AdaThinking-E: One-Token Entropy Regulation for Adaptive Thinking

Zining Wang[1*], Tongkun Guan[2*], Boming Chen[1*], Zhentao Guo[1], Jianqiang Liu[1], Chao Jin[3], Chen Duan[1], Kai Zhou[1(✉)], Pengfei Yan[1], Wei Shen[2(✉)], and Xiaokang Yang[2]

[1] Meituan, China,
{wangzining03,chenboming,guozhentao,zhoukai03}@meituan.com
[2] MoE Key Lab of Artificial Intelligence, AI Institute, School of Computer Science, Shanghai Jiao Tong University, China, gtk0615@sjtu.edu.cn
[3] MAIS&NLPR, Institute of Automation, Chinese Academy of Sciences, China, jinchao2024@ia.ac.cn

Abstract. Multimodal large language models have demonstrated strong document reasoning capabilities by incorporating explicit thinking processes. While this capability significantly improves performance on challenging tasks, current models apply such deep reasoning uniformly to all questions, resulting in unnecessary computational overhead for simple task. This not only degrades user experience but also negatively impact accuracy on benchmark datasets. We identify the critical need for adaptive thinking mechanisms that can intelligently determine when to engage reasoning based on question complexity. To address this, we propose AdaThinking-E, a novel reinforcement learning framework that learns adaptive thinking through one-token entropy regulation. Our key insight is that model confidence in the decision to engage thinking (or not) can be quantified through entropy analysis of the predicted probability distribution at critical decision tokens. This observation motivates our entropy-governed reward mechanism: the training process naturally transitions from high-entropy exploration, where the model experiments with different thinking strategies, to low-entropy convergence with confident, generalizable decision-making policies. Crucially, this approach enables models to intrinsically discover when to think without requiring manual intervention or external difficulty labels. Extensive experiments demonstrate that our approach enables models to be both accurate on complex problems and efficient on simple ones across diverse document tasks. Code is available at https://github.com/PriNing/AdaThinking-E.

Keywords: Adaptive Thinking · Entropy · Document Understanding

1 Introduction

Multimodal large language models (MLLMs) [55–57] have demonstrated remarkable capabilities in document understanding, ranging from visual text recogni-

*Equal contribution. ✉Corresponding author.

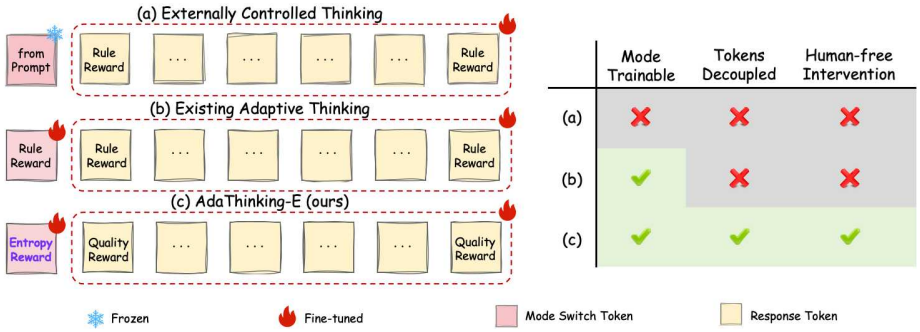

Fig. 1: Different RL training paradigms for the mode switch token in adaptive thinking models: (a) **Externally Controlled Thinking**, the prompt explicitly specifies the output mode (thinking or non-thinking); (b) **Existing Adaptive Thinking**, the model adaptively switches between thinking and non-thinking modes via special token, using the rule reward values involving subjective human judgments to learn when to think; (c) Our proposed **AdaThinking-E**, mode switch token and response tokens are optimized separately. Entropy regulation is employed to encourage the model to transition from autonomous exploration to confident decision-making, free from subjective human bias, thereby achieving truly adaptive thinking.

tion to sophisticated reasoning that solves solid geometry problems, interprets complex layouts, and answers logic-intensive questions. This progress is largely driven by recent advances in reinforcement learning [35, 59, 72], which enable models to generate step-by-step explanations for complex questions.

However, many everyday question don't require such deep reasoning. Consider a simple question like "Which department issued this memo?"—the answer can be directly extracted from the document header. When applying deep reasoning to all questions, they incur unnecessary computational overhead on simple queries and may even introduce hallucinations through excessive deliberation. Therefore, we argue document MLLMs require an adaptive thinking mechanism [66, 67, 78, 86] that intelligently determines when to engage thinking based on question complexity. This adaptive capability would enable models to be both accurate on complex problems and efficient on simple ones.

Despite its importance, adaptive thinking remains largely unresolved. Current mainstream MLLMs [5, 9, 76, 77] rely on manual intervention for mode selection, which lacks flexibility and requires predefined rules that fail to generalize, as illustrated in Fig. 1(a). Alternative approaches [8, 33, 41, 60] label question difficulty explicitly, but this introduces subjective judgments, incurs high annotation costs, and ultimately teaches models to follow external labels rather than develop genuine adaptive capabilities, as illustrated in Fig. 1(b).

To deeply analysis the adaptive thinking behavior, we revisit the probability distributions of predicted tokens, where the MLLM estimates the conditional probability of the next token based on previously generated tokens. Intuitively, the transition between modes is governed by special tokens within the output. If

the model confidently identifies a question as simple, it assigns a higher probability to the token representing the non-thinking mode. In contrast, for complex queries, tokens should have higher probabilities to initiate substantive reasoning. Entropy, which quantifies uncertainty in probability distributions, naturally serves as an indicator of the model's confidence in mode selection.

This insight motivates our formalization of adaptive thinking as an entropy-governed curriculum: transitioning from high entropy states with uncertain mode decisions (exploration) to low entropy states with confident decisions (convergence). During early training, high entropy at mode-switching tokens ensures balanced exploration of both thinking and direct-answer modes. During later training, low entropy ensures consistent, confident mode selection aligned with question complexity.

To achieve this goal, we propose AdaThinking-E, a reinforcement learning framework that learns adaptive thinking through targeted entropy regulation at critical decision tokens, as illustrated in Fig. 1(c). The core innovation lies in independently rewarding entropy dynamics at mode-switching positions: we assign higher advantages to high-entropy outputs in exploration stages and low-entropy outputs in convergence stages, enabling a principled transition from uncertainty to consistency. This mechanism encourages intrinsic self-discovery of when to think rather than relying on externally injected mode identifiers. Unlike existing adaptive thinking strategies, we depart from hand-crafted definitions of whether a query necessitates an internal thinking process. Instead, we encourage the model to autonomously explore when to think by regulating the entropy. Furthermore, our decoupled optimization scheme ensures that the actual content of the response focuses exclusively on accuracy, remaining independent of whether the optimal mode selection is achieved.

While our entropy-based framework provides the mechanism for adaptive thinking, effective training requires a dataset that captures the full spectrum of question complexity in document understanding—from simple extraction queries to complex reasoning tasks. To address this gap, we construct AdaThinking-Doc, a document-oriented adaptive thinking dataset designed specifically for training models to develop complexity-aware reasoning capabilities.

In summary, our contributions are three-fold:

- We propose **AdaThinking-E**, a novel RL framework that learns adaptive reasoning through entropy-regulated curriculum learning. Unlike prior work requiring manual mode selection or difficulty labels, our approach enables models to intrinsically discover when to think through entropy rewards at decision points.
- We introduce **AdaThinking-Doc**, the first document understanding dataset designed for adaptive thinking.
- Extensive experiments demonstrates that AdaThinking-E achieves state-of-the-art performance across multiple benchmarks, while matching the effectiveness of thinking models with significantly lower token consumption.

2 Related Work

2.1 MLLMs for Document Understanding

MLLMs demonstrate substantial potential in visual document understanding (VDU). Based on how multimodal features are extracted, these methods can be broadly categorized into OCR-dependent MLLMs [19–24, 38, 39, 42, 47, 48, 64, 68] and OCR-free MLLMs [11, 16, 17, 25, 26, 30, 31, 40, 43, 58, 70, 75, 81, 84, 87, 90]. OCR-dependent MLLMs leverage existing OCR models to extract textual and layout information. In contrast, OCR-free MLLMs enable end-to-end VDU by directly processing document images. To strengthen perceptual and reasoning capabilities, researchers have explored both fine-tuning [17, 25, 28, 30, 31, 36, 40, 43, 58, 70, 81] and reinforcement learning approaches [83]. While prior work [1, 2, 49, 50, 77] has improved reasoning efficiency by compressing output length of reasoning model, these approaches rely on static strategies, limiting their ability to generalize across varying levels of question difficulty. This underscores the need for an adaptive thinking framework capable of employing dynamic reasoning strategies.

2.2 MLLMs with adaptive thinking

Reasoning LLMs [35, 72] typically generate a chain of thought (CoT) with intermediate steps before arriving at the final answer, offering clear benefits for tasks involving complex computation and logical reasoning. However, excessively long CoT significantly increases inference costs and degrades user experience [9, 13]. Recent studies [18, 37, 63, 73, 85, 88] have explored how generation length affects model performance, revealing that the optimal reasoning length varies across tasks. Therefore, the reasoning model should adapt its reasoning depth according to the difficulty of the task.

These methods can be broadly categorized into two types: externally controlled reasoning model [5, 9, 76, 77] and adaptive thinking model [1, 8, 12, 33, 41, 46, 49, 60, 66, 67, 74, 78, 86]. Externally controlled reasoning model switches between short-form response and long-chain reasoning using external mechanisms such as prompt designs. Llama-Nemotron [5] adopts fixed prompt formats like "reasoning on/off" to switch response mode. In contrast, adaptive thinking models autonomously select the appropriate response mode based on the question, without relying on handcrafted prompt templates. Some methods [8, 33, 41, 60] explicitly estimate question difficulty or token budget using another model, while others [12, 74] use problem-solving rate or observed generation length as implicit indicators of difficulty. More recent work has further investigated models capable of switching between thinking/non-thinking modes, most of them [66, 78, 86] adopt a strategy of explicitly training the model to select between reasoning modes. During the cold-start phase, a dual-mode annealing mechanism equips the model with both reasoning and non-reasoning capabilities, while in the RL stage, carefully crafted reward functions guide the decision of whether to engage in reasoning. Heavily influenced by the trainer's subjective intent, such

approaches fail to reflect a genuine understanding and autonomous exploration of mode selection by the adaptive thinking model. However, existing methods fail to examine the core mechanism of mode switching: the probability distribution when outputting mode-switching tokens. Our method introduces the entropy of the model-switching token into the reward function. By steering entropy variations, it guides the thinking/non-thinking mode switching.

3 Methodology

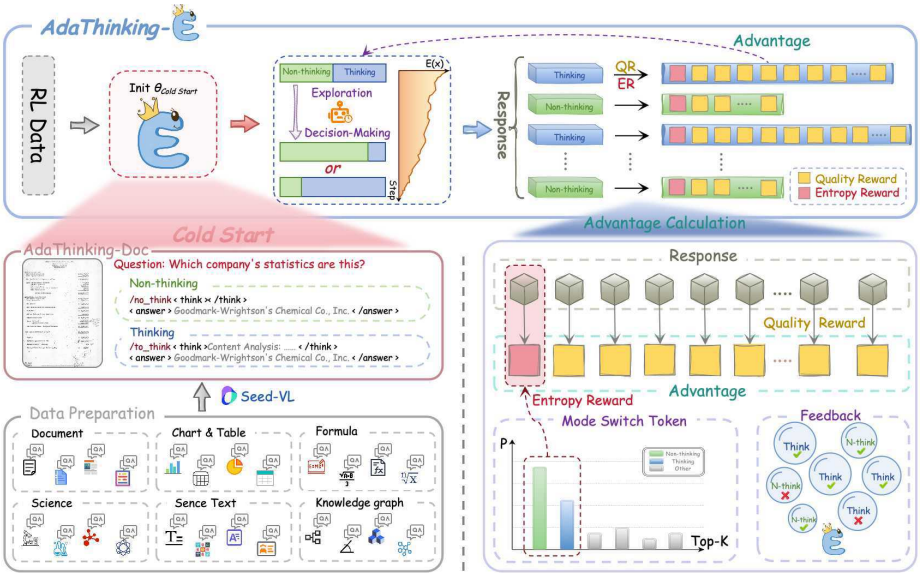

Fig. 2: Overview of **AdaThinking-E**, a reinforcement learning method for adaptive thinking via entropy regulation. The framework has two stages: (1) **Cold Start**: the model is fine-tuned with AdaThinking-Doc to generate outputs in both thinking and non-thinking modes; (2) **RL**: the model optimizes two token types separately using cold-start weights as initialization. The mode-switching token transitions from **high-entropy exploration** to **low-entropy decision-making**, controlling the sampling ratio between modes. The response tokens (remaining tokens) optimize for answer quality. In the bottom-right, the mode-switching token uses **Entropy Reward**, while response tokens use content-focused **Quality Reward**.

In the section, we introduce Adathinking-E, an entropy regulation for refining model behavior, as illustrated in Fig. 2. The proposed method comprises two distinct training stages: (1) Cold-start stage, where MLLM is fine-tuned to support both concise and detailed response modes, and (2) Reinforcement learning stage, where distinct optimization objectives are designed for different types of tokens. Specifically, the mode switch token regulates the sampling ratio across different modes by controlling the variations in its entropy, thereby learning when to think during this process. Meanwhile, the response tokens (i.e.,

the remaining tokens) strive to generate correct answers to the maximum extent possible, regardless of the mode selected by the mode switch token.

3.1 AdaThinking-Doc for Cold-start

To ensure that our proposed model can stably output both thinking and non-thinking modes in the cold start stage, we constructed AdaThinking-Doc dataset (sourced from existing public datasets) composed of these two modes. For data with thinking mode, we select Seed1.5-VL [27] to generate a detailed reasoning process, which is then validated for logical soundness by GPT-4o-mini [34]. Regarding the design of the data format, we define the mode switch tokens as `/no_think` or `/to_think`, representing the non-thinking and thinking modes, respectively. Meanwhile, the response tokens adhere to the widely adopted format: `<think>empty or think content</think><answer></answer>`.

Additionally, we recognized that manually specifying which questions need thinking, or using model distillation to make this decision, would impose external assumptions onto the model being trained. Such predetermined bias would hinder rather than help the model develop genuine insights during the reinforcement learning (RL) stage. Therefore, to ensure unbiased and random sampling during RL training, we generate both thinking mode and non-thinking mode responses for each data point.

3.2 Reinforcement Learning

Following the cold start stage, the model acquires the capability to generate both thinking and non-thinking responses for arbitrary questions, thereby providing sufficiently diverse samples for learning adaptive reasoning abilities during the reinforcement learning stage. To elicit the model's capacity to intelligently determine when to engage deep reasoning based on question complexity, we initially applied GRPO [59] algorithm directly. Unfortunately, as training progressed, we observed a systematic collapse: the model rapidly converged to exclusively generating non-thinking responses for all questions, regardless of their complexity. This behavior stems from a fundamental reward attribution problem in standard RL formulations. Specifically, when rewards are distributed uniformly across all tokens in a response, shorter sequences inherently receive higher per-token rewards than longer ones, creating an implicit bias against reasoning-intensive responses. DAPO [82] attempts to mitigate this length bias through reward balancing mechanisms, fail to capture the nuanced decision-making required for adaptive thinking, as they treat all tokens uniformly without considering their distinct roles in mode selection versus content generation.

To gain a deeper understanding of adaptive thinking behaviors, we revisit existing approaches capable of dual-mode output. Through empirical analysis, we observe that the critical determinant for mode selection lies in the heuristic-capable mode switch token, whereas the function of response tokens is merely to provide the answer deemed correct by the model. This observation implies that

the decision to initiate reasoning depends exclusively on the mode switch token and is independent of other tokens.

Building on this insight, we propose a novel reward attribution framework that explicitly disentangles the contributions of different token types in the response sequence. Specifically, we categorize tokens into two distinct groups: *mode-switching tokens*, which determine the adaptive thinking mode, and *response tokens*, which constitute the actual reasoning content. This decomposition allows us to formulate a dual-objective reward mechanism that optimizes each token type according to its functional role:

$$\mathcal{J}_{AdaThinking-E}(\theta) = \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\left[\mathbb{L}(\theta, E_{i,0}) + \sum_{t=1}^{|o_i|}\mathbb{L}(\theta, \hat{A}_{i,t})\right] \tag{1}$$

where θ denotes model parameters, G represents the number of samples, o_i represents the i-th generated response with length $|o_i|$, and $\mathbb{L}(\theta, \cdot)$ is the loss function, which details refer to supplementary materials (Section 1). Critically, we distinguish between *response tokens* ($t \neq 0$) optimized via advantage estimates $\hat{A}_{i,t}$ and the *mode-switching token* ($t = 0$) optimized via $E_{i,0}$ to separately reward content quality and adaptive thinking mode selection.

3.3 Entropy Reward for Mode-Switching Token

The entropy magnitude reflects the prediction uncertainty; *e.g.*, higher entropy of the mode-switching token indicates greater decision uncertainty between the two modes. Therefore, we regulate the entropy of the mode-switching token to guide the model's transition from an initial high-entropy exploratory stage to a low-entropy decision-making stage.

Entropy in Mode Switch Token. Following the standard GRPO algorithm, for each input question q in the training batch, we sample a group of G responses from the current policy θ, where each sample independently draws the mode-switching token according to its predicted probability distribution. This repeated sampling enables us to observe the model's mode selection behavior across multiple trials, thereby providing a robust estimate of its decision entropy. Let p_{tk} and p_{ntk} denote the probabilities of the mode-switching token belonging to thinking mode and non-thinking mode, respectively. For the i-th sampled response, the entropy reward of its mode-switching token is defined as:

$$H_{i,0} = -\frac{1}{\ln 2}\sum_{m\in\{tk,ntk\}} p_m^{(i)} \ln\left(p_m^{(i)}\right) \tag{2}$$

where $H_{i,0}$ denotes the normalized entropy of the mode-switching token for the i-th sample among the G samples generated for the current question, $p_m^{(i)}$ represents the probability of mode m in sample i.

We formulate the learning process of adaptive thinking as a two-stage curriculum: (1) High-Entropy Reward for Exploration. In this phase, we aim to

maintain a relatively high entropy in mode selection, reflecting uncertainty regarding when to initiate the thinking process. Consequently, the probabilities assigned to the two modes, denoted as p_{tk} and p_{ntk}, should be approximately equivalent. We refrain from imposing manual intervention regardless of the mode selected by the model. This strategy encourages extensive exploration for each query, allowing the model to discover the most suitable mode autonomously. (2) Low-Entropy Reward for Decision-Making. Following the comprehensive evaluation facilitated by the exploration phase, the process gradually transitions towards convergence to reinforce decisive mode selection. In this stage, the model is expected to yield consistent response patterns for specific inputs. Therefore, we incentivize the mode switch token to generate a low-entropy distribution, where the probability of one mode (i.e., p_{tk} or p_{ntk}) significantly dominates the other. This shift from high to low entropy indicates that the model has successfully identified the optimal mode for a given problem type, marking a transition from uncertain exploration to confident specialization.

Dynamic Scoring Mechanism. To facilitate a seamless transition from the exploration phase to the convergence phase, we introduce an adaptive coefficient α_k that varies with the iteration step k. This coefficient ensures that high-entropy rewards are amplified during the initial stages. Conversely, as k increases, the gains from high entropy gradually diminish while those from low entropy increase. The specific formulation is as follows:

$$\alpha_k = \frac{1}{1 + \exp\left[\beta \times \left(\gamma - \frac{k}{K}\right)\right]} \tag{3}$$

where K represents the total number of steps, k denotes the current step index ($0 \leq k \leq K$), γ signifies the proportion of the total steps at which the inflection point of the transition between two entropy states occurs, and β is the smoothness coefficient controlling the steepness of the transition.

Decision Feedback. Besides encouraging the model to explore when to think, we also need to make sure this exploration is reliable. To achieve this, we design a feedback mechanism inspired by confidence-based voting. For each input, we sample multiple outputs under two modes: thinking and non-thinking. We then evaluate which mode tends to give more correct answers, and use this as training feedback. This enables the model learn when thinking is useful by itself, rather than relying on predefined human instructions regarding whether to invoke the thinking process, thus fully unlocking its potential for autonomous exploration. Specifically, for each sampled rollout group, we label every rollout by (1) its mode (thinking/non-thinking) and (2) whether its final answer is correct. From this, we compute for each mode $m \in \{tk, ntk\}$:

- r_m: how often this mode appears in the sampled group (occurrence ratio);
- acc_m: accuracy of this mode within the group.

Subsequently, we divide accuracy into three levels using thresholds ϵ_l and ϵ_h: low, medium, and high. When acc_m in both modes falls within the same tier, an accuracy guidance term $\mathbb{A}$ is triggered to regulate the optimization. Specifically, when accuracy is low in both modes, it guides the model to invoke reasoning;

when both acc_{tk} and acc_{ntk} are medium-tier or both are high-tier, it guides the model to bypass reasoning.

Furthermore, we observe that when the model demonstrates a clear propensity in responding to a sample (defined as $|r_{tk} - r_{ntk}| > \Delta_r$), imbalanced sampling may lead to unreliable comparisons that compromise training stability. For instance, consider a sampling instance where the model achieves 13/15 correct answers ($acc_{tk} = 86.7\%$) in thinking mode versus 1/1 ($acc_{ntk} = 100\%$) in non-thinking mode. Naive $\mathcal{R}_{Ada}$ would steer optimization toward the non-thinking direction, contradicting model's demonstrated correct decisions, which results in training instability or even optimizing collapse. Therefore, we introduce a propensity guidance term $\mathbb{P}$ which guides the model towards its propensity when it demonstrates a clear propensity and both modes achieve high accuracy.

Finally, the weight for mode m is computed as:

$$w_m = \frac{acc_m}{acc_{\neg m}} + \mathbb{A}(acc, m) + \mathbb{P}(acc, r, m, \Delta_r)$$

More details about each term and its value in different cases are provided in Supplementary Materials (Section 2).

Entropy Reward. Finally, the overall entropy reward function is formulated in the following:

$$E_{i,0} = w_m \left[\underbrace{(1-\alpha_k)\, H_{i,0}}_{high-entropy} + \underbrace{\alpha_k\, (1 - H_{i,0})}_{low-entropy} \right] + p_{tk} + p_{ntk} \tag{4}$$

where the additive term $p_t + p_{nt}$ ensures that the cumulative probability of sampling a mode switch token at the first position approaches 1, thereby precluding the possibility of the first token being any other alternative identifier.

3.4 Quality Reward for Response Tokens

While entropy regulation governs the mode-switching decision, it remains agnostic to the quality of reasoning and problem-solving within each mode. Specifically, the entropy reward only determines *whether* thinking mode is activated, but cannot assess *how well* the model reasons in thinking mode or *how accurately* it responds in non-thinking mode. To address this, we introduce a content-focused reward function that directly evaluates the problem-solving effectiveness of the response tokens, thereby ensuring that appropriate mode selection is complemented by high-quality content generation.

Format Reward. The format reward R_f ensures strict adherence to the predefined template structure. This is critical for our entropy-based mode-switching mechanism that relies on specific token positions. We define `isValidFormat`(S) to verify template compliance:

$$R_f = \begin{cases} 1, & \text{if isValidFormat}(S), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Accuracy Reward. The accuracy reward R_{acc} evaluates whether the final answer $\hat{y}$ matches the ground truth y, independent of the reasoning mode:

$$R_{acc} = \begin{cases} 1, & \text{if } \hat{y} = y, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

4 Experiments

4.1 Training Details

Datasets. All data utilized by AdaThinking-Doc are sourced from the training sets of publicly available datasets, encompassing both simple information extraction and complex document reasoning scenarios. Detailed information regarding the data distribution is provided in the Supplementary Material.

Training Details. We used QwenVL-2.5-7B [4] or QwenVL-3-8B [3] as the base model. For the Cold-start stage, we fine-tuned using SWIFT [89] with learning rate 1e-6, 1 epoch, and batch size 4. We then conducted RL training using VeRL [61] with learning rate 1e-6, batch size 32, 16 samples per prompt, and 1 epoch for 2000 steps. All experiments used 8 NVIDIA A100 GPUs.

Evaluation. We evaluated AdaThinking-E against existing MLLMs, including non-thinking (NT), thinking (TK), and adaptive thinking (ATK) models, using VLMEvalKit [15] across multiple benchmarks [32, 44, 51, 53, 54, 62, 71]. GPT-4o-mini [34] served as the evaluator for tasks requiring large model assessment. All ablation studies are conducted using Qwen2.5-VL-7B [4] as the base model.

4.2 Main Results

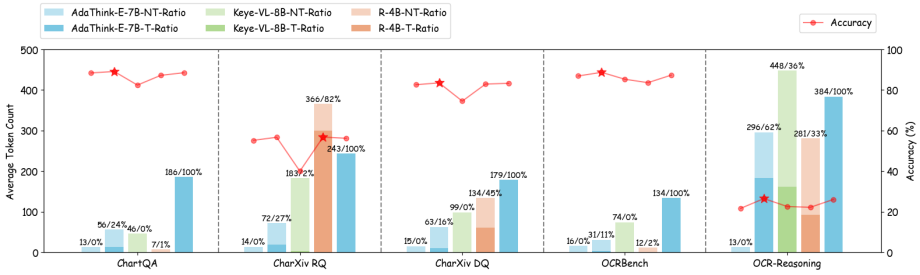

Fig. 3: Comparison of multiple metrics for the adaptive thinking model across document understanding benchmarks. Results of AdaThinking-E include non-thinking, adaptive thinking, and thinking modes. Bar height represents average output tokens, internal color ratio indicates thinking ratio, and the line plot shows accuracy. In the blue bars, 0% and 100% thinking ratios correspond to the non-thinking and thinking modes of AdaThinking-E, respectively.

VQA-based Document Understanding. As shown in Tab. 1, AdaThinking-E is compared against three types of MLLMs with distinct reasoning modes

Table 1: Performance comparison of MLLM on on diverse document benchmarks. "NT", "TK", and "ATK" denote the Non-Thinking, Thinking, and Adaptive Thinking modes, respectively. "*" indicates that the corresponding mode switch token is proactively inserted at the input stage according to the selected mode during the inference phase."†" denotes results reproduced from the original implementation.

Model	Mode	DocVQA	InfoVQA	Text_{Val}	ChartQA	CharXiv_{RQ}	CharXiv_{DQ}	OCR-Bench	OCR-Reasoning
DocOwl-1.5-8B [29]	NT	81.6	50.4	68.8	70.5	-	-	59.9	-
Monkey-10B [40]	NT	66.5	36.1	67.6	65.1	-	-	51.4	-
TextMonkey-8B [45]	NT	73.0	28.6	65.6	66.9	-	-	56.1	-
MiniMonkey-2B [31]	NT	87.4	60.1	75.7	76.5	-	-	80.2	-
HRVDA-7B [43]	NT	72.1	43.5	77.3	67.6	-	-	-	-
InternVL2.5-7B [10]	NT	93.0	77.6	79.1	84.8	32.9	68.6	82.2	-
InternVL3-8B [91]	NT	92.7	76.8	80.2	86.6	37.6	73.6	88.0	11.5
QwenVL2.5-8B [4]	NT	95.7	82.6	84.9	87.3	42.5	73.9	86.4	15.7
QwenVL3-8B [3]	NT	96.1	83.1	81.9	89.6	46.4	83.0	89.6	18.5
TextHawk2-7B [84]	NT	89.6	67.8	75.1	81.4	-	-	78.4	-
Marten-7B [70]	NT	92.0	75.2	74.4	81.7	-	-	82.0	-
AlignVLM-8B [52]	NT	81.2	53.8	64.6	75.0	-	-	-	-
TokenVL-7B [25]	NT	94.2	76.5	79.9	86.6	-	-	86.0	14.3
DocMark-2B [75]	NT	-	-	74.8	79.8	-	-	81.3	7.4
Docopilot-8B [16]	NT	92.0	73.3	-	83.3	-	-	-	11.6
OpenVLThinker-7B† [14]	TK	94.6	76.5	81.8	82.7	42.4	66.2	81.4	23.3
R1-Onevision-7B† [80]	TK	92.4	76.6	75.4	82.3	37.1	54.9	-	21.2
VLAA-Thinker-7B† [6]	TK	95.7	80.1	82.9	87.3	38.9	73.3	83.0	14.4
Kimi-VL-A3B-Thinking [65]	TK	-	-	-	-	47.7	75.4	82.5	20.5
VL-Rethinker-7B† [69]	TK	95.9	78.9	82.7	87.1	43.3	64.9	86.4	14.6
DocThinker-7B [83]	TK	-	-	83.6	-	-	-	-	-
QwenVL3-8B [3]	TK	95.3	86.0	78.7	88.6	53.0	85.9	81.9	48.4
R-4B† [79]	ATK	94.1	67.0	76.6	87.2	56.8	82.9	83.6	22.2
Keye-VL-8B† [66]	ATK	90.5	62.6	78.6	82.4	40.0	74.5	85.3	22.6
Mimo-VL-7B† [76]	ATK	96.1	79.7	80.6	87.4	56.5	**86.8**	86.6	-
ARES-7B† [7]	ATK	92.1	72.2	80.5	88.2	51.0	79.3	83.1	23.4
AdaThink-E-7B-Qwen2.5*	NT	96.3	82.8	84.9	88.4	55.2	82.6	86.9	21.7
AdaThink-E-7B-Qwen2.5*	TK	95.1	80.8	84.5	88.5	56.2	83.3	87.3	26.1
AdaThink-E-7B-Qwen2.5	ATK	96.3	82.9	**85.3**	89.1	56.7	83.5	88.7	26.5
AdaThink-E-8B-Qwen3*	NT	96.2	84.7	82.6	89.3	54.2	83.9	88.9	31.2
AdaThink-E-8B-Qwen3*	TK	95.7	86.2	80.9	89.9	56.9	86.3	87.7	49.4
AdaThink-E-8B-Qwen3	ATK	**96.4**	**86.3**	82.8	**90.1**	**57.3**	86.6	**89.8**	**49.7**

across multiple document understanding benchmarks, consistently achieving superior performance. On chart-based datasets requiring analytical computation, such as CharXiv, models equipped with a thinking mode (TK) outperform those with a Non-Thinking mode. However, in information extraction scenarios, the advantage of the thinking mode becomes less pronounced and can even lead to performance degradation due to overthinking-induced hallucinations.

Our proposed AdaThinking-E achieves near SOTA performance across both information extraction and analytical computation tasks. AdaThinking-E's adaptive thinking surpasses the baseline Qwen2.5-VL by 0.6% on DocVQA, and achieves improvements of 1.8% and 10.8% on ChartQA and OCR-Reasoning, respectively. Even compared to R-4B, which also features adaptive thinking, AdaThinking-E-Qwen2.5 delivers gains of 1.9% and 4.3% on these two datasets. These results clearly demonstrate AdaThinking-E-Qwen2.5's superior adaptive thinking capabilities in document understanding tasks. AdaThinking-E-Qwen3, trained on Qwen3-VL-Thinking, outperforms Qwen3-VL-Thinking across various benchmarks and even surpasses the higher-performing Qwen3-VL-Instruct on certain tasks. Additionally, in the last six rows of Tab. 1, we compare the performance of AdaThinking-E under different modes. The thinking and non-Thinking modes are actively switched via mode switch token. It is evident that

the adaptive thinking mode consistently matches or exceeds the performance of each standalone mode across various benchmarks, further validating the adaptive thinking capabilities of AdaThinking-E.

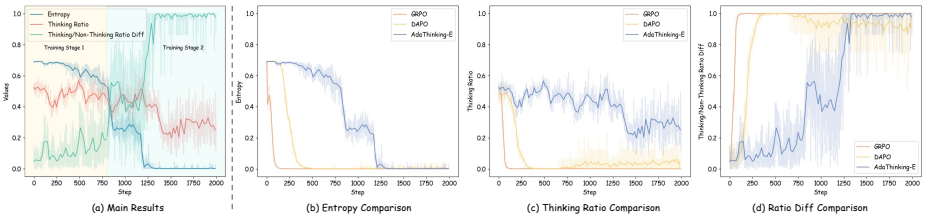

Fig. 4: Illustrate the trends in entropy and thinking ratio changes during the training phase. (a) displays the variation curves of three metrics of AdaThinking-E as the training steps increase. Specifically, "Entropy" represents the average entropy of mode switch tokens within a step, "Thinking Ratio" indicates the proportion of thinking samples within a step, and "Thinking/Non-Thinking Ratio" denotes the average difference in the ratio of thinking and non-thinking modes within the group. (b), (c), and (d) showcase the comparison of the three metrics between AdaThinking-E, GRPO, and DAPO under the same reward conditions

Token Count and Thinking Ratio. As illustrated in Fig. 3, we analyzed the average output length, thinking ratio, and accuracy of each query under different modes in AdaThinking-E-Qwen2.5, and compared these metrics with advanced adaptive thinking MLLMs. To ensure fairness in output token length, we uniformly added the phrase "Answer the question using a single word or phrase." in each query, ensuring that the token length of the answer part remains relatively balanced apart from the thinking content. On the CharXiv$_{RQ}$ dataset, AdaThinking-E achieved a comparable level to R-4B with only a 27% thinking ratio (72 tokens) compared to R-4B's 82% thinking rate (366 tokens). Moreover, on CharXiv$_{DQ}$, AdaThinking-E outperformed R-4B by 0.6% in performance with a lower thinking rate (11% compared to R-4B's 45%). This demonstrates that AdaThinking-E exhibits superior document understanding capabilities compared to other adaptive thinking models.Additionally, on the OCR-Reasoning benchmark, AdaThinking-E's thinking ratio increased to 62%, with an accuracy rate of 26.5%—0.4% higher than its own thinking mode's 26.1%—while the average token count decreased by 88. In contrast, R-4B and Keye-VL showed only 33% and 38% thinking ratios, respectively, and their performance decreased by 4.3% and 3.9% compared to AdaThink-E due to excessive non-thinking decisions. These findings collectively prove AdaThinking-E has a more accurate understanding of document complexity, maintaining performance while ensuring efficiency, and shows robust capabilities.

4.3 Training Analyze

To further analyze how entropy influences the learning of mode-switching strategies in adaptive thinking models, in Fig. 4, we visualized the indicators related to

entropy and thinking ratio during the RL stage. Fig. 4(c) illustrates the changes in thinking ratios of three strategies. Both GRPO and DAPO quickly tend towards the non-thinking mode during training, resulting in mode collapse.

Fig. 4(a) demonstrates how AdaThinking-E effectively addresses the mode collapse issue by regulating entropy during training. AdaThinking-E decouples mode switch token from response tokens, linking the gradient of mode switch token solely to entropy levels, preventing entropy rewards from affecting response tokens. During the exploratory stage, entropy remains at a high level, ensuring a balanced sampling with low differences in the number of thinking and non-thinking modes within the group, which provides high-quality samples for optimizing mode-switching strategies. As training progresses into the decision-making stage, mode collapse does not occur, and the proportion of thinking and non-thinking within the step remains relatively stable.

4.4 Ablation Study

Hyperparameter Setting in Decision Feedback. We conduct comprehensive ablation studies on three key hyperparameters in decision feedback. As shown in Tab. 2, two benchmarks of complementary difficulty levels, ChartQA (easier) and OCR-Reasoning (more challenging) are selected to evaluate scenario specific impacts. In simple scenarios, a lower ϵ_h increases the activation probability of $\mathbb{A}$, thereby amplifying the influence of $\mathbb{P}$. Higher Δ_r impose stricter criteria for determining model propensity, which impedes activation of $\mathbb{P}$, resulting in a pronounced reduction in the model's reasoning frequency. In challenging scenarios, the model's lower accuracy impedes activation of $\mathbb{P}$, while higher ϵ_l settings facilitate triggering of $\mathbb{A}$, enhancing both thinking frequency and accuracy. Furthermore, we investigate the performance when relying solely on *acc* as decision feedback, omitting $\mathbb{P}$ and $\mathbb{A}$. While this configuration achieves comparable scores in terms of performance, its thinking ratio is significantly higher than the results obtained with $\mathbb{P}$ and $\mathbb{A}$. This suggests that $\mathbb{P}$ and $\mathbb{A}$ effectively facilitate a deeper exploration of the transition from thinking to non-thinking modes.

Table 2: Hyperparameter Ablation in Decision Feedback. "Think (%)" represents the ratio of instances where the thinking mode is activated. "β" serves as the smoothing coefficient. "-/-" indicates that the parameter is not used.

method	ϵ_l	ϵ_h	Δ_r	**ChartQA**		**OCR-Reasoning**	
				Acc	Think(%)	Acc	Think(%)
w/o $\mathbb{A}$&$\mathbb{P}$	-/-	-/-	-/-	88.1	39.9	25.7	83.4
w $\mathbb{A}$&$\mathbb{P}$	0.2	0.8	0.625	<u>88.6</u>	8.7	<u>25.6</u>	55.1
			0.5	88.5	12.5	25.5	54.7
			0.375	**89.0**	23.7	**25.8**	56.6
w $\mathbb{A}$&$\mathbb{P}$	0.3	0.7	0.625	88.4	3.4	25.9	59.7
			0.5	<u>88.9</u>	13.1	<u>26.2</u>	60.3
			0.375	**89.1**	24.3	**26.5**	61.9

Entropy Regulation. Tab. 3 compares the performance and thinking ratio on document benchmarks with various parameter configurations. The first row of the table represents the cold-start baseline. Compared with Qwen2.5-VL, the model fine-tuned using AdaThinking-Doc during the cold-start phase achieves performance gains of 0.6% on ChartQA and 5.9% on OCR-Reasoning, respectively, demonstrating its effectiveness. In rows 2–3, we simulate scenarios that encourage either high or low entropy throughout the entire process. It can be observed that maintaining consistently high entropy for mode switch tokens leads to unstable thinking ratios and misallocates challenging queries to the non-thinking mode, which results in a performance drop of approximately 1.5%. Conversely, in low-entropy states, the mode-switching tokens gain an overwhelming bias toward non-thinking patterns, causing the model to converge prematurely to a non-thinking state. This leads to mode collapse and a significant reduction in the overall score by about 5%.

We explore the configurations of the inflection point γ and the smoothing coefficient β during the exploration and convergence phases. When γ is larger, the model accuracy is nearly consistent with that when γ is smaller, but the overall thinking ratio is 18% higher. This indicates that during the exploratory stage, due to balanced sampling, the progress of exploring the non-thinking mode is slowed. An excessively small smoothing coefficient renders the model insensitive to entropy regulation, leading to suboptimal performance. For a more comprehensive analysis, please refer to the Supplementary Material.

Table 3: Performance comparison of entropy regulation under different parameter configurations. "γ" represents the inflection point of the transitionand, and "β" serves as the smoothing coefficient. "-/-" indicates that the parameter is not used.

stage	β	γ	**ChartQA**		**OCR-Reasoning**	
			Acc(%)	Think(%)	Acc(%)	Think(%)
Cold-start	-/-	-/-	87.6	9.2	21.6	10.4
RL	-/-	1.0	88.1	75.6	23.7	48.4
		0.0	87.8	0.0	18.7	0.0
RL	5	0.6	<u>88.4</u>	50.7	25.5	63.3
		0.5	88.3	48.3	<u>25.7</u>	67.2
		0.4	**88.6**	40.6	**26.1**	59.7
RL	10	0.6	<u>88.9</u>	39.3	<u>26.7</u>	83.8
		0.5	88.8	30.6	**26.9**	75.6
		0.4	**89.1**	24.3	26.5	61.9

5 Conclusion

In this work, we introduce **AdaThinking-E**, a reinforcement learning framework that enables multimodal large language models to dynamically switch between thinking and non-thinking modes according to task complexity. By leveraging one-token entropy regulation, our method incorporates an entropy-based reward

to guide the transition from exploratory to convergent thinking strategies. This adaptive thinking allows the model to intrinsically determine when to think without relying on external difficulty annotations, thereby improving efficiency on simple tasks while preserving accuracy on complex ones. Extensive experiments across diverse document understanding benchmarks validate the effectiveness of our approach, demonstrating that AdaThinking-E achieves a favorable balance between computational efficiency and reasoning performance.

Acknowledgements

This work was supported by NSFC 62322604 and NSFC 62576207.

References

1. Aggarwal, P., Welleck, S.: L1: Controlling how long a reasoning model thinks with reinforcement learning. arXiv preprint arXiv:2503.04697 (2025)
2. Aytes, S.A., Baek, J., Hwang, S.J.: Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. arXiv preprint arXiv:2503.05179 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Bercovich, A., Levy, I., Golan, I., Dabbah, M., El-Yaniv, R., Puny, O., Galil, I., Moshe, Z., Ronen, T., Nabwani, N., et al.: Llama-nemotron: Efficient reasoning models. arXiv preprint arXiv:2505.00949 (2025)
6. Chen, H., Tu, H., Wang, F., Liu, H., Tang, X., Du, X., Zhou, Y., Xie, C.: Sft or rl? an early investigation into training r1-like reasoning large vision-language models (2025), https://arxiv.org/abs/2504.11468
7. Chen, S., Guo, Y., Ye, Y., Huang, S., Hu, W., Li, H., Zhang, M., Chen, J., Guo, S., Peng, N.: Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. arXiv preprint arXiv:2510.08457 (2025)
8. Chen, W., Yuan, J., Jin, T., Ding, N., Chen, H., Liu, Z., Sun, M.: The overthinker's diet: Cutting token calories with difficulty-aware training. arXiv preprint arXiv:2505.19217 (2025)
9. Chen, X., Xu, J., Liang, T., He, Z., Pang, J., Yu, D., Song, L., Liu, Q., Zhou, M., Zhang, Z., et al.: Do not think that much for 2+ 3=? on the overthinking of o1-like llms. arXiv preprint arXiv:2412.21187 (2024)
10. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
11. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
12. Cheng, X., Li, J., Zhang, Z., Tang, X., Zhao, W.X., Kong, X., Zhang, Z.: Incentivizing dual process thinking for efficient large language model reasoning. arXiv preprint arXiv:2505.16315 (2025)
13. Cuadron, A., Li, D., Ma, W., Wang, X., Wang, Y., Zhuang, S., Liu, S., Schroeder, L.G., Xia, T., Mao, H., et al.: The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. arXiv preprint arXiv:2502.08235 (2025)
14. Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv preprint arXiv:2503.17352 (2025)
15. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11198–11201 (2024)

16. Duan, Y., Chen, Z., Hu, Y., Wang, W., Ye, S., Shi, B., Lu, L., Hou, Q., Lu, T., Li, H., et al.: Docopilot: Improving multimodal models for document-level understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4026–4037 (2025)
17. Feng, H., Liu, Q., Liu, H., Tang, J., Zhou, W., Li, H., Huang, C.: Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. Science China Information Sciences **67**(12), 220106 (2024)
18. Ghosal, S.S., Chakraborty, S., Reddy, A., Lu, Y., Wang, M., Manocha, D., Huang, F., Ghavamzadeh, M., Bedi, A.S.: Does thinking more always help? understanding test-time scaling in reasoning models. arXiv preprint arXiv:2506.04210 (2025)
19. Guan, T., Gu, C., Lu, C., Tu, J., Feng, Q., Wu, K., Guan, X.: Industrial scene text detection with refined feature-attentive network. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6073–6085 (2022)
20. Guan, T., Gu, C., Tu, J., Yang, X., Feng, Q., Zhao, Y., Shen, W.: Self-supervised implicit glyph attention for text recognition. In: CVPR. pp. 15285–15294 (2023)
21. Guan, T., Lin, C., Shen, W., Yang, X.: Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In: European Conference on Computer Vision. pp. 130–147. Springer (2025)
22. Guan, T., Shen, W., Yang, X.: CCDPlus: Towards accurate character to character distillation for text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
23. Guan, T., Shen, W., Yang, X., Feng, Q., Jiang, Z., Yang, X.: Self-supervised character-to-character distillation for text recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19473–19484 (2023)
24. Guan, T., Shen, W., Yang, X., Wang, X., Yang, X.: Bridging synthetic and real worlds for pre-training scene text detectors. In: European Conference on Computer Vision. pp. 428–446. Springer (2024)
25. Guan, T., Wang, Z., Fu, P., Guo, Z., Shen, W., Zhou, K., Yue, T., Duan, C., Sun, H., Jiang, Q., et al.: A token-level text image foundation model for document understanding. arXiv preprint arXiv:2503.02304 (2025)
26. Guan, T., Yang, Z., Wan, J., Yang, M., Guo, Z., Hu, Z., Luo, R., Chen, R., Jiang, S., Wang, P., et al.: Codepercept: Code-grounded visual stem perception for mllms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33542–33552 (2026)
27. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
28. Guo, Z., Duan, C., Guan, T., Wang, Z., Zhou, K., Yan, P.: Vitexqa: A multi-frame temporal perception dataset for video text question answering (2026), `https://arxiv.org/abs/2606.24602`
29. Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3096–3120 (2024)
30. Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5817–5834 (2025)
31. Huang, M., Liu, Y., Liang, D., Jin, L., Bai, X.: Mini-monkey: Alleviate the sawtooth effect by multi-scale adaptive cropping. arXiv e-prints pp. arXiv–2408 (2024)

32. Huang, M., Shi, Y., Peng, D., Lai, S., Xie, Z., Jin, L.: Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. arXiv preprint arXiv:2505.17163 (2025)
33. Huang, S., Wang, H., Zhong, W., Su, Z., Feng, J., Cao, B., Fung, Y.R.: Adactrl: Towards adaptive and controllable reasoning via difficulty-aware budgeting. arXiv preprint arXiv:2505.18822 (2025)
34. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
35. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
36. Jiang, L., Wu, X., Huang, S., Dong, Q., Chi, Z., Dong, L., Zhang, X., Lv, T., Cui, L., Wei, F.: Think only when you need with large hybrid-reasoning models (2025), https://arxiv.org/abs/2505.14631
37. Jin, Z., Li, X., Ji, Y., Peng, C., Liu, Z., Shi, Q., Yan, Y., Wang, S., Peng, F., Yu, G.: Recut: Balancing reasoning length and accuracy in llms via stepwise trails and preference optimization. arXiv preprint arXiv:2506.10822 (2025)
38. Kim, G., Lee, H., Kim, D., Jung, H., Park, S., Kim, Y., Yun, S., Kil, T., Lee, B., Park, S.: Visually-situated natural language understanding with contrastive reading model and frozen large language models. arXiv preprint arXiv:2305.15080 (2023)
39. Lee, B.K., Park, B., Won Kim, C., Man Ro, Y.: Moai: Mixture of all intelligence for large language and vision models. In: European Conference on Computer Vision. pp. 273–302. Springer (2024)
40. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26763–26773 (2024)
41. Li, Z., Dong, Q., Ma, J., Zhang, D., Jia, K., Sui, Z.: Selfbudgeter: Adaptive token allocation for efficient llm reasoning. arXiv preprint arXiv:2505.11274 (2025)
42. Liao, W., Wang, J., Li, H., Wang, C., Huang, J., Jin, L.: Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4038–4049 (2025)
43. Liu, C., Yin, K., Cao, H., Jiang, X., Li, X., Liu, Y., Jiang, D., Sun, X., Xu, L.: Hrvda: High-resolution visual document assistant. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15534–15545 (2024)
44. Liu, Y., Li, Z., Yang, B., Li, C., Yin, X., Liu, C.l., Jin, L., Bai, X.: On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895 (2023)
45. Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., Bai, X.: Textmonkey: An ocr-free large multimodal model for understanding document. arXiv preprint arXiv:2403.04473 (2024)
46. Lou, C., Sun, Z., Liang, X., Qu, M., Shen, W., Wang, W., Li, Y., Yang, Q., Wu, S.: Adacot: Pareto-optimal adaptive chain-of-thought triggering via reinforcement learning. arXiv preprint arXiv:2505.11896 (2025)
47. Lu, J., Yu, H., Wang, Y., Ye, Y., Tang, J., Yang, Z., Wu, B., Liu, Q., Feng, H., Wang, H., et al.: A bounding box is worth one token-interleaving layout and text in a large language model for document understanding. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 7252–7273 (2025)

48. Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., Yao, C.: Layoutllm: Layout instruction tuning with large language models for document understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15630–15640 (2024)
49. Luo, H., Shen, L., He, H., Wang, Y., Liu, S., Li, W., Tan, N., Cao, X., Tao, D.: O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. arXiv preprint arXiv:2501.12570 (2025)
50. Ma, X., Wan, G., Yu, R., Fang, G., Wang, X.: Cot-valve: Length-compressible chain-of-thought tuning. arXiv preprint arXiv:2502.09601 (2025)
51. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
52. Masry, A., Rodriguez, J., Zhang, T., Wang, S., Wang, C., Feizi, A., Kalkunte Suresh, A., Puri, A., Jian, X., Noel, P.A., et al.: Alignvlm: Bridging vision and language latent spaces for multimodal document understanding. Advances in Neural Information Processing Systems **38**, 145684–145710 (2026)
53. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
54. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
55. OpenAI: Chatgpt (2023), available at: `https://openai.com/chatgpt`
56. OpenAI: Gpt-4 (2023), available at: `https://openai.com/gpt-4`
57. OpenAI: Gpt-4v(ision) system card (2023), available at: `https://cdn.openai.com/papers/GPTV_System_Card.pdf`
58. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. CoRR (2024)
59. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024), `https://arxiv.org/abs/2402.03300`
60. Shen, Y., Zhang, J., Huang, J., Shi, S., Zhang, W., Yan, J., Wang, N., Wang, K., Liu, Z., Lian, S.: Dast: Difficulty-adaptive slow-thinking for large reasoning models. arXiv preprint arXiv:2503.04472 (2025)
61. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv:2409.19256 (2024)
62. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
63. Sui, Y., Chuang, Y.N., Wang, G., Zhang, J., Zhang, T., Yuan, J., Liu, H., Wen, A., Zhong, S., Zou, N., et al.: Stop overthinking: A survey on efficient reasoning for large language models. arXiv preprint arXiv:2503.16419 (2025)
64. Tanaka, R., Iki, T., Nishida, K., Saito, K., Suzuki, J.: Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 19071–19079 (2024)

65. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
66. Team, K.K.: Kwai keye-vl technical report (2025), https://arxiv.org/abs/2507.01949
67. Tu, S., Lin, J., Zhang, Q., Tian, X., Li, L., Lan, X., Zhao, D.: Learning when to think: Shaping adaptive reasoning in r1-style models via multi-stage rl. arXiv preprint arXiv:2505.10832 (2025)
68. Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., Liu, X.: Docllm: A layout-aware generative language model for multimodal document understanding. In: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers). pp. 8529–8548 (2024)
69. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837 (2025)
70. Wang, Z., Guan, T., Fu, P., Duan, C., Jiang, Q., Guo, Z., Guo, S., Luo, J., Shen, W., Yang, X.: Marten: Visual question answering with mask generation for multi-modal document understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14460–14471 (2025)
71. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., et al.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems **37**, 113569–113697 (2024)
72. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
73. Wu, Y., Wang, Y., Ye, Z., Du, T., Jegelka, S., Wang, Y.: When more is less: Understanding chain-of-thought length in llms. arXiv preprint arXiv:2502.07266 (2025)
74. Xiang, V., Blagden, C., Rafailov, R., Lile, N., Truong, S., Finn, C., Haber, N.: Just enough thinking: Efficient reasoning with adaptive length penalties reinforcement learning. arXiv preprint arXiv:2506.05256 (2025)
75. Xiao, H., Xie, Y., Tan, G., Chen, Y., Hu, R., Wang, K., Zhou, A., Li, H., Shao, H., Lu, X., et al.: Adaptive markup language generation for contextually-grounded visual document understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29558–29568 (2025)
76. Xiaomi, L.C.T.: Mimo-vl technical report (2025), https://arxiv.org/abs/2506.03569
77. Xu, S., Xie, W., Zhao, L., He, P.: Chain of draft: Thinking faster by writing less. arXiv preprint arXiv:2502.18600 (2025)
78. Yang, Q., Ni, B., Xiang, S., Hu, H., Peng, H., Jiang, J.: R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning. arXiv preprint arXiv:2508.21113 (2025)
79. Yang, Q., Ni, B., Xiang, S., Hu, H., Peng, H., Jiang, J.: R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning (2025), https://arxiv.org/abs/2508.21113
80. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)

81. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., et al.: Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 2841–2858 (2023)
82. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.Y., Zhang, Y.Q., Yan, L., Qiao, M., Wu, Y., Wang, M.: Dapo: An open-source llm reinforcement learning system at scale (2025), https://arxiv.org/abs/2503.14476
83. Yu, W., Yang, Z., Liu, Y., Bai, X.: Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 837–847 (2025)
84. Yu, Y.Q., Liao, M., Zhang, J., Wu, J.: Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. arXiv preprint arXiv:2410.05261 (2024)
85. Zeng, Z., Huang, X., Li, B., Zhang, H., Deng, Z.: Done is better than perfect: Unlocking efficient reasoning by structured multi-turn decomposition. arXiv preprint arXiv:2505.19788 (2025)
86. Zhang, J., Lin, N., Hou, L., Feng, L., Li, J.: Adaptthink: Reasoning models can learn when to think, 2025. URL https://arxiv. org/abs/2505.13417
87. Zhang, R., Lyu, Y., Shao, R., Chen, G., Guan, W., Nie, L.: Token-level correlation-guided compression for efficient multimodal document understanding. arXiv preprint arXiv:2407.14439 (2024)
88. Zhao, W., Sui, X., Guo, J., Hu, Y., Deng, Y., Zhao, Y., Qin, B., Che, W., Chua, T.S., Liu, T.: Trade-offs in large reasoning models: An empirical analysis of deliberative and adaptive reasoning over foundational capabilities. arXiv preprint arXiv:2503.17979 (2025)
89. Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., Zhou, W., Chen, Y.: Swift:a scalable lightweight infrastructure for fine-tuning (2024), https://arxiv.org/abs/2408.05517
90. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
91. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)