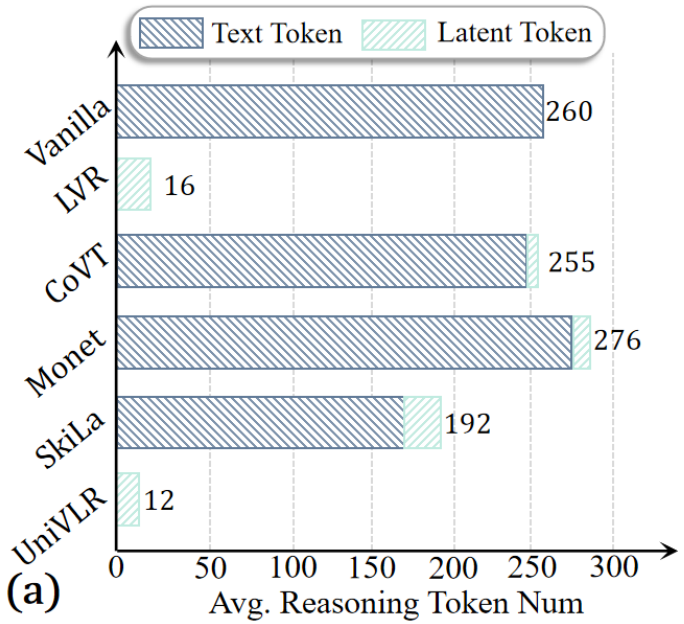
4 **Results & Discussion**

# Motivation: Vision as a Unified Source

## 1. What Goes Wrong in the Interleaved Design?




(a) Attention of later CoT tokens to the auxiliary image




(b) Generated reasoning tokens per sample

### 1 Limited efficiency

Interleaved methods still emit **190–276 tokens** per sample, mostly explicit text, so latents add little compression.

### 2 Fragmented information

Text CoT and visual latents are heterogeneous. Later CoT tokens show **sparse, diffuse attention** on auxiliary images, and the model falls back on text.

**Root cause: latent reasoning lacks a unified information source.**

UNIVLR (OURS), FIGURE 1(b) & FIGURE 3(a)

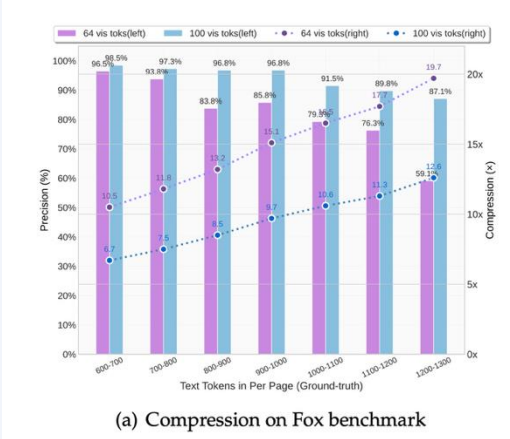
# Motivation: Vision as a Unified Source

## 2. Core Insight: Vision as the Unified Information Source

### Human cognition

- Vision is our **dominant channel** for taking in information, and even reading is a visual act
- In hard problems we lay out **notes, sketches and diagrams on one shared workspace** instead of switching channels
- The **visual field is a natural canvas** where text and images coexist

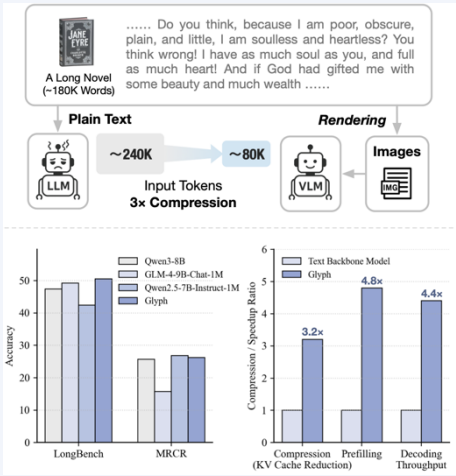
### MLLMs can already read pixels



(a) Compression on Fox benchmark

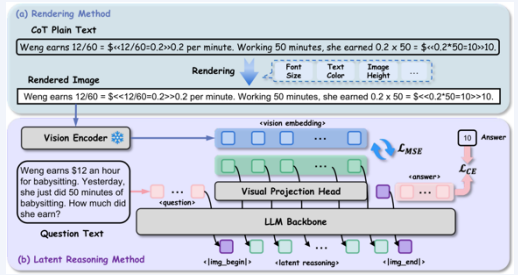
**DeepSeek-OCR**

<10× optical compression, **~97%** OCR precision




**Glyph**

Rendered long text: **~3×** fewer tokens




**Render-of-Thought**

Rendered CoT **supervises latents**

**One canvas for text CoT and images: vision as the unified source for latent reasoning.**

TVERSKY, VISUALIZING THOUGHT (2013); GOEL, SKETCHES OF THOUGHT (1995); DEEPSEEK-OCR (2025); GLYPH (2025); RENDER-OF-THOUGHT (2026)

# CONTENT


1 **Background:** *From Latent to Latent Visual Reasoning*

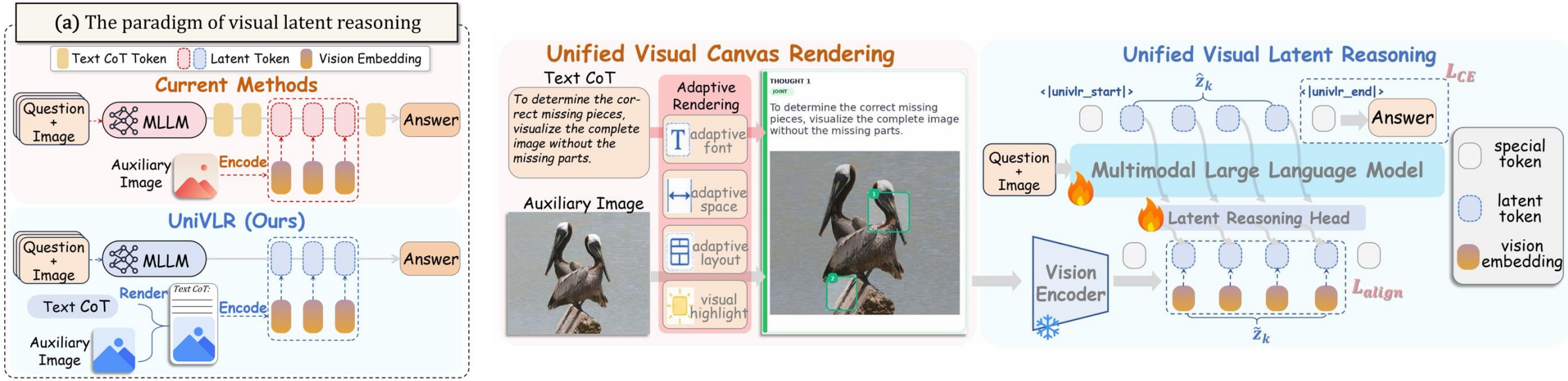
2 **Motivation:** *Vision as a Unified Source*

3 **Method:** *UniVLR*

4 **Results & Discussion**

# Method: UniVLR

## Overview: Unified Visual Latent Reasoning




### 1 Unified Visual Canvas

Render text CoT + auxiliary image into **one canvas** and encode it with the frozen vision encoder

### 2 Two-Stage Latent Alignment

Visual Latent Grounding, then **Text–Vision Unified Alignment** on canvas targets

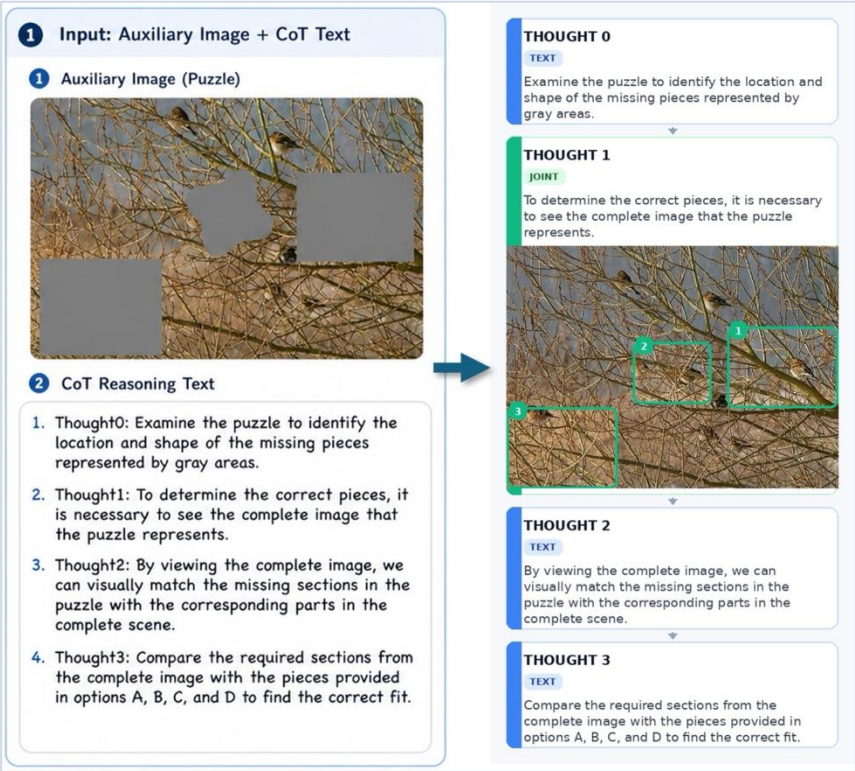
### 3 Latent-only Inference

**12 visual latent tokens**, then a direct answer, with no text CoT and no tool calls

UNIVLR (OURS), FIGURE 1(a) & FIGURE 2

# Method: UniVLR

## 1. Unified Visual Canvas Rendering




Vertical layout: auxiliary image + CoT text → one canvas

**Compose text trace R and auxiliary images U**

$$c = \Phi(\mathcal{R}, \mathcal{U}).$$

**Encode the canvas with the frozen vision encoder**

$$\mathbf{F} = E_{\text{vis}}(c) \in \mathbb{R}^{H_f \times W_f \times d},$$

**Aspect-aware 2D pooling preserves spatial layout**

$$(h^\star, w^\star) = \arg\min_{hw=K} \left| \log \frac{w}{h} - \log \frac{W_f}{H_f} \right|.$$

**→ K latent supervision targets**

$$\tilde{\mathbf{Z}} = \Pi_{h^\star, w^\star}(\mathbf{F}) = [\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K] \in \mathbb{R}^{K \times d}.$$

**Adaptive rendering:** font size, wrapping, spacing, layout and visual highlights, a general interface rather than a task-specific tool.

**Text CoT becomes a visual signal for the same vision pathway.**

UNIVLR (OURS), SEC. 2.1, FIGURE 6, APPENDIX B.5

# Method: UniVLR

## 2. Two-Stage Latent Alignment & Latent-only Inference

### Stage I · Visual Latent Grounding

Target: **auxiliary-image** embeddings $\tilde{Z}^{(1)}$

Learn to enter latent mode, emit K continuous visual tokens and exit · Visual-CoT 118.6K

### Stage II · Text–Vision Unified Alignment

Target: **unified canvas** embeddings $\tilde{Z}^{(2)}$

Absorb text CoT + visual evidence into one latent channel · 30.7K (Zebra : Visual-CoT = 7 : 3)

### Training objective (with latent teacher forcing)

$$\mathcal{S} = [\mathcal{X}, \texttt{<|univlr_start|>}, \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K, \texttt{<|univlr_end|>}, \mathcal{A}],$$

$$\hat{\mathbf{z}}_k = g_\phi(\mathbf{h}_{k-1,i}).$$

$$\mathcal{L}_{\text{align}} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{d}\left\|\text{LN}(\hat{\mathbf{z}}_k) - \text{LN}(\tilde{\mathbf{z}}_k)\right\|_2^2 + 1 - \cos\left(\text{LN}(\hat{\mathbf{z}}_k), \text{LN}(\tilde{\mathbf{z}}_k)\right)\right].$$

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}.$$

### Inference: reason only in visual latents

$\mathbf{e}_k = \hat{\mathbf{z}}_k.$ *latent fed back as the next input*

- Hidden state → lightweight **MLP head** g_φ
- L_CE only on <start> / <end> and answer tokens
- **K_infer = 12** latent steps
- At test time: **no rendering, no tools, no text CoT**

UNIVLR (OURS), SEC. 2.2–2.3, APPENDIX B.3

# Method: UniVLR

## 3. Training Data: UniVLR-SFT-140K

| Data Source | Problem Domain | Visual Operation Type | Amount |
|---|---|---|---|
| Visual-CoT | Real-world, Documents, Charts | Abstract visual operations (e.g., bounding boxes) | 118.6K |
| Zebra-CoT Visual Search | 2D Visual Reasoning | Object localization, local focus highlighting | 8.7K |
| Zebra-CoT Jigsaw | 2D Visual Reasoning | Spatial geometry, piece alignment | 8.8K |
| Zebra-CoT Maze | Visual Logic & Strategic Games | Pathfinding, spatial constraint satisfaction | 4.0K |
| **Total (UniVLR-SFT-140K)** | | | **140.1K** |

### Two-stage curriculum

- **Stage I:** full Visual-CoT (118.6K) warm-up
- **Stage II:** 30.7K, filtered Zebra-CoT : Visual-CoT = 7 : 3

### Bidirectional filtering keeps samples whose answer truly depends on the canvas

**1 Lower-bound filter**

Qwen2.5-VL-7B already answers **without** the canvas → drop (shortcut / too easy)

**2 Upper-bound filter**

Qwen2.5-VL-72B fails **even with** the canvas → drop (noisy or flawed trace)

**3 Aspect-ratio filter**

Extreme aspect ratios → drop, keeping canvases **compact and information-dense**

**Only 21.5K of ~40K Zebra-CoT samples survive; unfiltered data drops V* to 79.1.**

UNIVLR (OURS), APPENDIX C, TABLE 5; ZEBRA-COT (2025); VISUAL-COT (MONET-SFT-125K SUBSET)

# CONTENT


1 **Background:** *From Latent to Latent Visual Reasoning*

2 **Motivation:** *Vision as a Unified Source*

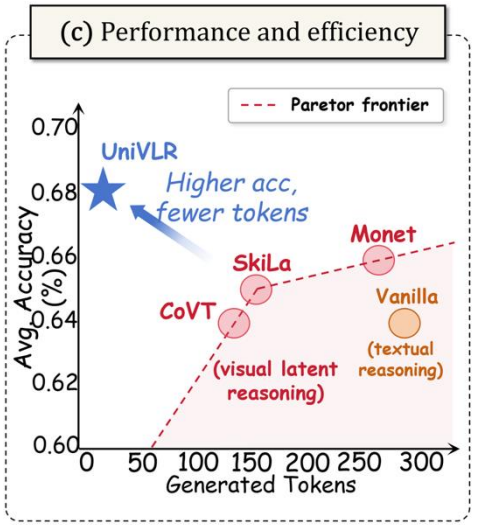
3 **Method:** *UniVLR*

4 **Results & Discussion**

# Results

## RQ1: Accuracy and Efficiency

| Model | V* | | | HRBench4K | | | HRBench8K | | | MME-RealWorld-Lite | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Attr. | Spa. | Overall | FSP | FCP | Overall | FSP | FCP | Overall | Rea. | Perc. |
| **Textual Reasoning** | | | | | | | | | | | | |
| GPT-4o [27] | 67.5* | 72.2* | 60.5* | 59.0* | 70.0* | 48.0* | 55.5* | 62.0* | 49.0* | 52.0* | 48.3* | 54.4* |
| Qwen2.5-VL-7B [23] | 77.4 | 78.3 | 76.3 | 69.0 | 85.8 | 52.2 | 66.0 | 80.3 | 51.8 | 46.2 | 43.1 | 48.2 |
| + vanilla SFT | 75.4 | 80.0 | 68.5 | 69.1 | 81.3 | 57.0 | 63.6 | 73.3 | 54.0 | 45.5 | 39.9 | 49.1 |
| **Tool-based Visual Reasoning** | | | | | | | | | | | | |
| PixelReasoner [9] | 80.6* | 83.5* | 76.3* | 72.9* | 86.0* | 60.3* | 66.9* | 80.0* | 54.3* | 49.7* | 44.5* | 53.1* |
| DeepEyes [8] | 83.3* | 84.4* | 81.6* | 71.3* | 83.8* | 58.8* | 65.1* | 77.0* | 53.3* | 54.3* | 50.5* | 56.6* |
| **Visual Latent Reasoning** | | | | | | | | | | | | |
| LVR [11] | 80.6 | 81.7 | 79.8 | 69.9 | 84.7 | 55.7 | 66.9 | 77.7 | 56.2 | 39.1 | 37.5 | 40.1 |
| Monet [12] | 79.1 | 81.7 | 75.0 | 71.9 | **89.3** | 54.5 | 63.5 | 76.5 | 50.5 | 46.9 | 40.3 | 51.2 |
| SkiLa [13] | 80.1 | 79.1 | 81.6 | 70.3 | 84.7 | 55.7 | 62.9 | 77.5 | 48.3 | 45.6 | 36.3 | 51.6 |
| CoVT [15] | 78.0 | 79.1 | 76.3 | 71.9 | 84.2 | 59.5 | **69.7** | **85.4** | 54.0 | 48.2 | 42.9 | 51.6 |
| **UniVLR** | **82.7** | **83.5** | **81.6** | **73.3** | 86.0 | **60.5** | 68.8 | 78.8 | **58.8** | **50.7** | **44.7** | **54.5** |



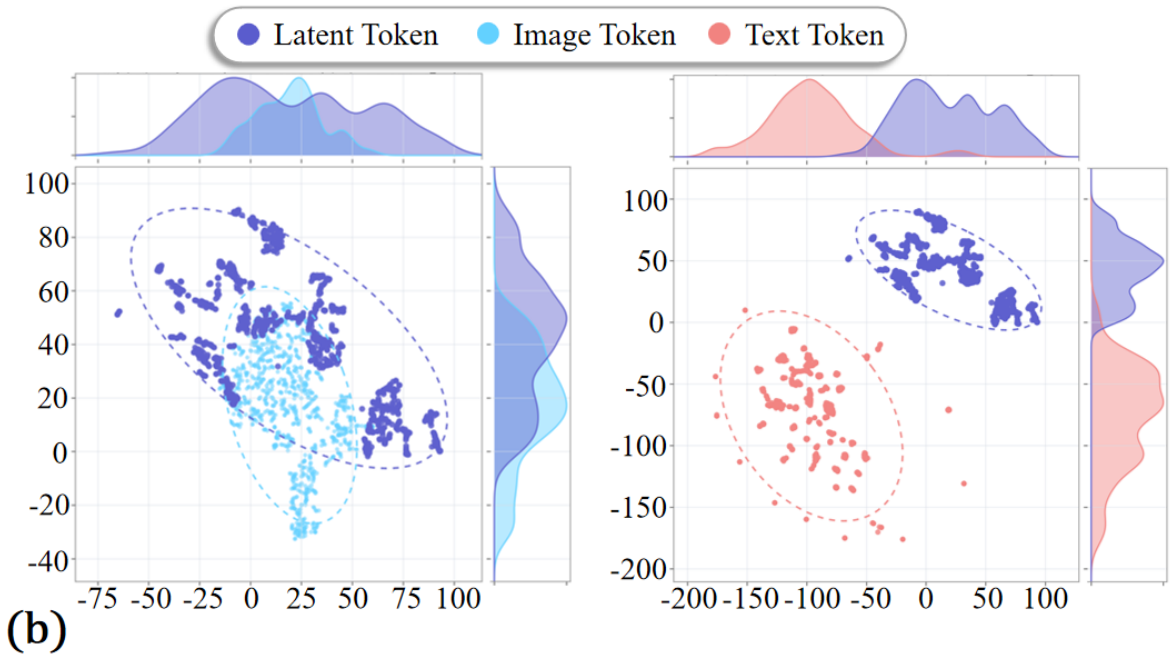

1. **Accuracy** best among latent methods; **+5.4%** on average
2. **Efficiency** 12 latents, **15.2×** fewer tokens, 33.8% lower latency

**Without text CoT, unified visual latents give higher accuracy with far fewer tokens.**

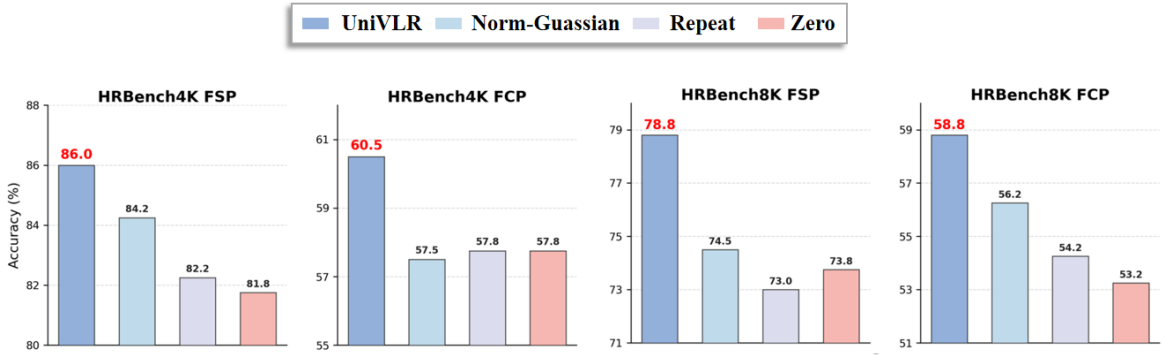
UNIVLR (OURS), TABLE 1 & FIGURE 1(c), TABLE 9. BACKBONE: QWEN2.5-VL-7B; * = REPORTED BY PRIOR WORK

# Results

## RQ2: What Do the Latent Tokens Encode?




(a) Last-hidden-state distribution: latent vs. image vs. text tokens




(b) Perturbing latents at inference: Gaussian noise / repeat / zero

1. **Visually grounded** latent states **overlap with image tokens** and stay clearly separated from text tokens

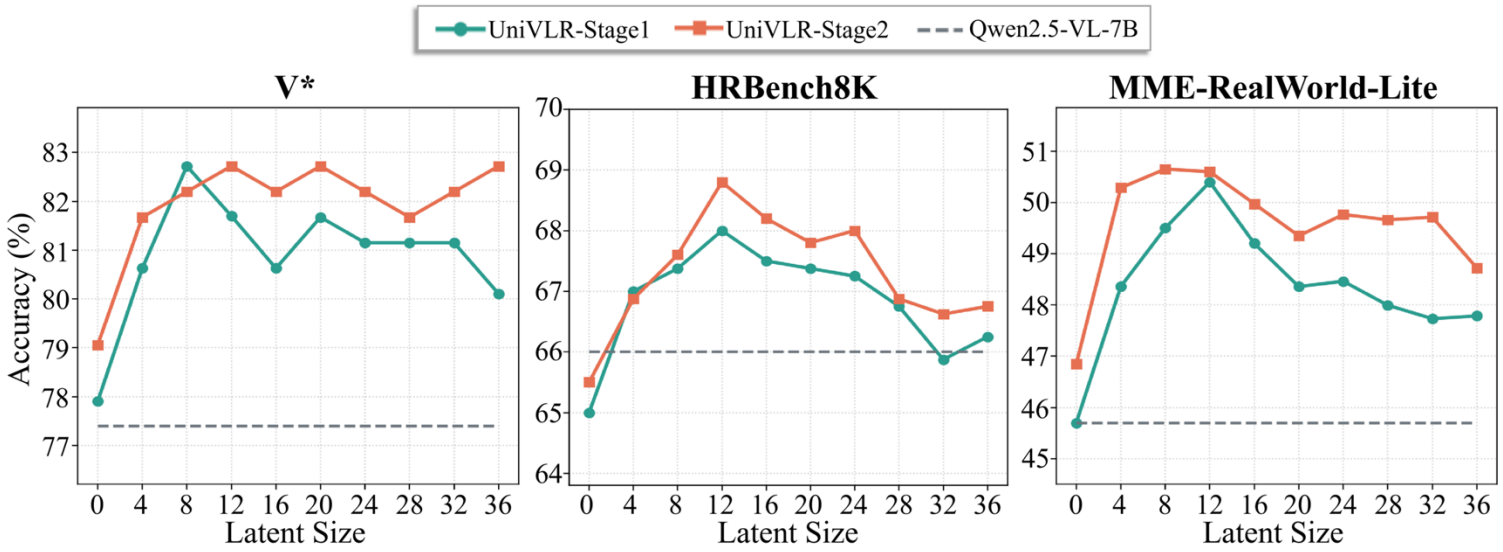
2. **Causally used** zeroing latents drops HR8K-FCP **58.8 → 53.2**; repeating the first latent drops FSP 78.8 → 73.0

**Unified latents lie on the visual representation manifold and carry task-relevant reasoning information.**

UNIVLR (OURS), FIGURE 3(b) & FIGURE 4

# Results

## RQ3: Latent Budget & Training Dynamics




| Model | V* | | | HRBench4K | | | HRBench8K | | | MME-RealWorld-Lite | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Attr. | Spa. | Overall | FSP | FCP | Overall | FSP | FCP | Overall | Rea. | Perc. |
| $\lambda_{\text{align}} = 0.1$ | **82.7** | **83.5** | **81.6** | **73.3** | **86.0** | **60.5** | **68.8** | **78.8** | **58.8** | **50.7** | **44.7** | **54.5** |
| $\lambda_{\text{align}} = 0.3$ | 82.7 | 83.5 | 81.6 | 72.8 | 86.0 | 59.5 | 66.6 | 75.5 | 57.8 | 46.9 | 42.0 | 50.1 |
| $\lambda_{\text{align}} = 0.5$ | 81.7 | 83.5 | 78.9 | 72.5 | 85.5 | 59.5 | 65.3 | 75.0 | 55.5 | 25.4 | 22.4 | 27.3 |
| $\lambda_{\text{align}} = 0.7$ | 82.2 | 82.6 | 81.6 | 71.8 | 85.5 | 58.0 | 65.5 | 75.3 | 55.8 | 29.0 | 28.8 | 29.2 |

1. **More ≠ better** K = 0 is weakest; accuracy peaks at a **moderate budget (8–12)**, then drifts
2. **Stage II stabilizes** unified alignment yields a **wider stable plateau** across K
3. **Light alignment** $\lambda = 0.1$ is best; $\lambda \geq 0.5$ collapses MME-RW (50.7 → 25.4)
4. **Full teacher forcing** beats scheduled sampling (HR8K 68.8 vs. 66.8): a stable latent manifold matters more

UNIVLR (OURS), FIGURE 5, TABLE 6 (λ_ALIGN ABLATION), TABLE 8 (TEACHER FORCING)

# Results

## RQ4: Ablation & Generalization

| Stage I: REPA Ablations | V* | HRBench8K | MME-RealWorld-Lite | | |
|---|---|---|---|---|---|
| | Overall | Overall | **Overall** | Rea. | Perc. |
| **UniVLR-Stage1 (Default)** | **81.7** | **68.0** | **50.4** | 42.4 | **55.6** |
| *Extraction Strategy (Default: Avg 2D Pool)* | | | | | |
| Avg 1D Pool | 79.1 | 66.3 | 46.5 | 38.0 | 51.9 |
| MLERP 2D Pool | 71.2 | 59.1 | 48.5 | 41.0 | 53.2 |
| *Latent Reasoning Head (Default: MLP)* | | | | | |
| w/o Head | 81.2 | 66.5 | 48.7 | **43.2** | 52.2 |
| GLU Head | 81.2 | 64.4 | 48.4 | 41.9 | 52.5 |
| *Hidden States Aligned Layer (Default: Middle)* | | | | | |
| Front Layer | 81.5 | 66.3 | 49.2 | 40.7 | 54.7 |
| Last Layer | 81.2 | 67.3 | 48.5 | 41.2 | 53.1 |

| Stage II: RFC Ablations | V* | HRBench8K | MME-RealWorld-Lite | | |
|---|---|---|---|---|---|
| | Overall | Overall | **Overall** | Rea. | Perc. |
| **UniVLR-Stage2 (Default)** | **82.7** | **68.8** | **50.7** | 44.7 | 54.5 |
| *Training Curriculum (Default: Two Stages)* | | | | | |
| Stage I Only (Warm-up) | 81.7 | 68.0 | 50.4 | 42.4 | **55.6** |
| Single-Stage Mixed | 77.5 | 64.8 | 47.1 | 42.1 | 50.2 |
| *Render Strategy (Default: Vertical Layout1)* | | | | | |
| Left-Right Layout2 | 82.1 | 67.8 | 50.1 | 43.5 | 54.5 |
| Fixed Adaptive Wrap3 | 81.6 | 66.5 | 48.8 | 44.9 | 51.2 |
| *Data Ablation (Default: Filtered Zebra + VisCoT)* | | | | | |
| Unfiltered Zebra-CoT | 79.1 | 65.5 | 49.9 | **47.1** | 51.8 |
| Pure Visual-CoT | 81.5 | 66.3 | 49.8 | 43.9 | 53.6 |

| Model | TextVQA | $MME_{translation}$ | POPE | | | | $WeMath_{loose}$ |
|---|---|---|---|---|---|---|---|
| | | | **Overall** | Adversarial | Popular | Random | |
| Qwen2.5-VL-7B [23] | 77.5 | 185.0 | 86.4 | 85.5 | 86.5 | 87.2 | **52.1** |
| LVR [11] | 75.6 | 200.0 | 84.9 | 84.3 | 84.8 | 85.7 | 48.4 |
| SkiLa [13] | 79.3 | **200.0** | 87.1 | 86.2 | 87.2 | 87.9 | 46.2 |
| CoVT [15] | 66.5 | 192.5 | 88.7 | 87.2 | 88.6 | 90.3 | 50.1 |
| UniVLR-Stage1 | 79.1 | 183.0 | 84.2 | 83.7 | 84.2 | 84.8 | 48.6 |
| **UniVLR (Ours)** | **80.4** | **200.0** | **88.8** | **87.3** | **88.9** | **90.4** | 49.4 |

1. **Spatial topology** aspect-aware 2D pooling > 1D / MLERP (MME-RW Rea. 42.4 vs. 38.0)
2. **Curriculum & layout** two stages > single-stage mix (V* 82.7 vs. **77.5**); vertical layout is best
3. **Head, layer & data** MLP head, middle-layer alignment and filtered data all help
4. **No forgetting** best TextVQA (80.4) and POPE (88.8); slight WeMath cost (52.1 → 49.4)

UNIVLR (OURS), TABLE 2 & TABLE 7. REPA = REPRESENTATION ALIGNMENT; RFC = REASONING FORMULATION & CURRICULUM

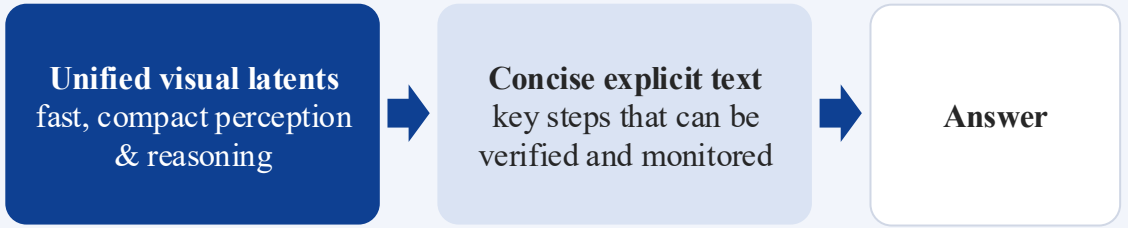
# Discussion

## Limitations & Outlook

### Limitations

- **OCR / layout dependence:** gains rely on the vision encoder reading rendered text
- **Less inspectable:** latents are not human-readable, so monitorability matters (cf. the GPT-6 Astra debate)
- **Fixed latent budget:** the same K for easy and hard questions
- **Not a tool replacement:** exact measurement and executable manipulation still need tools

### Outlook: hybrid latent–explicit reasoning




- **Adaptive latent budget**: halting or looped depth, as in recurrent-depth LMs
- **RL on visual latents** (e.g., VLPO-style) beyond SFT
- **Decode canvases back** from latents for interpretability

**Next step: let models think in a unified visual latent space and speak only when it matters.**

UNIVLR (OURS), SEC. 5; GPT-6 ASTRA MONITORABILITY DISCUSSION (AHEAD OF AI / LESSWRONG, 2026)

# Discussion

## Takeaways

1. **Trend:** Latent reasoning is a viable route beyond verbose CoT, and frontier models are moving toward it.
2. **Problem:** Existing visual latent reasoning keeps text and vision as **two separate sources**: costly and fragmented.
3. **Insight:** Vision is a natural **unified information source**, for humans and for MLLMs that read pixels.
4. **Method:** UniVLR renders text CoT + auxiliary images onto one canvas → **12 unified visual latents** → answer.
5. **Results:** **+5.4%** accuracy with **15.2×** fewer tokens, plus grounded latents and robust generalization.

Code: github.com/Warrenustc1958/UniVLR

*Thanks for Listening!*