CONFIDENT THINKING: MULTI-LAYER ENTROPY-GUIDED LATENT REASONING FOR MULTIMODAL LONG-CONTEXT RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal long-context retrieval-augmented generation (MLRAG) is increasingly important for knowledge-intensive question answering, as it enables models to retrieve and reason over extensive textual and visual evidence. However, during answer generation, multimodal large language models remain prone to hallucination when synthesizing retrieved cross-modal evidence, particularly when the evidence is heterogeneous, redundant, or conflicting. Existing methods commonly employ explicit chain-of-thought (CoT) to facilitate evidence reasoning, but its token-by-token generation forces the model to commit to discrete intermediate states, allowing unreliable predictions to propagate through subsequent steps and ultimately yield hallucinated answers. To address this problem, we propose **Confident Thinking**, a confidence-guided framework that switches from explicit to latent reasoning upon detecting genuine reasoning uncertainty, preserving candidate semantics in continuous space while incorporating retrieved evidence to guide subsequent updates. First, we introduce a multi-layer entropy consensus mechanism that evaluates predictive uncertainty across the last few Transformer layers and triggers latent reasoning only when all these layers exhibit high entropy, reducing unnecessary switches caused by final-layer-specific ambiguity. Second, we develop an evidence-calibrated soft-thinking mechanism that dynamically selects question- and state-relevant multimodal evidence and injects it into soft tokens through uncertainty-aware gating, constraining recursive latent updates to mitigate reasoning drift and error accumulation. Extensive experiments on multiple multimodal long-context question-answering benchmarks demonstrate that Confident Thinking consistently improves answer accuracy and overall performance across different settings.

1 INTRODUCTION

With advances in multimodal retrieval and generation, multimodal long-context retrieval-augmented generation has become increasingly important for knowledge-intensive question answering (Chen et al., 2022; Yu et al., 2024). By providing multimodal large language models with external retrieved textual and visual results, this paradigm supports answering questions whose required knowledge extends beyond model parameters. (Ma et al., 2024a; Dong et al., 2025) It requires models to integrate information distributed across lengthy contexts, where text, images, tables, and charts may contribute complementary retrieved results (Ma et al., 2024b; Chia et al., 2025). Such capability is essential for understanding complex documents and synthesizing answers from heterogeneous knowledge sources (Wu et al., 2025; Yu et al., 2023).

To support the synthesis of retrieved multimodal results, reasoning-based approaches employ explicit chain-of-thought (CoT), generating intermediate natural-language steps before producing an answer (Wei et al., 2022; Sun et al., 2025b). Although effective in decomposing complex questions, explicit CoT requires the model to select a discrete token at each step and propagate its embedding into subsequent reasoning (Tan et al., 2025; Su et al., 2025). In multimodal long-context RAG, retrieved results can be heterogeneous, redundant, or conflicting, making intermediate interpretations uncertain (Asai et al., 2024; Niu et al., 2024). Committing to an unreliable interpretation, such as an incorrect association between a textual statement and a visual observation, can bias subsequent

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.

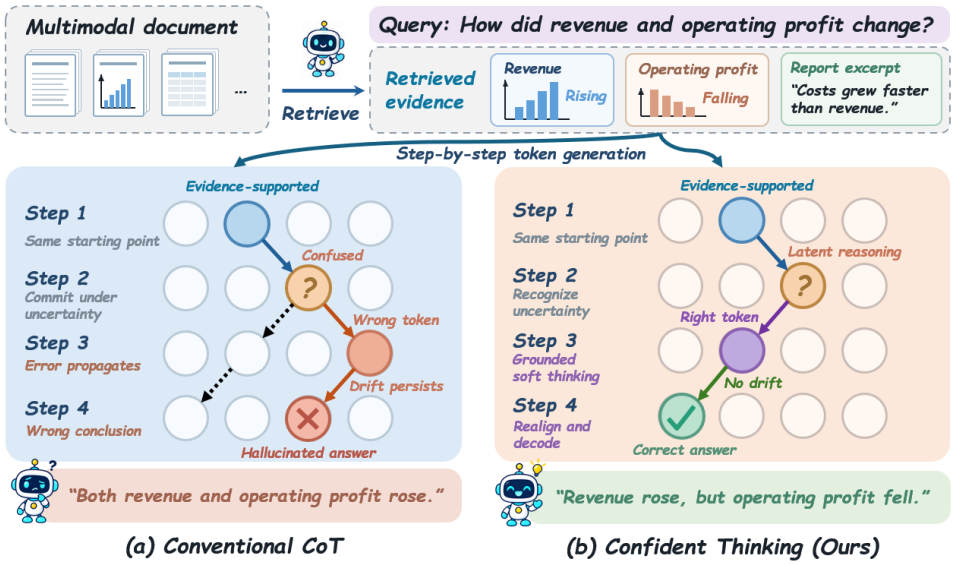


Figure 1: **Motivation for Confident Thinking.** (a) Conventional CoT commits to discrete tokens at each step, allowing unreliable intermediate predictions to propagate under uncertainty and lead to hallucinations. (b) Confident Thinking introduces latent reasoning to retain candidate semantics in continuous space and defer premature discrete commitments, helping mitigate error propagation and reduce hallucinations.

steps and allow errors to accumulate along the reasoning chain. Consequently, access to relevant retrieved results alone does not ensure faithful answers, as hallucinations can arise while the model interprets and combines that retrieved results.

This observation motivates delaying discrete commitments when the model encounters uncertainty during reasoning over retrieved results. Recent latent reasoning methods propagate continuous representations, allowing candidate continuations to contribute to subsequent computation before selecting a single token embedding (Hao et al., 2025; Zhang et al., 2025), as illustrated in Fig. 1. Such continuous propagation naturally allows the model to defer premature discrete commitments when uncertainty arises during reasoning over retrieved results. However, realizing this potential in multimodal long-context RAG requires addressing two key challenges. First, uncertainty-based switching methods use the entropy of the model's output distribution to decide when to invoke latent reasoning (Lv et al., 2026). Yet final-layer uncertainty can reflect local prediction perturbations, including those associated with alignment preferences, while a nearby layer still yields a more concentrated distribution (Zhang et al., 2026). More broadly, layer-wise decoding studies have shown that useful predictive information can emerge before the final Transformer layer (Chuang et al., 2024; Zhang et al., 2026). This suggests that final-layer entropy alone may trigger unnecessary switches without establishing that uncertainty persists across the model's late-layer predictions. We therefore address **Challenge 1:** ***How to distinguish final-layer-specific uncertainty from persistent cross-layer uncertainty and determine when latent reasoning should be activated?*** Furthermore, delaying discrete commitment does not itself establish retrieved results grounding. Soft-thinking methods recursively construct input embeddings from the model's own predictive distributions, without directly incorporating retrieved results into those soft embeddings. When predictions become unreliable, these updates can carry unsupported interpretations forward even in continuous space, while the retrieved results needed to resolve uncertainty may change as reasoning progresses. Prior work on adaptive retrieval similarly shows that a model's information needs can evolve during generation, motivating dynamic decisions about when and what to retrieve (Su et al., 2024; Jeong et al., 2024). Therefore, we aim to address **Challenge 2:** ***How to dynamically select multimodal retrieved results according to the question and current reasoning state to calibrate soft token feedback and limit reasoning drift?***

To address these challenges, we propose **Confident Thinking**, a confidence-guided, training-free decoding framework for multimodal long-context RAG. Confident Thinking combines selective la-

tent reasoning with direct evidence calibration to support reliable answer generation, as summarized in Fig. 1. To tackle **Challenge 1**, we introduce a Multi-Layer Entropy Consensus (MEC) mechanism that jointly evaluates predictive uncertainty across the last few Transformer layers. MEC routes discrete predictions to a sufficiently lower-entropy candidate layer when final-layer uncertainty is recoverable, and activates soft thinking only when all candidate layers remain highly uncertain over consecutive decoding steps. This criterion operationally distinguishes persistent confusion from final-layer-specific pseudo-confusion. To tackle **Challenge 2**, we develop Evidence-Calibrated Soft Thinking (ECST), which temporarily replaces discrete token feedback with probability-weighted token embeddings calibrated by a dynamically updated evidence anchor. At each active soft-thinking step, ECST jointly evaluates evidence relevance to the question and the current reasoning state, reselects supporting evidence to update the anchor, and injects it into soft token feedback through an uncertainty-aware gate. Soft-thinking phases have a bounded duration and terminate early when a low-entropy prediction re-emerges, final answers use conventional discrete decoding. Together, these mechanisms delay premature commitments at uncertain reasoning steps and guide subsequent computation with retrieved evidence.

The main contributions of this work are summarized as follows:

- We propose **Confident Thinking**, a training-free decoding framework that combines confidence-guided reasoning control and dynamic evidence calibration for multimodal long-context RAG, enabling adaptive transitions between discrete token feedback and evidence-calibrated soft propagation.
- We develop a **Multi-Layer Entropy Consensus** mechanism that recovers discrete predictions from lower-entropy candidate layers and activates soft thinking under persistent cross-layer uncertainty, thereby filtering final-layer-specific pseudo-confusion and reducing unnecessary switches to latent reasoning.
- We introduce **Evidence-Calibrated Soft Thinking**, which dynamically selects multimodal evidence based on the question and evolving reasoning state to calibrate probability-weighted token feedback through an uncertainty-aware gate, helping delay premature discrete commitments while limiting reasoning drift.
- Extensive experiments on multimodal long-context question-answering benchmarks demonstrate that Confident Thinking consistently improves answer accuracy and achieves superior overall performance over existing reasoning strategies.

2 Related Work

Multimodal Long-Context RAG. Multimodal long-context RAG supports question answering over textual and visual information distributed across pages and documents. ColPali (Faysse et al., 2025) enables visual page retrieval, while VisRAG2.0 (Sun et al., 2025b) combines explicit evidence collection with multi-image reasoning to mitigate visual hallucinations. M3DocRAG (Cho et al., 2024) and VisDoM (Suri et al., 2025) further explore multimodal retrieval and fusion for multi-document question answering. DocAgent (Sun et al., 2025a) combines structured document navigation with agent-based answer verification, while LatentRAG (Zheng & Worring, 2026) performs reasoning and retrieval in continuous space to reduce autoregressive overhead. These advances improve information access, evidence utilization, and inference efficiency, yet relevant retrieved content alone does not ensure faithful generation. When the content is heterogeneous, redundant, or conflicting, explicit CoT may commit to unreliable intermediate predictions that propagate through subsequent steps and produce hallucinated answers. Confident Thinking addresses this generation-stage challenge through multi-layer entropy-guided switching and evidence-calibrated latent updates, aiming to mitigate error propagation and hallucinations.

Latent Reasoning. Traditional Chain-of-Thought (CoT) reasoning (Wei et al., 2022) improves complex problem solving by explicitly generating intermediate steps, but its token-by-token autoregressive process repeatedly collapses a predictive distribution into a single discrete choice, potentially causing premature commitment when multiple reasoning paths remain plausible. This limitation has motivated a growing body of work on continuous and latent reasoning. Coconut (Hao et al., 2025) directly feeds continuous hidden states back into the language model, allowing reasoning to proceed without decoding every intermediate state into natural-language tokens. SoftCoT (Xu et al., 2025)

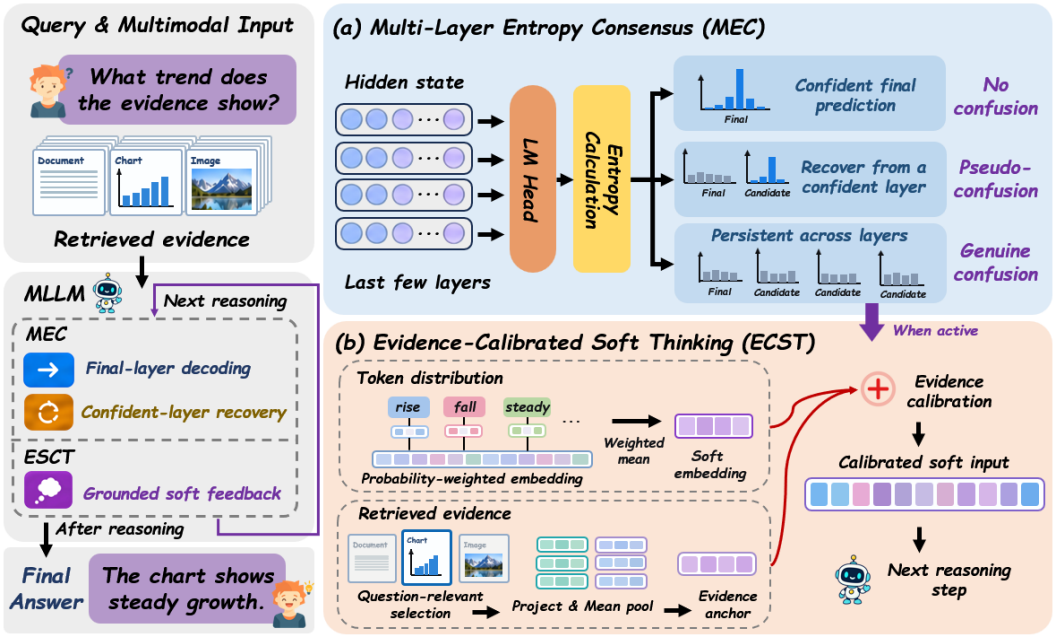


Figure 2: **Overview of Confident Thinking.** (a) MEC distinguishes genuine confusion from final-layer-specific pseudo-confusion. (b) ECST integrates retrieved evidence into soft embeddings to guide subsequent reasoning.

generates instance-specific soft thought tokens with an auxiliary model and maps them into the target LLM's representation space, while CODI (Shen et al., 2025) distills explicit CoT trajectories into continuous hidden representations through self-distillation. More closely related to our setting, Soft Thinking (Zhang et al., 2025) constructs continuous concept tokens as probability-weighted mixtures of vocabulary embeddings, preserving multiple candidate semantics rather than committing to a single token. DMLR (Liu et al., 2025) further extends latent reasoning to multimodal settings by refining latent think tokens through confidence-guided optimization and dynamically injecting relevant visual information. However, two key challenges remain when latent reasoning is applied to multimodal long-context RAG: how to selectively invoke latent reasoning only under genuine reasoning uncertainty, and how to keep latent reasoning grounded in relevant multimodal evidence as the reasoning process evolves. Therefore, we propose Confident Thinking, which combines Multi-Layer Entropy Consensus (MEC) for selective latent-reasoning activation with Evidence-Calibrated Soft Thinking (ECST) for dynamically calibrated soft token feedback.

3 Confident Thinking

To address the above challenges, we propose **Confident Thinking**, a training-free decoding framework for multimodal long-context retrieval-augmented generation. It consists of two complementary components: **Multi-Layer Entropy Consensus (MEC)** and **Evidence-Calibrated Soft Thinking (ECST)**. MEC distinguishes final-layer-specific pseudo-confusion from persistent cross-layer uncertainty to selectively activate latent reasoning, while ECST dynamically selects multimodal evidence according to the question and the current reasoning state to calibrate soft token feedback and mitigate reasoning drift. The overall framework is illustrated in Fig. 2.

3.1 Multi-Layer Entropy Consensus

Multi-Layer Entropy Consensus (MEC) aims to determine whether the uncertainty observed at the current reasoning step reflects final-layer-specific pseudo-confusion or persistent cross-layer uncertainty. A high-entropy final-layer prediction does not necessarily indicate genuine reasoning uncertainty, since a nearby layer may still produce a substantially more concentrated distribution. By jointly examining predictive entropy across multiple late Transformer layers, MEC filters such

final-layer-specific ambiguity and activates latent reasoning only when uncertainty persists across the candidate layers.

For a backbone with L Transformer layers, we consider the last w layers as the candidate set $\mathcal{C} = \{L - w + 1, \ldots, L\}$. Let $\mathbf{h}_t^{(\ell)}$ denote the hidden state at decoding step t from layer ℓ. For each candidate layer, we obtain the corresponding next-token predictive distribution $\mathbf{p}_t^{(\ell)}$ using the backbone's final normalization and language-model head. We quantify its uncertainty using Shannon entropy and identify the candidate layer with the most concentrated prediction as

$$\mathcal{H}_t^{(\ell)} = -\sum_{v \in \mathcal{V}} p_t^{(\ell)}(v) \log p_t^{(\ell)}(v), \qquad \ell_t^\star = \arg\min_{\ell \in \mathcal{C}} \mathcal{H}_t^{(\ell)}, \tag{1}$$

where $\mathcal{V}$ denotes the vocabulary and $\ell_t^\star$ is the candidate layer with the lowest predictive entropy. At each decoding step t, MEC evaluates whether the observed uncertainty is confined to the final layer or persists across the candidate layers. We identify the current prediction as *final-layer-specific pseudo-confusion* when

$$\mathcal{H}_t^{(L)} \geq \tau_H, \qquad \mathcal{H}_t^{(\ell_t^\star)} < \min\Big\{\tau_L,\, \mathcal{H}_t^{(L)} - \delta\Big\}, \tag{2}$$

where τ_H and τ_L denote the high- and low-entropy thresholds, respectively, and δ specifies the minimum entropy gap. Accordingly, MEC preserves discrete reasoning by adopting the prediction from $\ell_t^\star$. However, when high uncertainty is observed across all candidate layers, it can no longer be attributed to the final layer alone. To rule out ambiguity arising only at the current decoding step, MEC further examines whether such cross-layer uncertainty persists over K consecutive steps. We define

$$\mathcal{A}_t = \Big\{\ell \in \mathcal{C} \;\Big|\; \mathcal{H}_i^{(\ell)} \geq \tau_H,\ \forall\, i \in \{t - K + 1, \ldots, t\}\Big\}. \tag{3}$$

Here, $\mathcal{A}_t$ denotes the candidate layers whose predictive entropy remains above τ_H over the most recent K decoding steps. Accordingly, $\mathcal{A}_t = \mathcal{C}$ indicates that this high uncertainty is shared by all candidate layers, which we refer to as *persistent cross-layer uncertainty*. The subsequent reasoning is handled by ECST, as detailed in the following subsection.

3.2 Evidence-Calibrated Soft Thinking

Evidence-Calibrated Soft Thinking (ECST) is designed to address the persistent cross-layer uncertainty identified by MEC. Under such uncertainty, the predictive distribution still retains useful information about multiple possible continuations, which may be lost when reasoning is immediately reduced to a single discrete token. To preserve this information, ECST represents the predictive distribution as a probability-weighted token embedding and propagates it to the next reasoning step. Since this representation is derived from an uncertain prediction, it is further calibrated with retrieved multimodal evidence to guide subsequent reasoning and reduce reasoning drift.

Let $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ denote the token embedding matrix, where d is the embedding dimension and $\mathbf{E}(v) \in \mathbb{R}^d$ denotes the embedding of token v. We define the soft token $\mathbf{s}_t \in \mathbb{R}^d$ as

$$\mathbf{s}_t = \sum_{v \in \mathcal{V}} p_t^{(L)}(v) \mathbf{E}(v). \tag{4}$$

Each candidate token contributes to $\mathbf{s}_t$ according to its predictive probability, allowing information from multiple possible continuations to be retained during subsequent reasoning. We refer to this continuous token propagation as *soft thinking*.

However, soft thinking alone does not guarantee that subsequent reasoning remains grounded in the retrieved evidence, and the evidence most relevant to the current step may change as reasoning evolves. At each step, ECST therefore re-evaluates the relevance of the retrieved multimodal evidence to the input question and the current reasoning state, which captures the reasoning context accumulated up to step t. It then retains the k evidence items most relevant to the current state, where k denotes the number of selected items. These selected evidence items are used to calibrate the soft token.

Let $\mathbf{u}_{t,j} \in \mathbb{R}^d$ denote the representation of the j-th selected evidence item at step t, where $j \in \{1, \ldots, k\}$. We summarize the selected evidence into an evidence anchor $\mathbf{a}_t \in \mathbb{R}^d$, which is then

used to calibrate $\mathbf{s}_t$. The resulting calibrated soft token $\widetilde{\mathbf{s}}_t \in \mathbb{R}^d$ is given by

$$\mathbf{a}_t = \frac{1}{k}\sum_{j=1}^{k} \mathbf{u}_{t,j}, \qquad \widetilde{\mathbf{s}}_t = \mathbf{s}_t + \lambda\left(\mathbf{a}_t - \mathbf{s}_t\right), \tag{5}$$

where $\lambda \in [0, 1]$ controls the strength of evidence calibration. As the selected evidence changes with the current reasoning state, the evidence anchor is updated accordingly at each step. The calibrated soft token therefore combines the distributional information retained in $\mathbf{s}_t$ with evidence most relevant to the current reasoning step, providing a more grounded representation for subsequent reasoning.

The calibrated soft token $\widetilde{\mathbf{s}}_t$ is then propagated to the next reasoning step, where the model produces updated predictive distributions across the candidate layers. MEC subsequently re-evaluates these distributions to determine whether the cross-layer uncertainty has been resolved. Once a sufficiently confident candidate prediction re-emerges, with $\min_{\ell \in \mathcal{C}} \mathcal{H}_{t+1}^{(\ell)} < \tau_L$, soft thinking terminates and discrete token feedback resumes. Otherwise, soft thinking continues while the uncertainty persists, but for no more than M consecutive steps, where M denotes the maximum soft-thinking length, after which the final answer is generated using conventional final-layer discrete decoding.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate MEvi-Agents on two long-document multimodal benchmarks: MMLongBench-Doc (Ma et al., 2024b) and DocBench (Zou et al., 2025). MMLongBench-Doc contains 1,082 questions over 135 long PDFs from seven domains, covering multimodal, cross-page, and unanswerable reasoning. DocBench includes 1,102 questions over 229 real-world PDFs from five domains, spanning Text-only, Multimodal, Metadata, and Unanswerable settings.

Evaluation Metrics. For MMLongBench-Doc, we follow the official protocol and report generalized Accuracy (ACC) and F1, together with domain-wise results across Academic, Brochure, Financial, Guidebook, Industry, Report, and Tutorial documents. For DocBench, we use the official GPT-based binary evaluation and report Accuracy for Text-only, Multimodal, Metadata, and Unanswerable settings.

Baselines. We use DocAgent (Sun et al., 2025a) as the primary baseline and conduct controlled comparisons on three open-weight multimodal long-context backbones: Gemma4-31B-it, Qwen3.6-27B, and GLM-4.6V-Flash. For broader comparison, we also report published results from representative proprietary MLLMs, including Gemini 2.0 Flash, GPT-4o, and Claude 3.5 Sonnet on MMLongBench-Doc, and GPT-4o, GLM-4, Qwen-2.5, and ERNIE-3.5 on DocBench.

Implementation Details. We use the last four Transformer layers ($w = 4$) for uncertainty estimation, with high- and low-entropy thresholds $\tau_H = 1.0$ and $\tau_L = 0.2$. The uncertainty persistence threshold is set to $K = 2$, and each soft-thinking phase is limited to at most $M = 3$ steps. The same configuration is used across all backbones and datasets. Confident Thinking requires no additional training and is applied only at inference time. All experiments are conducted on two NVIDIA 80 GB A800 GPUs.

4.2 Main Results

We evaluate Confident Thinking on MMLongBench-Doc and DocBench using three different backbone models. Across both benchmarks, the proposed method consistently improves overall performance, demonstrating its effectiveness in reasoning over multimodal long-context retrieved evidence across different model backbones.

Performance on MMLongBench-Doc. As shown in Table 1, Confident Thinking improves the overall accuracy of Qwen3.6-27B, Gemma-4-31B-it, and GLM-4.6V-Flash from 50.01 to 54.77, 60.67 to 65.99, and 40.77 to 43.72, corresponding to gains of **+4.76**, **+5.32**, and **+2.95** points, respectively. The improvement is not concentrated in a small subset of document types: all seven document domains improve for all three backbones. For Qwen3.6-27B, the largest gains appear on Industry (+13.67), Brochure (+9.33), Tutorial (+6.62), and Financial (+6.11) documents. GLM-4.6V-

Table 1: **Performance comparison on MMLongBench-Doc.** We evaluate Confident Thinking with three multimodal long-context backbones across seven document categories. Best results within each backbone are highlighted in bold.

Method	Document Domain							ACC	F1
	Academic	Brochure	Financial	Guidebook	Industry	Report	Tutorial		
Proprietary MLLMs									
Gemini 2.0 Flash									
Gemini w/OCR	36.9	30.6	45.4	39.9	44.1	41.9	37.4	39.6	37.2
Gemini w/Page Images	27.9	31.5	31.4	43.3	44.6	44.3	43.2	38.4	36.1
Gemini w/CoA	28.0	28.6	44.2	34.6	43.2	45.2	33.2	37.2	31.9
Gemini w/RAG	19.3	24.4	7.7	22.9	40.8	24.9	18.8	22.1	11.1
GPT-4o									
GPT-4o w/OCR	–	–	–	–	–	–	–	30.1	30.5
GPT-4o w/Page Images	35.9	42.7	46.5	44.7	53.3	43.9	53.0	42.8	44.9
PDFTriage	40.6	32.8	45.9	40.0	42.8	41.7	31.2	39.6	35.6
Claude 3.5 Sonnet									
Claude 3.5 Sonnet w/OCR	46.4	36.6	62.8	43.3	46.4	43.4	37.3	44.9	42.4
Claude 3.5 Sonnet w/Page Images	26.5	30.9	30.3	38.2	57.0	44.4	49.4	39.0	36.7
Open-source Offline MLLMs									
Gemma-4-31B-it									
DocAgent	57.39	61.30	59.85	70.67	53.61	62.36	61.53	60.67	60.22
Confident Thinking	**61.89**	**65.31**	**63.18**	**74.58**	**58.71**	**69.07**	**67.69**	**65.99**	**61.15**
qwen3.6-27b									
DocAgent	51.69	53.42	43.60	68.90	41.30	49.45	45.23	50.01	49.42
Confident Thinking	**51.91**	**62.75**	**49.71**	**72.22**	**54.97**	**50.51**	**51.85**	**54.77**	**55.48**
GLM-4.6V-Flash									
DocAgent	40.12	45.03	32.47	50.30	45.63	36.98	39.68	40.77	37.53
Confident Thinking	**41.07**	**57.00**	**41.17**	**55.56**	**46.86**	**40.75**	**42.52**	**43.72**	**39.49**

Table 2: **Performance comparison on DocBench.** We report results across source categories and document types using three multimodal long-context backbones.

Model	Method	Source Category					Document Type				Overall
		Aca.	Fin.	Gov.	Laws	News	Meta.	MM	Text	Unans.	
Proprietary MLLMs											
GPT-4o	LLM-based	56.4	56.3	73.0	65.5	75.0	50.4	62.7	85.0	17.7	63.1
GLM-4	LLM-based	55.8	35.4	61.5	62.8	82.0	48.8	50.3	73.1	33.1	56.5
Qwen-2.5	LLM-based	42.9	29.9	51.4	55.5	69.2	36.0	31.8	61.7	58.1	46.9
ERNIE-3.5	LLM-based	56.4	37.5	54.7	58.1	58.1	36.8	47.7	63.6	54.0	51.8
Open-source Offline MLLMs											
Qwen3.6-27B	DocAgent	81.36	85.34	76.30	83.50	86.96	68.93	86.24	92.03	78.70	82.73
	Confident Thinking	**84.17**	**86.46**	**84.44**	**86.50**	**89.76**	**76.67**	**89.34**	**92.24**	**82.41**	**85.93**
Gemma-4-31B-it	DocAgent	**84.41**	85.03	78.89	82.15	86.22	73.75	**86.82**	90.70	78.70	83.56
	Confident Thinking	83.73	85.03	**79.26**	**83.38**	**87.09**	**74.21**	86.05	**91.50**	78.70	**83.77**
GLM-4.6V-Flash	DocAgent	64.24	61.92	**66.15**	70.50	**80.55**	52.83	66.47	75.95	76.85	67.36
	Confident Thinking	**68.57**	**66.12**	65.11	**76.15**	80.00	**56.71**	71.43	**78.48**	76.85	**70.56**

Flash shows particularly large improvements on Brochure (+11.97) and Financial (+8.70), whereas Gemma-4-31B-it obtains more balanced gains across domains, including Report (+6.71), Tutorial (+6.16), and Industry (+5.10). F1 scores also increase for all three backbones. These results suggest that Confident Thinking provides consistent benefits across diverse document structures rather than relying on a specific document category.

Performance on DocBench. Table 2 shows a similar overall trend on DocBench, although the gains are more backbone-dependent. Confident Thinking improves Qwen3.6-27B from 82.73 to 85.93 and GLM-4.6V-Flash from 67.36 to 70.56, yielding **+3.20** points for both models. For Qwen3.6-27B, the improvement is broad, covering every source category and all three document types, with particularly notable gains on Government (+8.14) and Metadata (+7.74). GLM-4.6V-Flash also benefits across most categories, including Laws (+5.65), multimodal documents (+4.96), Academic (+4.33), and Financial (+4.20), although small decreases are observed on Government and News. In contrast, the stronger Gemma-4-31B-it baseline improves only marginally from 83.56 to 83.77, with mixed changes across individual categories. Thus, while Confident Thinking remains beneficial in

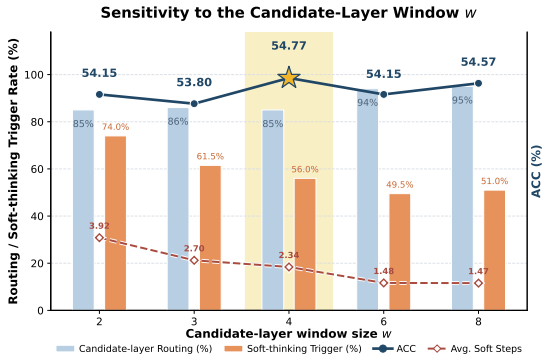


Figure 3: Sensitivity to the candidate-layer window w.

Table 3: **Ablation of MEC and ECST Components.** ECST is decomposed into soft thinking and evidence calibration (EC) for fine-grained analysis.

Variant	MEC	Soft	EC	Score
Standard CoT	–	–	–	50.01
FL Entropy + Soft	–	✓	–	48.49
MEC Only	✓	–	–	51.65
MEC + Soft	✓	✓	–	53.92
Ours	✓	✓	✓	**54.77**

FL: final-layer; Soft: soft thinking; EC: evidence calibration.

overall accuracy for all three backbones, its improvement on DocBench is substantially larger for Qwen3.6-27B and GLM-4.6V-Flash than for Gemma-4-31B-it.

4.3 Ablation Studies

We conduct ablation studies on **MMLongBench-Doc** with **Qwen3.6-27B** to evaluate the two core components of Confident Thinking: Multi-Layer Entropy Consensus (MEC) and Evidence-Calibrated Soft Thinking (ECST). For a finer-grained analysis, we further decompose ECST into its two constituents, *soft thinking* and *evidence calibration* (EC), and examine their individual and combined contributions.

Effectiveness of MEC and ECST Components. Table 3 evaluates the contribution of each component in Confident Thinking. Standard CoT achieves a score of 50.01, while directly triggering soft reasoning using final-layer entropy decreases performance to 48.49, indicating that soft reasoning alone is not sufficient and may be harmful when triggered by unreliable uncertainty signals. Introducing MEC improves the score to 51.65, showing the benefit of cross-layer uncertainty estimation even without soft reasoning. Combining MEC with soft thinking further improves performance to 53.92, and incorporating evidence calibration yields the best score of 54.77. Overall, these results show that MEC and ECST are complementary: MEC provides a more reliable basis for deciding when to intervene, while evidence-calibrated soft reasoning improves reasoning under genuine uncertainty.

Analysis of MEC Intervention Behavior. We further analyze the intervention behavior of different uncertainty-handling strategies in Appendix A.2 (Table 4). Using final-layer entropy alone triggers soft reasoning on 74.31% of samples with 6.47 soft steps per sample, yet performs worse than standard CoT (48.49 vs. 50.01), indicating substantial over-triggering. MEC alone improves the score to 51.65 through candidate-layer selection. When combined with soft thinking, the score further increases to 53.92, while the soft-triggered sample rate decreases to 63.30% and the average number of soft steps drops to 4.72. Confident Thinking achieves the best score of 54.77 with only 60.54% soft-triggered samples and 2.50 soft steps per sample. These results show that the gains come primarily from more selective uncertainty intervention rather than more frequent latent reasoning.

4.4 Hyperparameter Analysis

We analyze the sensitivity of Confident Thinking to three key hyperparameters: the candidate-layer window w, persistence threshold K, and soft-thinking budget M. Overall, the method exhibits stable performance across a reasonable range of settings, while moderate values provide the best balance between answer quality and intervention behavior.

Candidate-layer window w. Figure 3 shows the sensitivity to the candidate-layer window size w. Performance remains stable across different values, with $w = 4$ achieving the highest ACC of 54.77. As the window becomes larger, the routing rate generally increases, while the soft-trigger rate and average number of soft steps decrease. Further enlarging the window, however, does not yield additional performance gains. We therefore set $w = 4$ as the default.

Persistence threshold K. Increasing K consistently reduces both the soft-trigger rate and the average number of soft steps. Performance peaks at $K = 2$ with an ACC of 54.77; a smaller threshold leads to more frequent intervention, whereas larger values delay soft reasoning and reduce performance. We therefore set $K = 2$ by default. Detailed results are provided in Appendix A.3.

Soft-thinking budget M. Increasing M initially improves performance, with ACC reaching its maximum of 54.77 at $M = 3$, while a larger budget brings no further gains. This suggests that a short but sufficient soft-thinking phase is preferable to either premature termination or prolonged latent propagation. We therefore set $M = 3$ by default, with detailed results reported in Appendix A.3.

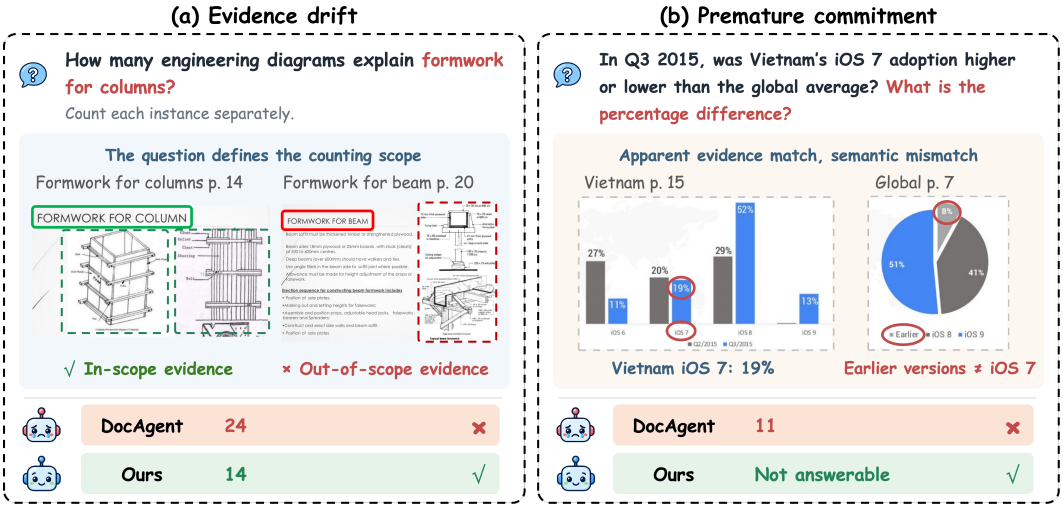


Figure 4: **Qualitative Results.** Confident Thinking mitigates evidence drift from out-of-scope content and avoids premature commitment under semantically mismatched evidence.

4.5 Qualitative Results

Figure 4 presents two representative cases illustrating reasoning errors during generation in multimodal long-context RAG. In the first case, the question restricts the evidence scope to column formwork diagrams, while a later section contains visually similar beam formwork diagrams. Baseline methods that include this section obtain an incorrect count of 24, reflecting a failure to preserve the question-specified scope when interpreting retrieved content. This allows irrelevant diagrams to influence the final count, whereas Confident Thinking follows the specified scope and returns 14. In the second case, the Vietnam chart reports an iOS 7 adoption rate of 19%, whereas the global chart only provides an aggregate value of 8% for earlier iOS versions. Baseline methods nevertheless proceed with the comparison by treating this aggregate as the global iOS 7 rate, introducing an unsupported intermediate assumption that propagates into the final answer. Confident Thinking recognizes the mismatch and determines that the requested comparison cannot be answered from the available evidence. This case illustrates how overinterpreting available evidence during generation can turn unsupported intermediate assumptions into hallucinated answers, motivating reasoning-time intervention when the model encounters uncertainty.

5 Conclusion

In this paper, we introduced Confident Thinking, a training-free framework for reliable reasoning in multimodal long-context RAG that selectively delays discrete commitments under genuine uncertainty. Confident Thinking combines Multi-Layer Entropy Consensus (MEC) for reliable uncertainty diagnosis with Evidence-Calibrated Soft Thinking (ECST) for grounding soft reasoning states in dynamically selected multimodal evidence. Extensive experiments on MMLongBench-Doc and DocBench demonstrate consistent improvements across multiple backbone models, while ablations show that better performance is achieved with fewer and more selective soft interventions. Overall, this work suggests that accurately identifying when to leave discrete reasoning is critical for reliable reasoning over heterogeneous long-context evidence.

AI USE STATEMENT

In line with the conference policy, we disclose that Large Language Models (LLMs) were used solely as writing aids. Their involvement was limited to improving grammar, refining sentence structure, and enhancing readability. All scientific contributions, including the development of ideas, methodology, experiments, and conclusions, were made exclusively by the authors, who take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our work. The source code for Confident Thinking, including the implementation of Multi-Layer Entropy Consensus (MEC) and Evidence-Calibrated Soft Thinking (ECST), will be made publicly available upon publication. Our framework is built upon publicly available models, and our experiments are conducted using three open-weight multimodal long-context backbones: Gemma4-31B-it, Qwen3.6-27B, and GLM-4.6V-Flash. All relevant citations for these models are provided in the main text. We evaluate our method on two standard and publicly available multimodal long-document question-answering benchmarks, MMLongBench-Doc and DocBench, following their official evaluation protocols. Critical hyperparameters and detailed experimental settings, including the candidate-layer window, entropy thresholds, uncertainty persistence threshold, and soft-thinking budget, are documented in Section 4.1. Confident Thinking requires no additional training and is applied entirely at inference time. Furthermore, the Appendix provides a detailed analysis of the intervention behavior of different uncertainty-handling strategies (Appendix A.2) and comprehensive sensitivity studies on the candidate-layer window w, uncertainty persistence threshold K, and soft-thinking budget M (Appendix A.3), further aiding the reproducibility of our results.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9233–9250, 2025.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. MM-DocRAG: Benchmarking retrieval-augmented multimodal generation for document question answering. *arXiv preprint arXiv:2505.16470*, 2025.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E. Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7036–7050, 2024.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. doi: 10.52202/068431-1613.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36407–36433. Curran Associates, Inc., 2023. doi: 10.52202/075280-1580.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space, 2025.

Fangrui Lv, Lei Wang, Ruixin Hong, Yong Du, Xiangyu Wu, Tingting Gao, Guorui Zhou, and Changshui Zhang. Beyond tokens: Dynamic latent reasoning via semantic residual refinement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32383–32391, 2026.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6492–6505, 2024a.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*, 2024b.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10862–10878, 2024.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377.

Dijia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57144–57163. PMLR, 2025.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12991–13013, 2024.

Li Sun, Liu He, Shuyue Jia, Yangfan He, and Chenyu You. Docagent: An agentic framework for multi-modal long-context document understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17712–17727, 2025a.

Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Sen Mei, Chi Chen, and Maosong Sun. VisRAG2.0: Mitigating visual hallucinations via evidence-guided multi-image reasoning in visual retrieval-augmented generation. *arXiv preprint arXiv:2510.09733*, 2025b. doi: 10.48550/arXiv.2510.09733. Revised version, July 2026.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6088–6109, 2025.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In *Advances in Neural Information Processing Systems*, volume 38, pp. 5220–5242, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley. CTRLS: Chain-of-thought reasoning via latent state transition. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2728–2736. PMLR, 02–05 May 2026.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. SoftCoT: Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23336–23351, Vienna, Austria, July 2025. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023. doi: 10.52202/075280-0517.

Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. Unified language representation for question answering over text, tables, and images. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4756–4765, 2023.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in LLMs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 333–356. Curran Associates, Inc., 2024. doi: 10.52202/079017-0011.

Xuanming Zhang, Sining Zhoubian, Yuxuan Chen, Tianyi Tang, An Yang, Sean Du, Chujie Zheng, Fei Huang, Dayiheng Liu, Gao Huang, and Jingren Zhou. Deeper is not always better: Mitigating the alignment tax via confident layer decoding. *arXiv preprint arXiv:2606.21906*, 2026.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5823–5840, Toronto, Canada, July 2023. Association for Computational Linguistics.

Yijia Zheng and Marcel Worring. Latentrag: Latent reasoning and retrieval for efficient agentic rag. *arXiv preprint arXiv:2605.06285*, 2026.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. Docbench: A benchmark for evaluating llm-based document reading systems. In *Proceedings of the 4th international workshop on knowledge-augmented methods for natural language processing*, pp. 359–373, 2025.

A Appendix

A.1 Additional Related Work

Chain-of-Thought (CoT) Reasoning. Chain-of-Thought (CoT) prompting improves complex reasoning by eliciting intermediate natural-language steps (Wei et al., 2022), while zero-shot CoT induces such reasoning without task-specific demonstrations (Kojima et al., 2022). Subsequent studies improve reasoning reliability through verification, reflection, and knowledge augmentation. Deductive verification (Ling et al., 2023) checks intermediate steps against their supporting premises, and Reflexion (Shinn et al., 2023) uses verbal feedback to refine behavior across successive attempts. Verify-and-Edit (Zhao et al., 2023) further improves factuality by post-editing generated reasoning chains using external knowledge. Beyond linear reasoning, Tree of Thoughts (ToT) (Yao et al., 2023) explores alternative reasoning paths through structured search, self-evaluation, and backtracking. Chain of Preference Optimization (CPO) (Zhang et al., 2024) fine-tunes models using stepwise preferences derived from tree search. CTRLS (Wu et al., 2026) guides stepwise generation through learned transitions between latent semantic states, supporting exploration of alternative reasoning trajectories. These methods improve how reasoning trajectories are constructed, evaluated, or optimized. Our work focuses on uncertainty during generation over retrieved multimodal content, where unreliable intermediate predictions may bias subsequent reasoning and produce hallucinated answers. Confident Thinking combines cross-layer entropy-guided switching with evidence-calibrated soft-token updates to defer premature discrete commitments and support evidence-grounded latent reasoning.

A.2 Analysis of MEC Intervention Behavior

Table 4 shows that the three MEC-based variants have nearly identical routing rates, indicating that the gains from soft thinking and evidence calibration do not come from more frequent candidate-layer switching.

The main difference lies in the soft-thinking phase. Adding evidence calibration reduces the average number of soft steps from 4.72 to 2.50 while further improving performance, suggesting that EC mainly helps the model recover from uncertainty more efficiently after activation.

Table 4: **Intervention Behavior Analysis.** Comparison of different uncertainty-handling strategies in terms of routing behavior, soft-thinking activation, and reasoning steps.

Policy	Score ↑	Routed Rate (%)	Soft-Trigger Rate (%)	Avg. Soft Steps
FL Entropy + Soft	48.49	–	74.31	6.47
MEC Only	51.65	85.86	–	–
MEC + Soft	53.92	85.64	63.30	4.72
Ours	**54.77**	85.40	60.54	2.50

FL denotes final-layer entropy; MEC denotes Multi-Layer Entropy Consensus. "–" indicates that the corresponding intervention is not applied.

A.3 Additional Hyperparameter Analysis Table

Candidate-layer window w. Table 5 shows that enlarging the candidate window mainly changes how uncertainty is resolved rather than substantially affecting accuracy. As w increases, candidate-layer routing becomes more frequent, while soft interventions generally decrease. This suggests that a broader candidate set allows more uncertain states to be handled through discrete routing instead of entering soft reasoning.

Persistence threshold K. Table 6 illustrates the role of K in controlling intervention timing. Increasing K consistently reduces both the soft-trigger rate and the average number of soft steps, indicating increasingly conservative activation. The non-monotonic accuracy trend suggests that temporal persistence should filter transient uncertainty without delaying intervention for genuinely difficult reasoning states.

Table 5: **Sensitivity to the candidate-layer window.**

w	ACC ↑	Route (%)	Soft Trigger (%)	Avg. Soft Steps
2	54.15	85.0	74.0	3.92
3	53.80	86.0	61.5	2.70
4	**54.77**	85.0	56.0	2.34
6	54.15	94.0	49.5	1.48
8	54.57	95.0	51.0	1.47

Table 6: **Sensitivity to the persistence threshold K.**

K	ACC ↑	Soft Trigger (%)	Avg. Soft Steps
1	51.24	68.30	4.80
2	**54.77**	60.54	2.50
3	52.27	57.55	2.33
4	50.95	51.10	1.80

Soft-thinking budget M. Table 7 further shows that the soft-thinking budget affects the structure of the intervention trajectory. Interestingly, smaller budgets result in more accumulated soft steps, which may indicate fragmented or repeated soft-thinking phases when a single intervention is insufficient to resolve uncertainty. Larger budgets allow each activated phase to proceed further, but excessive soft propagation does not necessarily improve the final prediction.

Overall, these results indicate that the three hyperparameters regulate different aspects of the reasoning process: w controls the opportunity for discrete recovery, K determines when soft reasoning is activated, and M controls the duration of each soft-thinking phase.

Table 7: **Sensitivity to the soft-thinking budget M.**

M	ACC ↑	Soft Trigger (%)	Avg. Soft Steps
1	48.30	68.15	7.50
2	52.63	63.26	5.62
3	**54.77**	60.54	2.50
4	53.85	60.03	2.30
5	52.37	58.65	1.97

A.4 Limitations

Confident Thinking is a white-box decoding framework that requires access to intermediate Transformer states and the language-model head, together with control over input embeddings. These requirements are essential for cross-layer uncertainty assessment and evidence-calibrated soft-token propagation, so the method cannot be directly applied to black-box API models. In addition, ECST operates on the context supplied by the upstream retriever, and its benefits may be limited when essential evidence is missing or misleading. Future work will focus on improving robustness to imperfect retrieval and strengthening the coordination between evidence acquisition and latent reasoning.