SwimBird: ELICITING SWITCHABLE REASONING MODE IN HYBRID AUTOREGRESSIVE MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have advanced perception and reasoning by bridging vision and language, yet most still rely on textual chain-of-thought (CoT), which is ill-suited to vision-intensive tasks. Recent methods inject a fixed number of continuous hidden states as visual thoughts to strengthen visual reasoning but often weakening text-based logical reasoning. We argue that the main limitation is a rigid, pre-defined reasoning pattern that cannot choose the most suitable thinking modality for each query. We introduce SwimBird, a reasoning-switchable MLLM that dynamically selects among three input-conditioned modes: (1) text-only reasoning, (2) vision-only reasoning with continuous visual thoughts, and (3) interleaved vision–text reasoning. SwimBird adopts a hybrid autoregressive formulation that unifies next-token prediction for textual thoughts with next-embedding prediction for visual thoughts, and uses a reasoning-mode curation strategy to construct SwimBird-SFT-92K, a diverse SFT dataset covering all three reasoning patterns. By enabling query-adaptive mode selection and adaptive latent-token allocation, SwimBird preserves textual logic while boosting vision-dense reasoning. Experiments across textual reasoning and challenging visual understanding benchmarks show that SwimBird achieves state-of-the-art results and robust gains over prior fixed-pattern reasoning methods.

1 INTRODUCTION

Building on the success of Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022) reasoning in LLMs, recent multimodal research has adopted step-by-step reasoning to decompose complex vision-and-language problems into intermediate steps. With textual CoT, Multimodal Large Language Models (MLLMs) (Zhang et al., 2023; Liu et al., 2025; Tong et al., 2025c) improve performance on tasks requiring symbolic manipulation, numerical calculation, and logical analysis. However, this success does not fully transfer to vision-dense tasks where the bottleneck lies in dense perception and spatial reasoning rather than logical structure (Fu et al., 2024). Typical examples include maze solving, fine-grained visual search, and other problems where accurate intermediate visual states are essential. On such tasks, purely textual CoT (Shen et al., 2025a) can be an ill-posed interface: the model must describe intermediate visual evidence in language even when language is not a faithful carrier, causing brittle reasoning and error accumulation (Yu et al., 2025). To address this, recent works introduce latent visual reasoning (Li et al., 2025b; Tong et al., 2025a) that supervises models to generate semantically grounded continuous hidden states as visual thoughts, enabling intermediate visual representations to be maintained and updated across steps, which substantially strengthens performance on vision-dense benchmarks.

Despite these advances, existing multimodal CoT designs remain rigid. Most methods (Wang et al., 2025a; Zhang et al., 2025a; Yang et al., 2025) choose a reasoning pattern in advance: text-only CoT, vision-only latent reasoning, or a fixed interleaving of visual and textual thoughts. As illustrated in Fig. 1, this creates a mismatch between the reasoning modality and the needs of the query. Text-centric questions may be harmed by unnecessary latent visual steps, while vision-dense questions may fail when constrained to language-only reasoning. Even interleaved reasoning can be inefficient if the alternation schedule is fixed rather than chosen by the model (Tong et al., 2025b).

We argue that the core limitation is the assumption that a single static reasoning template can generalize across heterogeneous multimodal queries. Different queries require different internal

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.

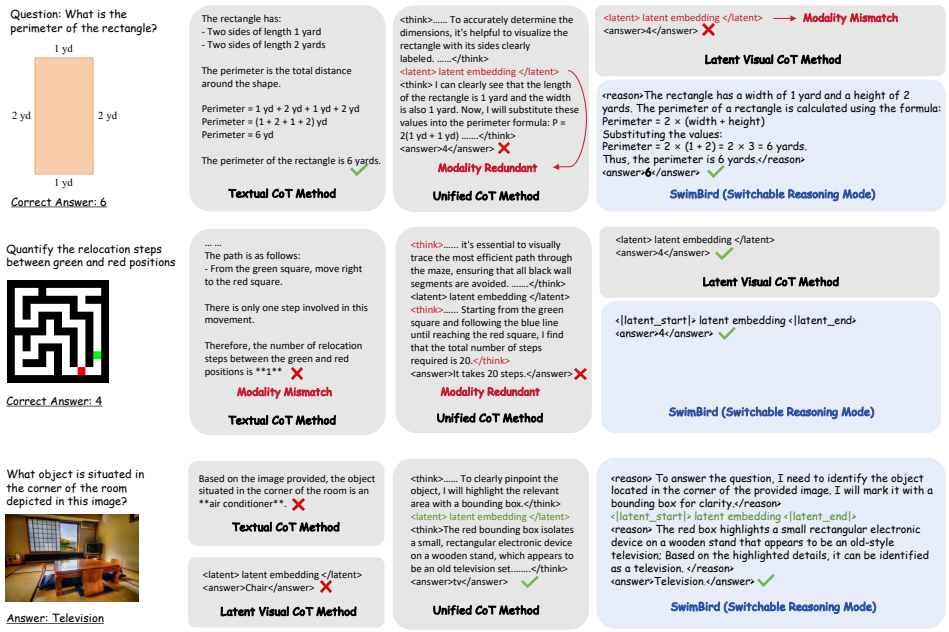


Figure 1: **SwimBird performs query-adaptive multimodal reasoning by switching among text-only, vision-only, and interleaved vision–text modes**. It avoids unnecessary latent computation for text-centric queries (Case 1), uses latent visual thoughts for vision-dense spatial reasoning (Case 2), and combines visual grounding with textual deduction when both are required (Case 3), reducing modality mismatch across diverse queries.

computation formats: some are best solved through discrete symbolic steps, some require latent visual transitions, and others require tight alternation between visual grounding and textual deduction. A stronger MLLM should therefore decide when to reason in language, when to reason in vision, and when to switch between the two.

Motivated by this, we propose **SwimBird**, a reasoning-switchable MLLM for query-adaptive multimodal reasoning. SwimBird is built on two key ideas derived from the limitations above. First, we adopt a hybrid autoregressive formulation that supports both (i) standard next-token prediction for textual thoughts and (ii) next-embedding prediction for continuous visual thoughts, providing a unified interface for switchable reasoning. Second, we address the data bias that reinforces fixed reasoning patterns. We design a curation strategy that filters and categorizes multimodal CoT samples by their visual dependency and reasoning characteristics. Through this strategy, we construct **SwimBird-SFT-92K**, a diverse SFT dataset covering text-only, vision-only, and interleaved vision–text patterns. With these designs, SwimBird can dynamically switch among three reasoning modes.

Importantly, SwimBird also removes the fixed-budget constraint in visual reasoning. Instead of generating a constant-length sequence of visual thought tokens, it dynamically determines the number of visual thought tokens during vision-only or interleaved reasoning, allocating more latent computation to vision-dense queries while avoiding redundant visual thoughts for text-centric problems. As a result, a single model can handle diverse query types, whereas fixed-pattern baselines may underperform when the required thinking modality or visual-thought budget deviates from their pre-defined design. Our contributions are as follows:

- We identify two key bottlenecks of prior multimodal CoT frameworks: fixed reasoning-mode templates and fixed visual-thought lengths, which can create a modality mismatch across heterogeneous queries.
- We introduce SwimBird, a hybrid autoregressive MLLM that dynamically switches among text-only, vision-only, and interleaved reasoning modes by combining next-token prediction with next-embedding prediction.
- We introduce adaptive visual-thought allocation, allowing SwimBird to determine the length of continuous visual reasoning spans according to query difficulty.

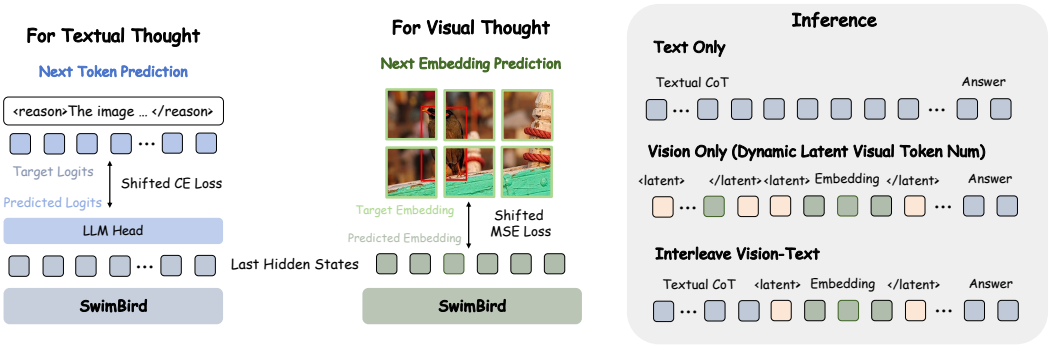


Figure 2: **SwimBird adopts a hybrid autoregressive formulation that performs next-token prediction for textual thoughts and switches to next-embedding prediction for visual thoughts**. During inference, SwimBird performs query-adaptive multimodal reasoning by dynamically selecting among three modes conditioned on the input: text-only, vision-only, and interleaved vision-text reasoning.

- We design a reasoning-mode curation strategy, construct SwimBird-SFT-92K, and show through extensive experiments that SwimBird improves vision-dense reasoning while preserving strong text-centric reasoning ability.

2 Related Works

2.1 Textual CoT in MLLMs

The integration of vision and language has evolved from discriminative tasks toward generative reasoning frameworks. Early MLLMs focus mainly on VQA through direct answer generation (Li et al., 2023; Liu et al., 2023; Wang et al., 2024a; Liu et al., 2024; Yan et al., 2025). With the success of step-by-step reasoning in LLMs, recent MLLMs incorporate explicit reasoning chains for complex multimodal problems (Bai et al., 2025; Wang et al., 2025d; Xu et al., 2025). These models generate intermediate textual explanations before final answers, improving mathematical word problems, scientific diagram understanding, and multi-hop visual reasoning (Yue et al., 2024; Wang et al., 2024b; Qiao et al., 2025). Despite their effectiveness on logic-heavy benchmarks, text-based reasoning approaches struggle when the core challenge is visual perception rather than logical decomposition (Shen et al., 2025b). Tasks involving spatial tracking, visual prediction, or fine-grained comparison expose this limitation: intermediate visual evidence must be compressed into language even when language is not a faithful or efficient carrier, leading to brittle reasoning and error accumulation.

2.2 Latent Visual Reasoning

Recognizing the constraints of language-only reasoning, researchers have explored alternative computational substrates for visual thinking (Qin et al., 2025; Wang et al., 2025c). Recent methods train models to produce continuous embeddings supervised by visual reconstruction objectives. For instance, Mirage (Yang et al., 2025) uses hidden states trained to approximate annotated helper images, while LVR (Li et al., 2025b) reconstructs cropped image regions. SkiLa (Tong et al., 2025a) proposes unified reasoning that alternates between latent visual tokens and discrete textual tokens. However, existing latent reasoning methods apply the same reasoning structure to all inputs: models trained with visual thoughts produce them even for purely textual queries, and use fixed-length latent tokens regardless of the deliberation required. SwimBird overcomes both limitations via dynamic mode selection and adaptive visual token budgets, enabling query-adaptive multimodal reasoning.

3 METHOD

SwimBird adopts a hybrid autoregressive formulation that supports both discrete textual tokens and continuous latent visual tokens. As in Fig. 2 (left), it performs next-token prediction for textual thoughts with a shifted cross-entropy loss, and next-embedding prediction for visual thoughts with an MSE loss that reconstructs the embeddings of intermediate thinking images. In inference (Fig. 2 right), it performs query-adaptive reasoning by generating, conditioned on the input, (i) text-only traces, (ii) vision-only traces with a variable-length latent span, or (iii) interleaved vision–text traces.

3.1 HYBRID AUTOREGRESSIVE MODELING

Textual thought as next-token prediction. For textual reasoning, SwimBird behaves like a standard language model. Given a token sequence $\{w_1, \ldots, w_T\}$, it outputs logits parameterizing

$$p_\theta(w_t \mid w_{<t}, \mathbf{x}), \tag{1}$$

where $\mathbf{x}$ denotes observed images (and prior context). We train spans with the cross-entropy loss:

$$\mathcal{L}_{\text{text}} = -\sum_{t=1}^{T} \log p_\theta(w_t \mid w_{<t}, \mathbf{x}). \tag{2}$$

It preserves the discrete symbolic manipulation and logical consistency of the language backbone, which is essential for text-centric reasoning tasks.

Visual thought as next-embedding prediction. For vision-only reasoning or visual segments inside interleaved reasoning, SwimBird generates a sequence of continuous latent tokens (visual thoughts) $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, each represented as a hidden-state embedding rather than a discrete word. We treat each visual-thought step as predicting the next embedding in an autoregressive manner:

$$\hat{\mathbf{z}}_k = f_\theta(\mathbf{z}_{<k}, w_{\leq T}, \mathbf{x}), \tag{3}$$

and supervise it with a shifted mean squared error (MSE) loss against target embeddings $\mathbf{z}_k$:

$$\mathcal{L}_{\text{vis}} = \sum_{k=1}^{K} \|\hat{\mathbf{z}}_k - \mathbf{z}_k\|_2^2 . \tag{4}$$

Here, the target embeddings are computed by encoding the intermediate thinking images with the same vision encoder (and projection) used by SwimBird, thus grounding latent visual thoughts in semantically meaningful visual states.

Unified training objective. A training instance may contain pure textual CoT, pure visual CoT, or interleaved segments. We optimize a unified objective that sums modality-specific losses over the activated segments:

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}}, \tag{5}$$

where λ_{text} and λ_{vis} are balancing coefficients. In practice, each sample only contributes to the losses of the modes it contains, enabling the model to learn all three reasoning patterns without forcing unnecessary supervision.

Mode switching with special delimiters. To enable controllable and learnable switching among reasoning modes, we introduce explicit delimiters in the target sequences. Specifically, we mark visual-thought spans using special tokens such as <|latent_start|> and <|latent_end|>. During training, these delimiters define where the model should produce continuous latent embeddings instead of textual tokens. During inference, SwimBird generates these delimiters autoregressively, which makes mode selection *query-adaptive*: the model can decide whether to enter a latent visual-thinking phase, remain in text-only reasoning, or alternate between the two (Fig. 2 right).

Data Source	All Mode	Text Only	Vision Only	Interleave	Problem Domain
Zebra-CoT	26.3K	0	5.9K	20.4K	Visual Search, Jigsaw, Maze, Geometry, Chess...
ThinkMorph	7.1K	0	1.2K	5.9K	Visual Search, Spatial Navigation, Jigsaw, Chart
MathCanvas	8.9K	0	1.7K	7.2K	Geometry, Algebra, Calculus, Statistics
OpenMMReasoner	50K	50K	0	0	General VQA, Math VQA, Text QA
Total	**92.3K**	**50K**	**8.8K**	**33.5K**	

Table 1: Detailed statistics of SwimBird-SFT-92K.

3.2 Dynamic Latent Token Budget

Prior latent visual reasoning methods use a fixed number of latent tokens (or a fixed pooling strategy) for all inputs, leading to two drawbacks: (1) capacity may be insufficient for vision-dense, high-resolution images yet wasted on simple ones; and (2) pooling intermediate images to a fixed token length discards spatial details, hindering the learning of semantically meaningful latent embeddings.

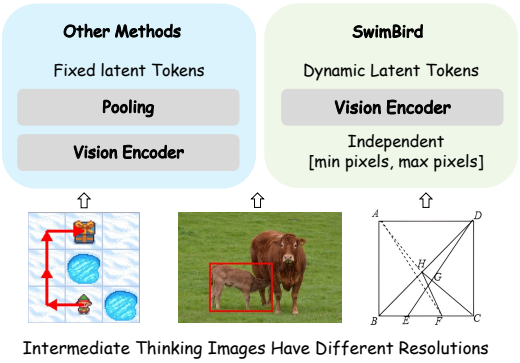


Figure 3: **Resolution-aware, dynamic tokens budget.**

As shown in Figure 3, SwimBird adopts a resolution-aware, dynamic latent-token budget. Leveraging the native dynamic-resolution capability of the Qwen ViT, we set a maximum pixel budget for intermediate thinking images, allowing the vision encoder to emit a variable number of visual tokens within the range $[N_{\min}, N_{\max}]$ depending on image resolution. This avoids aggressive pooling that would drop fine-grained evidence, while preventing overly long visual sequences from dominating computation. As a result, SwimBird preserves detailed visual information when needed and stays efficient on simpler cases. Building on this resolution-aware training, SwimBird learns to allocate latent computation dynamically at inference time. In the vision-only and interleaved modes, the number of latent tokens K is not fixed in advance: the model keeps producing latent embeddings until it emits `</latent>`. This variable-length latent span aligns the amount of visual thinking with the perceived difficulty of the query.

3.3 Switchable Reasoning SFT Dataset Construction

To enable switchable reasoning modes, we curate a diverse SFT dataset covering three reasoning patterns: (1) text-only CoT, (2) vision-only CoT, and (3) interleaved vision-text CoT. The curation pipeline consists of three stages:

Stage 1: Candidate collection and easy-instance filtering. We collect raw image-text interleaved CoT data from ThinkMorph (Gu et al., 2025), Zebra-CoT (Li et al., 2025a), and MathCanvas-Instruct (Shi et al., 2025). These datasets provide multimodal reasoning chains where each sample contains intermediate thinking images. To focus on cases where visual reasoning is useful, we remove instances that are already solvable from the original input: Qwen3VL-8B is evaluated on the question and the original image, and correctly answered samples are filtered out.

Stage 2: Reasoning-mode labeling via pass@8. For each remaining sample, we compute two pass@8 scores with Qwen3VL-8B: $\text{pass}_{\text{base}}$ uses only the question and problem image, while $\text{pass}_{\text{hint}}$ additionally provides the intermediate thinking images as visual hints, with answers judged by Qwen3-235B-Instruct. We keep samples with $\text{pass}_{\text{hint}} \geq \text{pass}_{\text{base}}$, indicating that intermediate thinking images provide non-negative gains. Among them, we label samples with $\text{pass}_{\text{hint}} \geq 0.75$ as vision-only, since the model can solve the problem with high probability using the intermediate thinking images without an explicit textual CoT. The remaining kept samples, where $\text{pass}_{\text{hint}} \geq \text{pass}_{\text{base}}$ but $\text{pass}_{\text{hint}} < 0.75$, are labeled as interleaved vision–text, since the images help but are insufficient for consistently correct solutions and textual reasoning is still needed.

Model	V* Bench	HR-Bench 4K	HR-Bench 8K	MME RealWorld	Avg.
Textual Reasoning Models					
GPT-4o (Hurst et al., 2024)	66.0	59.0	55.5	62.8	60.9
GPT-5-mini	63.9	66.3	60.9	-	-
Qwen2.5-VL-32B-Instruct	80.6	69.3	63.6	59.1	68.2
Qwen2.5-VL-7B-Instruct	75.3	65.5	62.1	56.8	64.9
Qwen3-VL-8B-Instruct *	83.8	76.5	71.3	61.9	73.4
Qwen3-VL-8B-Thinking	77.5	72.4	68.1	-	-
InternVL3-8B (Zhu et al., 2025)	81.2	70.0	69.3	-	-
LLaVA-OneVison (Li et al., 2024)	75.4	63.0	59.8	57.4	63.9
Vision-R1 (Huang et al., 2025)	80.1	64.8	57.0	-	-
Latent Visual Reasoning Models					
Monet (Wang et al., 2025c)	83.3	71.0	68.0	-	-
LVR (Li et al., 2025b)	81.7	69.6	66.1	-	-
SkiLa (Tong et al., 2025a)	84.3	72.0	66.5	-	-
Multimodal Agentic Models					
SEAL (Wu & Xie, 2024)	74.8	-	-	-	-
Pixel Reasoner (Wang et al., 2025b)	84.3	72.6	66.1	64.4	71.9
DeepEyes (Zheng et al., 2025)	83.3	73.2	69.5	64.1	72.5
Thyme (Zhang et al., 2025c)	82.2	77.0	72.0	64.8	74.0
DeepEyesV2 (Hong et al., 2025)	81.8	77.9	73.8	64.9	74.6
SwimBird	**85.5**	**79.0**	**74.9**	**65.3**	**76.2**

Table 2: Performance on visual understanding benchmarks. Here, * denotes the results are reproduced by ourselves.

Stage 3: Add text-only CoT data. To complete the three-mode training set, we sample 50K text-only CoT instances from OpenMMReasoner (Zhang et al., 2025b), which provides pass@8-filtered textual CoT traces. Combining them with the 42K samples from Stage 2 yields **SwimBird-SFT-92K**, covering text-only, vision-only, and interleaved vision–text patterns. Detailed statistics are reported in Table 1.

4 EXPERIMENTS

Training Details. We adopt Qwen3-VL-8B (Bai et al., 2025) as the base model and conduct supervised fine-tuning on our curated SwimBird-SFT-92K. Training is performed with a global batch size of 128. The vision encoder and multimodal projector are kept frozen, and only the LLM parameters are updated. A cosine learning rate scheduler is applied with an initial learning rate of 1e-5.

Baselines and Benchmarks To comprehensively assess the effectiveness of SwimBird, we compare it against three categories of baselines: (1) textual reasoning models, including advanced closed-source systems (e.g., GPT-4o and GPT-5-mini) and state-of-the-art open-source models; (2) latent visual reasoning models (e.g., Monet, LVR, SkiLa); and (3) multimodal agentic models that rely on explicit tool/workflow designs (e.g., Pixel Reasoner, DeepEyes, Thyme). Evaluation covers two benchmark groups: (i) fine-grained visual understanding (V* Bench (Wu & Xie, 2024), HR-Bench 4K/8K (Wang et al., 2025e), MME-RealWorld (Zhang et al., 2024b); Table 2), and (ii) general VQA and multimodal reasoning (MMStar (Chen et al., 2024), RealWorldQA (Dsouza, 2025), WeMath (Qiao et al., 2025), DynaMath (Zou et al., 2024), MathVerse (Zhang et al., 2024a)). Results marked with * are reproduced by ourselves.

4.1 MAIN RESULTS

Fine-grained Visual Understanding. Table 2 demonstrates that SwimBird achieves state-of-the-art performance on fine-grained and high-resolution perception. SwimBird outperforms strong textual reasoning baselines such as Qwen3-VL-8B-Instruct. Notably, Qwen3-VL-Thinking performs worse than Qwen3-VL-Instruct on visual perception, further supporting our claim that a mismatched reasoning mode can harm performance. Furthermore, SwimBird also outperforms current SoTA multimodal agentic models such as Thyme and DeepEyesV2, which enhance perception via explicit

Models	General VQA		Multimodal Reasoning		
	MMStar	RealWorldQA	WeMath	DynaMath	MathVerse_MINI
Qwen2.5-VL-32B-Instruct	70.3	-	-	-	48.5
Qwen2.5-VL-7B-Instruct	60.3	67.4	34.6	53.3	45.6
Qwen3-VL-8B-Instruct *	64.7	71.8	38.8	65.3	61.3
LLaVA-OneVision (Li et al., 2024)	61.9	69.9	20.9	-	19.3
DeepEyes (Zheng et al., 2025)	-	-	38.9	55.0	47.3
DeepEyesV2 (Hong et al., 2025)	-	-	38.1	57.2	52.7
SkiLa (Tong et al., 2025a)	64.8	69.3	-	-	-
SwimBird	**71.2**	**73.1**	**49.5**	**67.2**	**65.8**

Table 3: Performance on general VQA and multimodal reasoning tasks. * denotes the results are tested by ourselves.

cropping tools, highlighting that SwimBird can achieve stronger fine-grained perception without relying on complex tool pipelines. We attribute these gains to SwimBird's query-adaptive reasoning mode switching and adaptive latent-token allocation. Fine-grained visual tasks often require precise spatial evidence that is difficult to faithfully compress into text; meanwhile, forcing latent visual thoughts on text-centric steps can be redundant. By switching to vision-only reasoning when dense perception is needed (and allocating more latent computation for high-resolution inputs), SwimBird better preserves visual details and reduces modality mismatch, leading to consistently higher accuracy.

General VQA and Multimodal Reasoning. Beyond perception, SwimBird also shows strong improvements on general VQA and reasoning-heavy benchmarks. As in Table 3, SwimBird exceeds Qwen3-VL-8B-Instruct and even outperforms Qwen2.5-VL-32B-Instruct. More importantly, SwimBird delivers clear gains on multimodal reasoning, outperforming strong open-source methods and agentic models. These suggest that SwimBird's latent visual thoughts do not come at the cost of symbolic reasoning. Instead, it stays in text-only reasoning when the task is primarily linguistic or mathematical, and invokes vision-only or interleaved latent thinking only when additional visual evidence is beneficial. Learned from the multi-pattern supervision in SwimBird-SFT-92K, this query-adaptive selection avoids redundant visual thoughts that could interfere with textual logic, while still leveraging latent visual computation for vision-dependent subproblems.

Latent Tokens	HRBench4K	HRBench8K	RealWorldQA
16	76.4	71.4	73.1
32	79.0	74.9	73.1
64	77.8	73.4	72.7
128	76.0	71.8	72.7

Table 4: Impact of maximum latent tokens budget.

MSE Weight	HRBench4K	HRBench8K	RealWorldQA
0.1	79.0	71.8	72.8
0.2	79.0	74.9	73.1
0.5	77.8	75.9	72.0
1.0	79.4	73.8	71.9

Table 5: Impact of MSE loss weight coefficients.

4.2 Ablation Studies

Impact of the Maximum Latent Token Budget. We study how $N_{\max}$ influences performance under the dynamic range $[N_{\min}, N_{\max}]$, with $N_{\min} = 2$ fixed to preserve effective resolution for small images and $N_{\max} \in \{16, 32, 64, 128\}$. As shown in Table 4, increasing $N_{\max}$ from 16 to 32 yields clear gains on vision-dense benchmarks, showing that a moderate upper bound provides sufficient capacity for high-resolution perception. However, further expanding $N_{\max}$ to 64 or 128 does not help and even degrades performance, while RealWorldQA slightly drops. This suggests that an overly large latent budget may introduce redundant visual computation and interfere with overall reasoning. We thus use $N_{\max} = 32$ as the default.

Impact of the MSE Loss Weight Coefficient. We ablate the weight of the visual-thought reconstruction loss by varying λ_{vis} while keeping other settings fixed. As shown in Table 5, $\lambda_{\text{vis}} = 0.2$ achieves the most balanced performance across all benchmarks. A too-small weight (0.1) provides weak supervision on latent visual thoughts, causing a notable drop on the most vision-dense benchmark (HRBench8K: 71.8). Conversely, increasing λ_{vis} to 0.5 improves HRBench8K (75.9) but degrades RealWorldQA (72.0), suggesting that over-emphasizing MSE biases the model toward visual reconstruction at the expense of general multimodal reasoning. With $\lambda_{\text{vis}} = 1.0$, HRBench4K slightly increases (79.4), but both HRBench8K and RealWorldQA drop, indicating instability under overly

strong visual-loss weighting. Overall, we use $\lambda_{\text{vis}} = 0.2$ as the default, which best balances visual reasoning and text-centric reasoning.

4.3 Analysis of Switchable Reasoning Mode

Analysis of Reasoning-Mode Distribution. We analyze the distribution of SwimBird's reasoning modes across benchmarks (Fig. 4) to verify its query-adaptive behavior. On **text-logic-dominant** datasets (DynaMath and MathVerse_MINI), SwimBird almost always uses text-only reasoning, with vision-only and interleaved traces rarely triggered, suggesting it avoids redundant latent visual thoughts when symbolic and linguistic deduction are sufficient. On **vision-dense perception** benchmarks (V* Bench and HR-Bench 4K/8K), it frequently activates vision-only and especially interleaved reasoning, reflecting the need to alternate between visual grounding (e.g., tiny targets in high-resolution images) and explicit textual deduction. The vision-only ratio further grows from HR-Bench 4K to 8K, consistent with higher perceptual load at higher resolutions. WeMath exhibits a more **balanced** mixture of all three modes, as its problems vary between text-centric and vision-grounded. These results confirm that SwimBird does not follow a fixed template, but instead selects reasoning modes in an instance-dependent manner to mitigate modality mismatch.

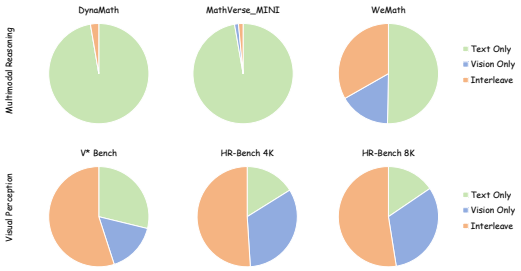


Figure 4: Distribution of reasoning mode across different benchmarks for SwimBird.

Mode	V*	HR4K	MMStar	MathVerse MINI
Text-Only	83.8	73.1	68.5	63.1
Visual-Only	85.6	78.9	70.7	57.2
Interleave	85.2	77.0	70.9	60.4
Adaptive	**85.5**	**79.0**	**71.2**	**65.8**

Table 6: By forcing the reasoning mode in benchmarks, the adaptive reasoning mode is the most effective one.

Token Budget	V*	HR4K	MMStar	RealWorldQA
Fixed Token Num (16)	76.0	71.5	62.5	72.7
Fixed Token Num (32)	78.2	74.3	61.9	72.7
Fixed Token Num (64)	77.5	74.6	61.4	72.1
Adaptive Token Num (0~32)	**79.0**	**74.9**	**71.2**	**73.1**

Table 7: By forcing the reasoning token budget, the adaptive reasoning budget is the most effective one.

Ablation on Reasoning Modes. To verify the benefit of adaptive mode selection, we compare the model using a single fixed reasoning mode at inference time against our adaptive strategy. As shown in the Tab. 6, fixed modes suffer from modality mismatch: vision-only excels on perception tasks (HR4K, V*) but degrades on MathVerse Mini, while text-only shows the opposite. The adaptive mode consistently matches or approaches the best single-mode performance across all task types.

Inference Time (Hour:Min:Second)	HR8K	MMStar
Qwen3-VL-8B-Instruct	41:10	1:21:30
SwimBird-8B	**25:35**	**45:27**

Table 8: Verification of lower inference costs.

Mean Response Length (Tokens)	HR8K	MMStar
Qwen3-VL-8B-Instruct	477.4	641.5
SwimBird-8B	**156.2**	**244.2**

Table 9: Verification of reduced reasoning redundancy.

Verification of Adaptive Budget. To verify the adaptability of the reasoning token budget, we conduct ablation experiments comparing fixed-budget baselines against our adaptive strategy in Tab.7. The results clearly confirm this trade-off: a budget of 32 yields the best performance on HR4K but falls behind 64 on HR8K, while larger budgets significantly hurt performance on MMStar compared to 16. No fixed budget achieves consistently strong results across all benchmarks. In contrast, our adaptive strategy dynamically allocates visual tokens according to each query's actual demand, achieving the best performance across all tasks.

Verification of Inference Cost. Latent reasoning does not add inference overhead; instead, it reduces inference cost for two reasons. First, a single latent hidden state can encode rich visual information that would otherwise require many text tokens, substantially reducing the number of generated tokens

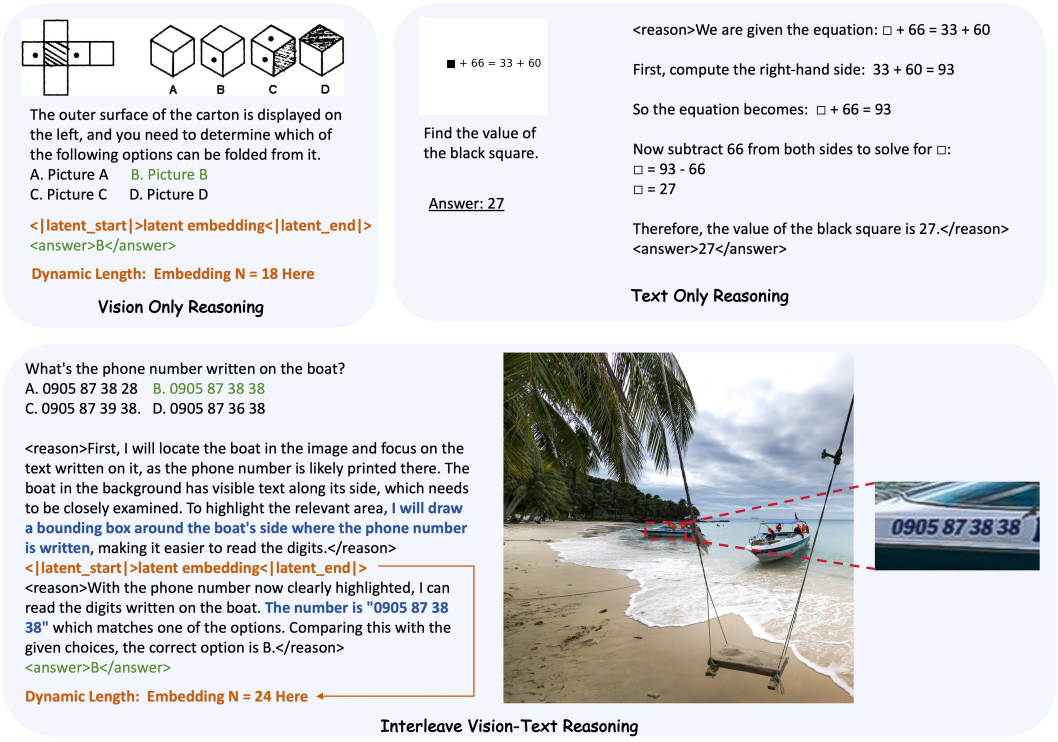


Figure 5: **Analysis of Different Reasoning-Mode Case.**

during reasoning. Second, because visual thoughts remain in latent space, the model does not need to decode intermediate representations into discrete tokens and then re-encode them as embeddings, further reducing computational overhead. We validate this empirically by comparing inference time and mean response token count between SwimBird-8B and Baseline, and we can achieve reduced reasoning lengths (Tab. 9) and inference costs (Tab. 8).

Analysis of Different Reasoning-Mode Cases. Fig. 5 provides qualitative examples of SwimBird's mode selection. For vision-only reasoning (top-left), the cube-net folding problem mainly requires spatial perception and mental rotation; SwimBird directly enters a latent visual-thought span and outputs the answer without unnecessary textual CoT, while allocating an appropriate latent length (e.g., $N{=}18$). For text-only reasoning (top-right), the arithmetic equation is purely symbolic; SwimBird solves it with textual deduction, avoiding redundant visual thoughts that could interfere with logical steps. For interleaved vision–text reasoning (bottom), reading a phone number from a small image region requires both precise visual localization and explicit option comparison; SwimBird first uses latent visual thoughts to focus on the relevant region, then switches back to text for verification and decision making, again with a dynamically allocated latent length (e.g., $N{=}24$). Together, these cases show that SwimBird mitigates modality mismatch by choosing when to think in vision versus language and by adaptively allocating visual-thought computation to match perceptual difficulty.

5 Conclusion

We present **SwimBird**, a reasoning-switchable MLLM that addresses the fixed reasoning pattern in current methods. Built on a hybrid autoregressive paradigm, SwimBird adaptively switches among text-only, vision-only, and interleaved vision–text reasoning, while dynamically allocating the latent visual-token budget. To enable multi-mode training, we construct **SwimBird-SFT-92K** via a systematic curation and mode-labeling pipeline. Extensive experiments show that SwimBird achieves SoTA performance on both text-centric reasoning and vision-dense tasks.

AI USE STATEMENT

In this work, we did not use generative AI tools for any task with required disclosure. Specifically, generative AI was not used to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or assist in writing proofs, propose or refine hypotheses, design or provide feedback on our research methodology or experiments, implement methods, perform translation, clean or reformat datasets, support qualitative or thematic analysis, or interpret results; these uses are not applicable to this work. Among tasks with recommended disclosure, we used generative AI tools solely to aid and polish the writing, i.e., correcting grammar and spelling and improving the readability and phrasing of author-written text. We did not use generative AI for other recommended-disclosure tasks such as creating scientific figures, writing or editing code, searching for or identifying literature, formatting references, or brainstorming research ideas. We have reviewed all AI-assisted edits to ensure they are correct and faithfully preserve the authors' intended meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Nigel Dsouza. Comparative analysis of leading generative ai conversational systems: Chatgpt, grok ai, gemini, and meta ai. *Authorea Preprints*, 2025.

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*, pp. 148–166. Springer, 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.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

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*, 2025.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025a.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6613–6629, 2025b.

Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*, 2025a.

Jintao Tong, Wenwei Jin, Pengda Qin, Anqi Li, Yixiong Zou, Yuhong Li, Yuhua Li, and Ruixuan Li. Flowcut: Rethinking redundancy via information flow for efficient vision-language models. *arXiv preprint arXiv:2505.19536*, 2025b.

Jintao Tong, Shiwei Li, Zijian Zhuang, Jinghan Hu, and Yixiong Zou. Emosync: Multi-stage reasoning with multimodal large language models for fine-grained emotion recognition. In *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing*, pp. 95–99, 2025c.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a.

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*, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*, 2025c.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025d.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025e.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024b.

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

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*, pp. 13084–13094, 2024.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

Shilin Yan, Jiaming Han, Joey Tsai, Hongwei Xue, Rongyao Fang, Lingyi Hong, Ziyu Guo, and Ray Zhang. Crosslmm: Decoupling long video sequences from lmms via dual cross-attention mechanisms. *arXiv preprint arXiv:2505.17020*, 2025.

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.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*, 2025a.

Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Ziwei Liu, Xingxuan Li, and Lidong Bing. Openmmreasoner: Pushing the frontiers for multimodal reasoning with an open and general recipe. *arXiv preprint arXiv:2511.16334*, 2025b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024b.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025c.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

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.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

A PROMPT

To guide SwimBird's query-adaptive reasoning mode selection, we design a system prompt that explicitly instructs the model on how to switch between textual and visual thinking modes. As shown in Figure 6, the prompt defines three reasoning patterns: text-only, vision-only, and interleaved, using structured tags (`<reason>` for textual thoughts and `<latent>` for visual thoughts), and allows the model to dynamically choose the most appropriate mode or combination based on the input query.

System Message

You are a multimodal reasoning assistant capable of thinking in textual and visual modes.

Use the following tags to switch your thinking mode:
1. **Textual Mode**: `<reason>Your textual reasoning process</reason>`
- For logical analysis, planning, and verbal thought.

2. **Visual Mode**: `<|latent_start|>Your visual reasoning process<|latent_end|>`
- For mental visualization, imagination and simulation.

Output Rules:
- Depending on the problem, you can use: textual reasoning only, visual reasoning only, or a mix of both (alternating multiple times as needed).
- After all thinking is complete, place the final answer inside `<answer>Your Final Answer</answer>`.

Figure 6: **System prompt used for SwimBird.**

B EFFECTIVENESS ACROSS MODEL SIZES

To assess the scalability of SwimBird, we extend it to a smaller backbone, Qwen3-VL-4B, and compare it against the corresponding instruction-tuned baseline. As shown in Table 10, SwimBird-4B achieves consistent gains across all benchmarks, covering both fine-grained perception (V*, HR-Bench 4K/8K) and general VQA (MMStar, RealWorldQA). This confirms that our switchable reasoning recipe generalizes well across model scales.

Method	V*	HR4K	HR8K	MMStar	RealWorldQA
Qwen3-VL-4B-Instruct	78.0	75.3	68.5	66.4	70.8
SwimBird-4B	82.7	76.2	70.1	71.0	71.5

Table 10: Performance of SwimBird applied to a smaller backbone (Qwen3-VL-4B).