

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 THINKOMNI: LIFTING TEXTUAL REASONING TO OMNI-MODAL SCENARIOS VIA GUIDANCE DECODING

Anonymous authors

Paper under double-blind review

## ABSTRACT

Omni-modal reasoning is essential for intelligent systems to understand and draw inferences from diverse data sources. While existing Omni-modal Large Language Models (OLLM) excel at perceiving diverse modalities, they lack the complex reasoning abilities of recent Large Reasoning Models (LRM). However, enhancing the reasoning ability of OLLMs through additional training presents significant challenges, including the need for high-quality data, task-specific adaptation, and substantial computational costs. To address these limitations, we propose **THINKOMNI**, a training-free framework that lifts textual reasoning to omni-modal scenarios. THINKOMNI introduces two key components: 1) *LRM-as-a-Guide*, which leverages off-the-shelf LRMs to guide the OLLM decoding process; 2) *Stepwise Contrastive Scaling*, which adaptively balances perception and reasoning signals without manual hyperparameter tuning. Experiments on six multi-modality reasoning benchmarks demonstrate that THINKOMNI consistently delivers performance improvements, with main results achieving 70.2% on MathVista and 75.5% on MMAU. Overall, THINKOMNI offers a flexible and generalizable solution for omni-modal reasoning and provides new insights into the generalization and application of reasoning capabilities.

## 1 INTRODUCTION

The advent of Large Language Models (LLM) has marked a significant shift from traditional *fast thinking* paradigms, which rely on immediate and intuitive responses, to *slow thinking*, which emphasizes reflective and iterative reasoning, known as Large Reasoning Models (LRM). Recent LRMs, such as DeepSeek-R1 (Guo et al., 2025) and o1 (OpenAI, 2025), have demonstrated exceptional performance in specialized reasoning tasks like mathematical problem-solving and code generation. Nonetheless, their effectiveness remains predominantly constrained to textual inputs, thus limiting their applicability to more complex, omni-modal real-world scenarios involving text, audio, images, and videos (see Fig. 1(a)).

Omni-modal reasoning is essential for synthesizing diverse data sources and enabling sophisticated inference in context-rich tasks. Strong omni-modal reasoning capabilities have profound implications for practical applications such as advanced virtual assistants (Zhang et al., 2025) and embodied robots (Gan et al., 2020). Although recent advances in Omni-modal Large Language Models (OLLM) (Xu et al., 2025; Li et al., 2025b; Liu et al., 2025c; Fu et al., 2025; Luo et al., 2025) have shown promise in comprehending various input modalities, these models typically fall short when tasked with intricate reasoning across modalities, as illustrated in Fig. 1(b). Therefore, a fundamental research challenge is how to effectively extend and elevate the reasoning capabilities of models from primarily textual inputs to truly omni-modal scenarios.

Actually, this is not a trivial problem, and despite considerable efforts, existing approaches to omni-modal reasoning are still limited in several critical aspects. Specifically, 1) Insufficient modality diversity. Current studies largely focus on specific modalities (e.g., image (Liu et al., 2025b;a; Lin et al., 2025), audio (Li et al., 2025a), or video (Wang et al., 2025)), rather than generalizing across arbitrary combinations of modalities. 2) Task-specific enhancement. Enhancements proposed for existing OLLMs (Zhao et al., 2025; Zhong et al., 2025; Rouditchenko et al., 2025; Yang et al., 2025b) remain confined to particular downstream tasks, lacking broader generalizability. 3) Data scarcity and high training costs. Current methods predominantly rely on extensive supervised finetuning

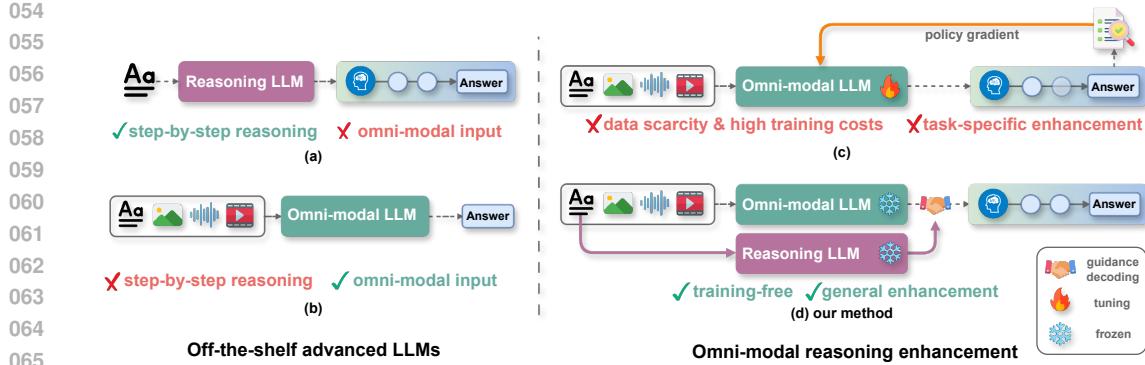


Figure 1: **Comparison between existing methods and THINKOMNI.** We integrate an OLLM with an LRM via guidance decoding, enabling advanced reasoning abilities with omni-modal input.

(SFT) (Xu et al., 2024; Yang et al., 2025c) (requiring tens of thousands of reasoning examples) or reinforcement finetuning (RFT) (Shao et al., 2024; Yu et al., 2025) approaches demanding training computational resources (e.g.,  $8 \times 40$ G VRAM for 7B model,  $16 \times 80$ G VRAM for 32B model). These challenges collectively motivate an important question: *Is it possible to overcome the constraints of data and training conditions to bring general reasoning abilities to omni-modal content?*

In this paper, we propose THINKOMNI, a novel training-free framework designed to lift textual reasoning to omni-modal scenarios (see Fig. 1(d)). Unlike existing approaches (see Fig. 1(c)) reliant on costly data annotation or additional model training, THINKOMNI directly leverages off-the-shelf LRM as decoding-time guides for OLLMs. Specifically, we first introduce the *LRM-as-a-Guide* strategy, enabling the integration of reasoning capabilities from LRMs into OLLMs. We further identify a potential issue: a fixed guidance weight is unsuitable for all the tasks, and manual, task-specific adjustment is impractical. To resolve this, we propose a *Stepwise Contrastive Scaling* module, adaptively balancing perceptual and reasoning signals based on real-time analysis of model predictions. This module adapts to various task types and facilitates coherent omni-modal reasoning.

Extensive experiments conducted on six challenging multi-modal reasoning benchmarks demonstrate the effectiveness of our method. Specifically, our method improves the state-of-the-art open source OLLM Qwen2.5-Omni (Xu et al., 2025) by substantial margins without additional training, as shown in Fig. 2, rivaling or surpassing models that undergo extensive RFT. Additionally, compared to other guidance decoding algorithms (Li et al., 2022; Liu et al., 2024), our method reduces the burden of multi-modal data input, thereby maintaining decoding efficiency.

THINKOMNI provides a flexible framework for lifting textual reasoning to a more diverse and enriched input space. By leveraging the strengths of OLLM and LRM, we explore the effective generalization of reasoning capabilities to omni-modal scenarios in a training-free manner. Besides, our method is not limited to current LRMs. As new LLM technologies emerge (often developing faster than multi-modal variants), our approach can be easily adapted to improve performance across multi-modal variants and other downstream domains.

## 2 PRELIMINARIES

### 2.1 NEXT TOKEN PREDICTION

Given an omni-modal input  $O$  (e.g., images, audios, videos) and a sequence of text tokens  $x_{<t} = (x_1, x_2, \dots, x_{t-1})$ , the OLLM  $M$  first computes the logits  $z_t$  for the next token  $x_t$ :

$$z_t = M(x_{<t}, O), \quad (1)$$

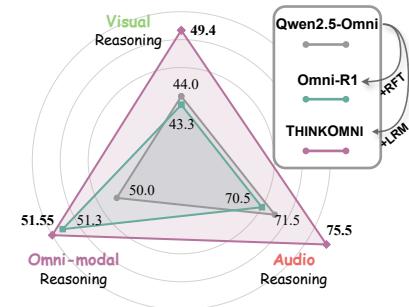


Figure 2: Performance comparison.

108 where  $z_t \in \mathbb{R}^V$  and  $V$  is the vocabulary size. The probability distribution  $P$  for  $x_t$  is then given by  
 109

$$110 \quad P(x_t | x_{<t}, O) = \text{Softmax}(z_t). \quad (2)$$

111 Then  $x_{<t+1} = (x_1, x_2, \dots, x_t)$ . The model computes a distribution and decodes a token at each  
 112 step, resulting in an auto-regressive generation process.  
 113

## 114 2.2 INFERENCE-TIME GUIDANCE DECODING

116 Finetuning large language models is time-consuming and costly, highlighting the need for methods  
 117 to modify or control models’ behaviors without additional training. In this subsection, we introduce  
 118 the following works to understand our method better: Contrastive Decoding (Li et al., 2022), Visual  
 119 Contrastive Decoding (Leng et al., 2024), ProxyTuning (Liu et al., 2024), and ProxyThinker (Xiao  
 120 et al., 2025). For models within the same family (i.e., sharing the same token vocabulary), these  
 121 methods guide base model decoding by introducing a contrastive pair at the logits level:  
 122

$$123 \quad \hat{z} = z^{\text{base}} + \alpha \cdot \underbrace{(z^+ - z^-)}_{\text{contrastive pair}}, \quad (3)$$

125 where  $\alpha$  controls the influence of the guidance signal. Here,  $z^+$  and  $z^-$  represent the logits from the  
 126 positive and negative references, respectively. These encourage or discourage certain behaviors in  
 127 the model’s output. This mechanism is analogous to a differential amplifier circuit, which amplifies  
 128 the desired signals while suppressing noise. Consequently, the model can reduce hallucinations or  
 129 achieve preference alignment during inference without additional training.

130 In Contrastive Decoding (Fig. 3(a)), the contrastive pair is formed by comparing the responses to the same  
 131 prompt from the original guiding model and an additional amateur model, with  $z^+$  set to  $z^{\text{base}}$ . In Visual  
 132 Contrastive Decoding (Fig. 3(b)), the contrastive pair is created by applying different input conditions to the same  
 133 model. Specifically,  $z^-$  is obtained by adding Gaussian noise to the input image and then performing inference.  
 134 In contrast to these approaches, ProxyTuning and ProxyThinker (Fig. 3(c))  
 135 construct contrastive pairs across different models within the same family, aiming to transfer behaviors from more minor, guiding models to larger, amateur models.  
 136

137 Existing guidance decoding methods are limited to scenarios with **consistent input modalities** and  
 138 **available expert models**. There is often no suitable expert model in omni-modal settings or other  
 139 downstream tasks, making it technically challenging to construct effective guidance signals. More-  
 140 over, the heterogeneity of modalities complicates the alignment and integration of guidance during  
 141 inference. Our work addresses these challenges by designing a framework for cross-modal guidance  
 142 decoding, enabling preference alignment without requiring modality-specific expert models.  
 143

144

## 145 3 METHOD

146

147 This section outlines the implementation roadmap of THINKOMNI, starting with a straightforward  
 148 guidance decoding approach. We first introduce *LRM-as-a-Guide*, which separates the input modalities of the  
 149 Omni-modal Large Language Model (OLLM) and incorporates an off-the-shelf Large Reasoning Model (LRM) as a guiding component. While this approach is practical, coordinating  
 150 fixed guidance decoding hyperparameters remains challenging due to the varying demands for reasoning  
 151 signals across different tasks and scenarios. To address this shortcoming, we propose *Stepwise  
 152 Contrastive Scaling*, a module that dynamically adjusts parameters based on real-time analysis of  
 153 model predictions, thereby adapting automatically to each decoding scenario. An overview of our  
 154 framework is provided in Fig. 4.  
 155

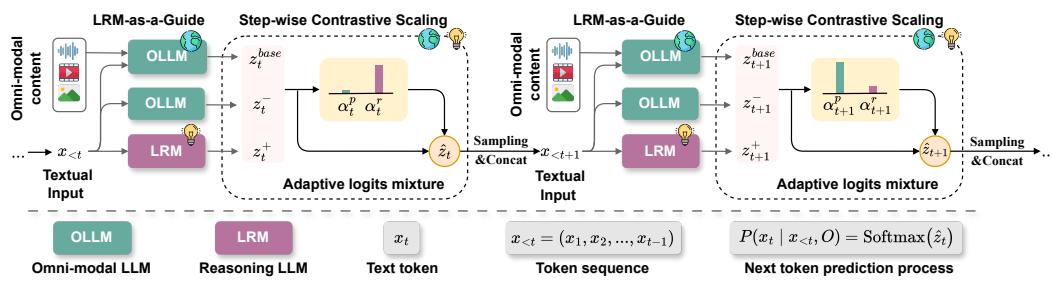


Figure 4: **Overview of THINKOMNI.** The framework begins by separating input modalities of the OLLM and introducing the LRM as a guiding model. Stepwise Contrastive Scaling dynamically adjusts guidance parameters based on real-time prediction analysis, enabling adaptive and effective decoding across diverse tasks.

### 3.1 LRM-AS-A-GUIDE

As discussed in Sec. 2.2, to address the gap where current guidance decoding approaches are limited to models with matched input modalities, we introduce the **LRM-as-a-Guide**, which lifts advanced textual reasoning into the omni-modal content through collaborative decoding with OLLM.

Let  $M_O$  denote the OLLM and  $M_R$  denote the LRM. As shown in Fig. 6(a), we compute the base logits with full omni-modal input,  $z^{\text{base}} = M_O(x_{<t}, O)$ . Then we discard the omni-modal content and feed  $M_O$  only the textual prefix  $x_{<t}$ . The results are treated as the negative logits  $z^- = M_O(x_{<t})$ . The positive logits are produced by the LRM on the same prefix  $z^+ = M_R(x_{<t})$ . As formulated in Eq. (3), the token probability distribution will then serve as

$$\hat{P} = \text{Softmax} \left[ M_O(x_{<t}, O) + \alpha \cdot (M_R(x_{<t}) - M_O(x_{<t})) \right], \quad (4)$$

where the scalar  $\alpha$  determines the extent to which the LRM influences the OLLM. After obtaining the mixed logits, we normalize them to probabilities and then sample the next token as usual.

Although the LRM cannot access omni-modal information, we mitigate this disadvantage and amplify the reasoning preference through the logits contrastive. During the generation process, the OLLM and LRM collaborate in a complementary manner. The OLLM, serving as the primary agent, extracts and integrates omni-modal clues, while the LRM provides deeper reasoning over the textual trace. As decoding progresses, the LRM can compensate for the lack of omni-modal information by leveraging the already decoded tokens, and the OLLM achieves logical reasoning through the reasoning preferences supplied by the LRM. Their strengths are seamlessly fused through logit mixing, resulting in a unified decoding framework that effectively integrates perception and reasoning.

### 3.2 STEPWISE CONTRASTIVE SCALING

While LRM-as-a-Guide effectively enables collaboration between the LRM and OLLM, there remains room for improvement regarding the choice of the fixed guidance weight  $\alpha$ . A fixed  $\alpha$  may not consistently achieve the optimal balance between perception and reasoning across different tasks.

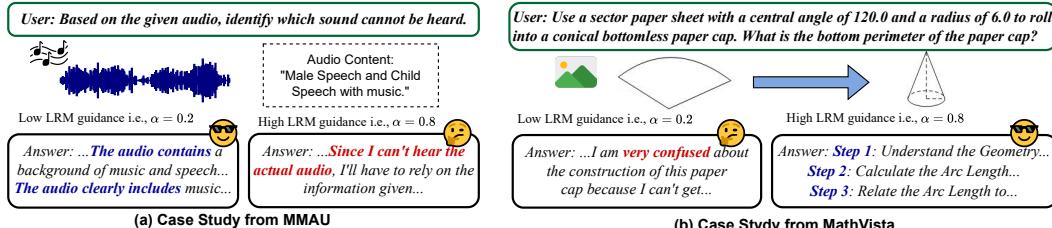


Figure 5: **Case studies from (a) MMAU and (b) MathVista.** (Sakshi et al., 2024; Lu et al., 2023) Tasks require different levels of LRM involvement. Using a fixed  $\alpha$  limits the ability of the model to optimally adapt to task-specific needs, highlighting the need for a more flexible approach.

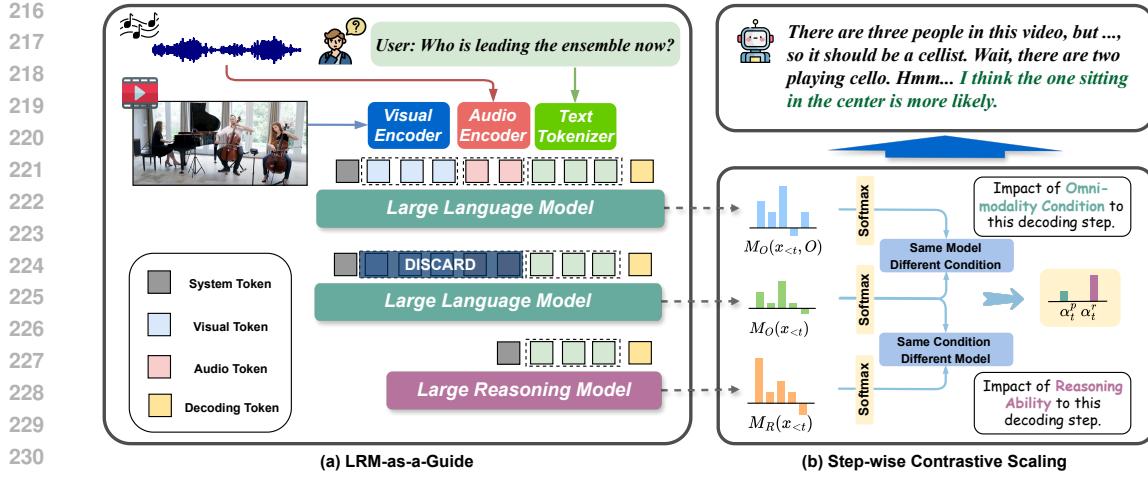


Figure 6: **Detailed process of THINKOMNI.** (a) The OLLM handles multi-modal inputs, while the LRM focuses on textual reasoning. By mixing their logits, the system effectively integrates perception and reasoning during token generation. (b) Each decoding step balances perception and reasoning dynamically by comparing logit distributions under different conditions and models.

The OLLM prefers a smaller  $\alpha$  to emphasize omni-modal cues, while the LRM benefits from a larger  $\alpha$  to strengthen its guidance. Additionally, since  $z^+$  and  $z^-$  do not possess comprehensive omni-modal content, excessive reliance on them (adopting a large  $\alpha$ ) can lead to recognition bias such as hallucination (Fig. 5(a)). Conversely, setting  $\alpha$  too low may diminish the effectiveness of guidance, thereby constraining the logical reasoning capabilities (see Fig. 5(b)). Motivated by this, we propose **Stepwise Contrastive Scaling**, which dynamically apportions a token’s prediction budget between perception and reasoning through online analysis of logits.

We introduce a stepwise influence metric to determine whether each decoding step is dominated by perception or reasoning. Specifically, all the generated logits are first transformed into probability distributions with a softmax function, let  $P_O$ ,  $P_R$ ,  $P$  denote the corresponding distributions for  $M_O(x_{<t}, O)$ ,  $M_R(x_{<t})$ , and  $M_O(x_{<t})$ , respectively. The pairwise distances between these distributions are then quantified by the Jensen–Shannon divergence, which is employed in DoLa (Chuang et al., 2023) to measure the similarity between two logits. This metric is symmetric, bounded, and numerically stable, making it well-suited for our purposes:

$$D_R = \text{JS}(P_R \parallel P), \quad D_P = \text{JS}(P_O \parallel P). \quad (5)$$

Intuitively,  $D_R$  reflects the unique influence of reasoning preference, whereas  $D_P$  captures the contribution from perceptual omni-modalities. A larger pairwise distance signifies that the corresponding factor (perception or reasoning) impacts the current decoding step more. Building on this metric, we proceed to reformulate Eq. (4) and introduce an additional contrastive logits term:

$$\hat{P} = \text{Softmax} \left[ M_O(x_{<t}, O) + \alpha_t^r \cdot (M_R(x_{<t}) - M_O(x_{<t})) + \alpha_t^p \cdot (M_O(x_{<t}, O) - M_O(x_{<t})) \right], \quad (6)$$

where  $\alpha_t^r$  acts as the original guidance weight, capturing enhanced reasoning capability, whereas the difference contributed by  $\alpha_t^p$  serves as an aggressive visual contrastive term (Leng et al., 2024) (i.e., by directly removing non-textual inputs rather than adding noise), reflecting augmented perceptual capability. To improve decoding stability, we normalize the coefficients as  $\alpha_t^r = D_R / (D_R + D_P)$  and  $\alpha_t^p = D_P / (D_R + D_P)$ , ensuring that the blended logits remain numerically stable. In practice, when both  $D_R$  and  $D_P$  are small, the adaptive mixing reduces to  $\hat{P} \approx P_O$ , recovering the original OLLM distribution. When both divergences are large, normalization induces a competitive dynamic, allocating a larger weight to the modality with greater divergence.

As shown in Fig. 4, the entire THINKOMNI procedure is training-free, requiring no additional finetuning or corpus statistics. Leveraging stepwise contrastive scaling, LRM-as-a-Guide can autonomously evaluate the relative contributions of perceptual and reasoning signals at each generation step, seamlessly balancing these complementary abilities without manual hyperparameter tuning.

270  
 271 **Table 1: Model performance on several omni-modal reasoning benchmarks.** Here, DeepSeek  
 272 refers to DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), and Qwen3 denotes Qwen3-8B (Yang  
 273 et al., 2025a). The numbers in parentheses indicate the performance changes compared to the base  
 274 OLLMs Qwen2.5-Omni-3B / 7B (Xu et al., 2025) and Omni-R1 (Zhong et al., 2025).

275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 Model	MathVista test-mini	MathVision test	MathVerse test-mini	MMAU(v05.15.25) test-mini	DailyOmni test	OmniBench test
<b>Close-Source Model</b>						
GPT-4o	63.8	30.4	50.8	62.5	56.5	-
Gemini-2.0-Flash	73.1	41.3	59.3	70.5	67.8	-
<b>Open-Source Omni Model</b>						
Baichuan-Omni-1.5	63.6	-	-	66.2	50.0	42.9
Ola	68.4	-	-	70.3	50.71	-
<b>Open-Source RFT Omni Model</b>						
Omni-R1	64.7	25.4	39.8	70.5	59.6	43.0
HumanOmniV2	68.8	25.4	37.3	75.3	58.5	41.9
<b>THINKOMNI-Qwen2.5-Omni-3B</b>						
Qwen2.5-Omni-3B	56.0	18.2	32.0	69.4	56.6	37.5
+ DeepSeek	56.1(+0.1)	20.2(+2.0)	33.5(+1.5)	70.1(+0.7)	57.1(+0.5)	39.9(+2.4)
+ Qwen3	58.1(+2.1)	25.3(+7.1)	38.8(+6.8)	70.6(+1.2)	57.3(+0.7)	39.5(+2.0)
<b>THINKOMNI-Qwen2.5-Omni-7B</b>						
Qwen2.5-Omni-7B	66.8	25.0	40.2	71.5	57.9	42.1
+ DeepSeek	68.8(+2.0)	28.2(+3.2)	42.0(+1.8)	73.8(+2.3)	59.8(+1.9)	43.2(+1.1)
+ Qwen3	70.2(+3.4)	32.9(+7.9)	45.1(+4.9)	75.5(+4.0)	59.5(+1.6)	43.6(+1.5)
<b>THINKOMNI-Omni-R1</b>						
Omni-R1	64.7	25.4	39.8	70.5	59.6	43.0
+ DeepSeek	66.1(+1.4)	27.0(+1.6)	43.1(+3.3)	73.1(+2.6)	60.3(+0.7)	43.5(+0.5)
+ Qwen3	71.3(+6.6)	31.5(+6.1)	45.2(+5.4)	75.4(+4.9)	59.8(+0.2)	43.4(+0.4)

## 4 EXPERIMENT

### 4.1 EXPERIMENT SETUP

**Models** To validate the effectiveness of THINKOMNI, we conduct experiments on three OLLMs: Qwen2.5-Omni-3B / 7B (Xu et al., 2025) and Omni-R1 (Zhong et al., 2025). We utilize the DeepSeek-R1-Distill series (Guo et al., 2025) and the Qwen3 series (Yang et al., 2025a), both in thinking mode, as our LRM to guide decoding.

**Benchmarks** To demonstrate the generalizability of THINKOMNI, we evaluate it on omni-modal scenarios using six benchmarks, comprising over 10,000 test samples in total: MathVista (test-mini) (Lu et al., 2023), MathVision (Wang et al., 2024), MathVerse (test-mini) (Zhang et al., 2024), MMAU-v05.15.25 (test-mini) (Sakshi et al., 2024), Daily-Omni (Zhou et al., 2025), OmniBench (Li et al., 2024). More details are provided in Appendix B.1.

**Evaluation** We first use template matching for multiple-choice questions to extract the option from the model’s output. If the answer cannot be extracted directly, we use GPT-4o to extract it and then compare the extracted answer to the gold answer. For free-form questions, we first use GPT-4o to extract the answer from the model’s output, then compare the extracted answer to the gold answer to determine if their meanings are consistent. This process is designed to account for various expressions in the answers.

### 4.2 MAIN RESULT

To evaluate the generality and scalability of THINKOMNI, we benchmark the improvements of different LRM guides on several OLLMs with varying capability levels. Our main results are presented in Tab. 1. The experiment result shows that THINKOMNI brings extensive improvements across all OLLMs, LRMs, and benchmarks. For example, with the Qwen3 guide, THINKOMNI brings remarkable improvement to Qwen2.5-Omni-7B on MathVision by 7.9%, achieving the final score of 32.9%. Since LRMs do not have access to omni-modal data contents, our results demonstrate that THINKOMNI indeed lifts the complex reasoning of LRMs to the omni-modal scenario.

324 Table 2: **Comparison with several training-free methods.** All are built upon Qwen2.5-Omni-7B.  
325

326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377	Method	326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377	326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377	MathVista test-mini	MMAU(v05.15.25) test-mini	OmniBench test
Base Model	66.8	71.5	42.1			
Average Logits Fusion	55.0( <b>-11.8</b> )	55.7( <b>-15.8</b> )	36.1( <b>-6.0</b> )			
Caption-then-Answer	61.0( <b>-5.8</b> )	59.7( <b>-11.8</b> )	32.3( <b>-9.8</b> )			
VCD	63.5( <b>-3.3</b> )	72.1( <b>+0.6</b> )	42.8( <b>+0.7</b> )			
ProxyTuning	67.0( <b>+0.2</b> )	72.3( <b>+0.8</b> )	43.0( <b>+0.9</b> )			
THINKOMNI (Ours)	68.8( <b>+2.0</b> )	73.8( <b>+2.3</b> )	43.2( <b>+1.1</b> )			

We compare our approach with methods trained using reinforcement learning finetuning (RFT) (i.e., Omni-R1 (Zhong et al., 2025) and HumanOmniV2 (Yang et al., 2025b)). Based on the same foundation model, Qwen2.5-Omni-7B, our DeepSeek-guided model achieves comparable performance, while our Qwen3-guided model consistently outperforms all the RFT-based methods. Moreover, our approach can be applied to models already undergoing RFT, further demonstrating broad performance improvements.

In addition, we observe differences in performance gains, which the following factors can explain: 1) the capabilities of LRM models, newer models like Qwen3, with stronger logical understanding and reasoning abilities, achieve greater improvements compared to DeepSeek under identical settings; 2) the training data of LRMs is biased towards scientific and mathematical content, leading to more pronounced gains on these tasks; 3) the tested tasks themselves differ in their demands for reasoning ability, with scientific and mathematical tasks typically requiring more reasoning than audio or general omni-modal tasks.

#### 349 4.3 COMPARE WITH TRAINING-FREE METHODS

We use the original evaluation results of the OLLMs as a baseline. In addition, we compare our method with several other training-free methods: 1) *Average Logits Fusion*, which directly averages the output logits of the OLLM and LRM during inference. 2) *Caption-then-Answer*, where the OLLM generates a detailed caption for the omni-modal input, and the LRM answers the question based on this caption. 3) *Visual Contrastive Decoding* (VCD) (Leng et al., 2024), which enhances perception by introducing noise to input content to generate negative logits. 4) *ProxyTuning* (Liu et al., 2024; Xiao et al., 2025), which enhances model performance by leveraging the logits from both the reinforcement finetuned small model and the original model, both of which are of the same type (LLM in ProxyTuning, VLM in ProxyThinker, and OLLM in our implementation).

As shown in Tab. 2, compared to the base OLLM, only *ProxyTuning* and THINKOMNI achieve performance improvements, with our method delivering a more substantial gain. For *Average Logits Fusion*, although simple mixing of logits allows the model to generate outputs, it negatively impacts answer accuracy due to improper integration. The *Caption-then-Answer* experiment demonstrates that when the LRM alone is responsible for answering, even with multi-modal information provided by the OLLM, performance drops significantly because information transmission is one-way. The OLLM cannot respond to the LRM’s specific needs. *VCD* is designed to enhance attention to multi-modal information rather than reasoning ability, so its performance declines on MathVista, which requires stronger reasoning skills.

#### 369 4.4 ABLATION STUDY

**Ablation study on fixed  $\alpha$  and adaptive  $\alpha^r$**  OLLM has limited capability in complex reasoning, while LRM cannot access multi-modal content. Over-reliance on either component leads to sub-optimal performance. As shown in Fig. 7(a), adjusting the fixed guiding weight  $\alpha$  markedly impacts results: when  $\alpha = 0$ , performance matches the original OLLM, and extreme  $\alpha$  values reduce scores on both benchmarks. In contrast, our Stepwise Contrastive Scaling (Full THINKOMNI) consistently achieves superior results across both benchmarks. Furthermore, Fig. 7(b) visualizes the distribution of the dynamic  $\alpha^r$ , revealing distinct shifts across different tasks and underscoring the adaptive nature of our method in autonomously tuning parameters to meet specific task requirements.

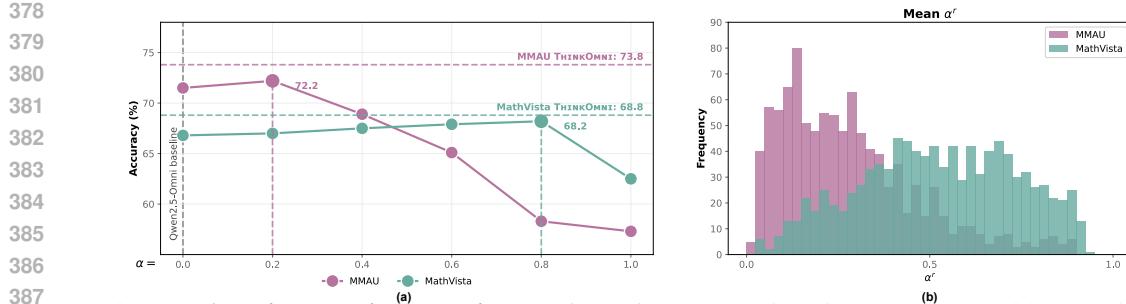


Figure 7: **Ablation on guidance weight.** Left: Performance varies with the constant guidance weight  $\alpha$ . Each task’s optimal  $\alpha$  range differs, with  $\alpha = 0$  as the OLLM baseline. Right: THINKOMNI uses adaptive dynamic weights, and the dynamic  $\alpha^r$  shows a similar distribution shift, indicating that stepwise contrastive scaling can flexibly adapt to different task requirements.

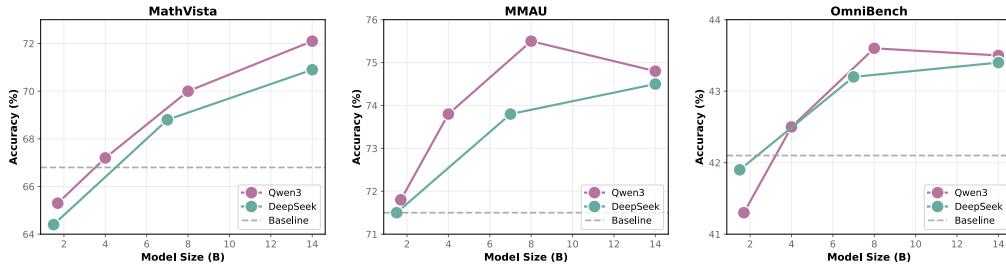


Figure 8: **LRM-as-a-Guide performance scaling.** We replace the Guide LRM on several benchmarks to study the impact of different LRM sizes and different LRM series on the performance of our method. The baseline refers to the performance of the Qwen2.5-Omni-7B

**Ablation study on different LRMs** Fig. 8 demonstrates the performance of Qwen2.5-Omni-7B under the guidance of LRM across different model sizes and series. When the LRM size is too small, performance on MathVista and OmniBench both degrades, as the language capabilities of LRM and the base OLLM are insufficiently matched. A sufficiently strong LRM positively impacts all benchmarks, with larger LRMs generally yielding better results than smaller ones. However, increasing the size to Qwen3-14B does not lead to further improvements on MMAU and OmniBench, suggesting that enhanced reasoning ability has a limited effect on these tasks, although the results still surpass the baseline.

## 4.5 ANALYSIS

**Qualitative analysis** Fig. 9 illustrates the generation process of THINKOMNI, where darker colors indicate greater LRM contributions to each token. These tokens, often logical connectives (e.g., “but”, “Therefore”) and key terms (e.g., “traditional”, “common”), are evenly distributed, showing that LRM consistently guides reasoning rather than merely supplementing OLLM. This highlights LRM’s role in analyzing multi-modal clues and driving logical inference. In contrast, lighter tokens are mainly function words and specific terms reflecting multi-modal content, suggesting that OLLM focuses on retrieving information and constructing fluent responses under LRM’s guidance. For more cases, see Appendix D.

**Failure case analysis** We identified several representative failure cases from the response. 1) THINKOMNI demonstrates correct multimodal perception, but conflicting information within the input leads to erroneous reasoning. As shown in Fig. 10(a), the model correctly recognizes that the highest visible marking on the beaker is 400ml. However, due to the beaker being labeled as 600ml, the model incorrectly infers that only a portion of the beaker is visible in the image, resulting in the wrong final answer. 2) Insufficient perceptual ability and limited sensitivity to subtle differences in the input lead to incorrect answers. As illustrated in Fig. 10(b), the model fails to accurately detect

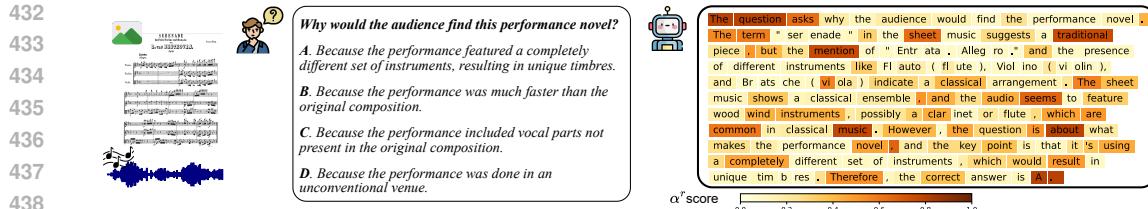


Figure 9: An OmniBench case study. This case study visualizes the reasoning process of THINKOMNI-Qwen2.5-Omni-7B, highlighting the stepwise contrastive scaling coefficient  $\alpha^r$ .

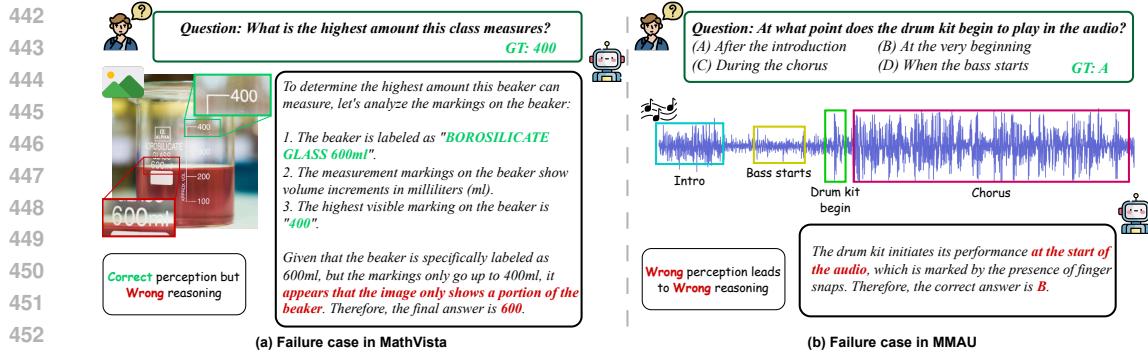


Figure 10: Failure case study. (a) THINKOMNI makes a reasoning error due to conflicting information between the visible beaker markings (400ml) and the label (600ml). (b) THINKOMNI fails because it cannot accurately detect the drum kit’s true start time in the audio.<sup>1</sup>

the actual onset of the drum kit in the audio and mistakenly identifies the beginning of the audio as the start of the drum kit’s performance, ultimately producing the wrong answer.

**Efficiency analysis** Although our method introduces some additional computation, it remains efficient. We measured generation latency on an H800-80G GPU with KV cache in the generate stage, benchmarking VCD (Leng et al., 2024), ProxyTuning (Liu et al., 2024), and THINKOMNI using 100 random OmniBench (Li et al., 2024) samples. VCD performs two forward passes per step, while ProxyTuning and THINKOMNI require three. As shown in Tab. 3, our method (7B+7B setting) incurs 1.38 $\times$  in the prefill stage (first token generation) and 2.88 $\times$  in the generate stage (response generation utilizing KV cache). Importantly, during decoding, the guiding model in THINKOMNI processes only text, which helps to reduce latency.

Table 3: Generation latency comparison. Prefill Time: first token generation latency; Generate Time: response generation with KV cache. Results are averaged over 100 samples from OmniBench.

Guidence Decoding Method	Model Size	Prefill Time $\downarrow$	Generate Time $\downarrow$
None (Baseline)	7B	0.138s	0.025s
VCD (Leng et al., 2024)	7B	0.262s (1.89 $\times$ )	0.050s (2.00 $\times$ )
ProxyTuning (Liu et al., 2024)	7B + 3B + 3B	0.406s (2.94 $\times$ )	0.086s (3.44 $\times$ )
THINKOMNI (Ours)	7B + 7B	0.191s (1.38 $\times$ )	0.072s (2.88 $\times$ )
THINKOMNI (Ours)	7B + 1.5B	0.191s (1.38 $\times$ )	0.069s (2.76 $\times$ )

## 5 RELATED WORK

**Omni-modal Large Language Models** With the rapid advancement of large language models, expanding their capabilities to omni-modal domains has become a key focus. Omni-modal Large Language Models (OMLLM) align and process information from multiple modalities, capturing richer

<sup>1</sup>For the reader’s understanding, the colored markings in Fig. 10 are visual aids added afterward and were not provided to the model.

486 semantics and context than single-modal systems. Proprietary models (Hurst et al., 2024; Deepmind,  
 487 2025) demonstrate impressive real-time multi-modal interaction, while open-source efforts such as  
 488 Qwen2.5-Omni (Xu et al., 2025; Li et al., 2025b; Liu et al., 2025c; Yang et al., 2025b; Fu et al.,  
 489 2025; Xie & Wu, 2024; Chen et al., 2024) are quickly closing the gap in modality alignment and  
 490 deployment.

491  
 492 **Large Reasoning Model** Recent advances in large-scale reasoning models, such as o1 (OpenAI,  
 493 2025) and DeepSeek-R1 (Guo et al., 2025), highlight the challenge of robust general reasoning.  
 494 Early methods used supervised fine-tuning with chain-of-thought data (Xu et al., 2024; Deng et al.,  
 495 2024), while recent work leverages reinforcement learning (Shao et al., 2024; Xiaomi et al., 2025;  
 496 Team et al., 2025; Yu et al., 2025; Tan et al., 2025; Zhu et al., 2025) for autonomous reasoning. As  
 497 OLLMs tackle complex cross-modal reasoning, bridging perception and reasoning remains a core  
 498 challenge.

499  
 500 **Decoding-time Algorithm** Decoding-time algorithms refine language model outputs at inference  
 501 in a training-free manner. Contrastive Decoding (O’Brien & Lewis, 2023) improves long-form gen-  
 502 eration by avoiding degenerate outputs, and Visual Contrastive Decoding (Leng et al., 2024) reduces  
 503 hallucination via visual input perturbation. ProxyTuning (Liu et al., 2024) combines expert outputs  
 504 and injects knowledge from finetuned models. ProxyThinker (Xiao et al., 2025) extends ProxyTun-  
 505 ing to multi-modal reasoning tasks. Beyond visual modalities, THINKOMNI adaptively integrates  
 506 reasoning with omni-modal perception, achieving a flexible fusion of fast and slow thinking.

## 507 6 CONCLUSION

511 We present THINKOMNI, a training-free inference-time framework that achieves robust and general-  
 512 izable reasoning enhancement by introducing an off-the-shelf LRM guide decoding with a stepwise  
 513 adaptive scaling mechanism. Across six challenging multi-modal benchmarks, it delivers consist-  
 514 ent gains, often matching or surpassing reinforcement-fine-tuned models, suggesting a general and  
 515 extensible paradigm for omni-modal reasoning.

516  
 517 **Limitation** THINKOMNI requires shared vocabularies between the OLLM and LRM for logit fusion  
 518 and introduces extra inference overhead due to additional forward passes. Nevertheless, we believe  
 519 our approach offers valuable insights for bridging the gap between multi-modal and textual LLMs  
 520 and provides a sustainable direction for future LLM improvements.

## 523 ETHICS STATEMENT

525 This work introduces a training-free, data-free inference framework using only publicly available  
 526 models (Qwen2.5-Omni (Xu et al., 2025), Omni-R1 (Zhong et al., 2025), DeepSeek-R1 (Guo et al.,  
 527 2025), Qwen3 (Yang et al., 2025a)) and benchmarks (see Appendix B.1), without collecting new hu-  
 528 man, biometric, or sensitive data. Risks stem from inherited biases, possible hallucinated cross-modal  
 529 attributions, and over-trust in generated reasoning chains; the method does not ensure factuality or  
 530 safety in high-stakes domains. Before deployment, we advise bias auditing, human oversight, and  
 531 external safety / factuality filters. Environmental impact is reduced relative to finetuning because no  
 532 additional training is performed.

## 534 535 REPRODUCIBILITY STATEMENT

536 Reproducibility is supported by: 1) the explicit inference formulation (Eq. 4 and Eq. 6; Algo-  
 537 rithm 1); 2) unified decoding hyperparameters (see Appendix B.1); 3) public benchmark splits and  
 538 the two-stage answer extraction pipeline follow VLMEvalKit (Duan et al., 2024); 4) hardware spec-  
 539 ification (H800-80G GPU, KV cache enabled). The code will be coming up soon upon acceptance.

540 **LLM USAGE**  
541

542 LLMs were used only for 1) minor code scaffolding/refactoring (boilerplate, log parsing), 2) linguistic  
 543 polishing after technical content was finalized, and 3) answer string normalization in the evaluation  
 544 pipeline (format extraction, not scoring). They were not used for ideas, model/method design,  
 545 experiments, analyses, claims, or interpretations. The researchers authored, validated, and cross-  
 546 checked all algorithms, equations, results, and conclusions. Every LLM-assisted output was manually  
 547 reviewed to prevent hallucination. Thus, LLM involvement provides no creative contribution and  
 548 does not affect the authenticity or reliability of the paper.

549 **REFERENCES**  
550

551 Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam,  
 552 Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, et al. Omnipxr: Evaluating omni-modality language  
 553 models on reasoning across modalities. *arXiv preprint arXiv:2410.12219*, 2024.

554 Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola:  
 555 Decoding by contrasting layers improves factuality in large language models. *arXiv preprint*  
 556 *arXiv:2309.03883*, 2023.

557 Google Deepmind. Gemini 2.5: Our most intelligent AI model —  
 558 blog.google. [https://blog.google/technology/google-deepmind/  
 559 gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking](https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking),  
 560 2025.

561 Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan  
 562 Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem  
 563 generation for geometric reasoning in large multimodal models. *arXiv e-prints*, pp. arXiv-2410,  
 564 2024.

565 Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong,  
 566 Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating  
 567 large multi-modality models. In *Proceedings of the 32nd ACM international conference on  
 568 multimedia*, pp. 11198–11201, 2024.

569 Chaoyou Fu, Haojin Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao,  
 570 Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech  
 571 interaction. *arXiv preprint arXiv:2501.01957*, 2025.

572 Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen,  
 573 and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on  
 574 Robotics and Automation*, pp. 9701–9707. IEEE, 2020.

575 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
 576 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
 577 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

581 Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong  
 582 Bing. Mitigating object hallucinations in large vision-language models through visual contrastive  
 583 decoding. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pp. 13872–  
 584 13882, 2024.

585 Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement  
 586 learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv  
 587 preprint arXiv:2503.11197*, 2025a.

588 Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke  
 589 Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization.  
 590 *arXiv preprint arXiv:2210.15097*, 2022.

594 Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Gu-  
 595 osheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*,  
 596 2025b.

597 Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng  
 598 Liu, Zekun Wang, Jian Yang, et al. Omnipbench: Towards the future of universal omni-language  
 599 models. *arXiv preprint arXiv:2409.15272*, 2024.

600 Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. Mind with eyes: from language  
 601 reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*, 2025.

602 Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A Smith. Tuning  
 603 language models by proxy. *arXiv preprint arXiv:2401.08565*, 2024.

604 Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-  
 605 zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint  
 606 arXiv:2503.06520*, 2025a.

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

609 Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola:  
 610 Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv  
 611 e-prints*, pp. arXiv-2502, 2025c.

612 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng,  
 613 Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning  
 614 of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

615 Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze  
 616 Chen, Jiaming Li, Lei Zhang, et al. Openomni: Advancing open-source omnimodal large lan-  
 617 guage models with progressive multimodal alignment and real-time self-aware emotional speech  
 618 synthesis. *arXiv preprint arXiv:2501.04561*, 2025.

619 Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models.  
 620 *arXiv preprint arXiv:2309.09117*, 2023.

621 OpenAI. OpenAI o1 System Card — openai.com. <https://cdn.openai.com/o1-system-card.pdf>, 2025.

622 Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio  
 623 Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? *arXiv  
 624 preprint arXiv:2505.09439*, 2025.

625 S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto,  
 626 Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio  
 627 understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

628 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,  
 629 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical  
 630 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

631 Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Ruihua Song. Think silently, think  
 632 fast: Dynamic latent compression of llm reasoning chains. *arXiv preprint arXiv:2505.16552*,  
 633 2025.

634 Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun  
 635 Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with  
 636 llms. *arXiv preprint arXiv:2501.12599*, 2025.

637 Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and  
 638 Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Proc.  
 639 of Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

648 Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju,  
 649 Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal  
 650 video grounding. *arXiv preprint arXiv:2503.13377*, 2025.

651

652 Zilin Xiao, Jaywon Koo, Siru Ouyang, Jefferson Hernandez, Yu Meng, and Vicente Ordonez. Prox-  
 653 ythinker: Test-time guidance through small visual reasoners. *arXiv preprint arXiv:2505.24872*,  
 654 2025.

655 LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang,  
 656 Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of  
 657 language model—from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025.

658

659 Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and  
 660 duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024.

661

662 Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language  
 663 models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

664

665 Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang  
 666 Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

667

668 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang  
 669 Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*,  
 670 2025a.

671

672 Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen  
 673 Yin, Xihan Wei, and Jingren Zhou. HumanomniV2: From understanding to omni-modal reasoning  
 674 with context. *arXiv preprint arXiv:2506.21277*, 2025b.

675

676 Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng  
 677 Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reason-  
 678 ing through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025c.

679

680 Qiyi Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong  
 681 Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at  
 682 scale. *arXiv preprint arXiv:2503.14476*, 2025.

683

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

684

685 Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. Stream-omni: Si-  
 686 multaneous multimodal interactions with large language-vision-speech model. *arXiv preprint*  
*arXiv:2506.13642*, 2025.

687

688 Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recog-  
 689 nition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

690

691 Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen,  
 692 and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system  
 693 collaboration. *arXiv preprint arXiv:2505.20256*, 2025.

694

695 Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal  
 alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

696

697 Linghao Zhu, Yiran Guan, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Bin Qin, Jian Luan, Yuliang  
 698 Liu, and Xiang Bai. Shuffle-r1: Efficient rl framework for multimodal large language models via  
 699 data-centric dynamic shuffle. *arXiv preprint arXiv:2508.05612*, 2025.

700

701

## 702 A A DETAILED EXPLANATION FOR THINKOMNI

704 At each decoding step  $t$ , the Omni-modal Large Language Model (OLLM) generates two token  
 705 probability distributions: conditioned on the full omni-modal input and the textual input only. In  
 706 contrast, the Large Reasoning Model (LRM) outputs a single distribution based solely on the textual  
 707 input.

708 We would like to decide how much of the next-token decision should be driven by perception ( $P_O^{(t)}$ )  
 709 and how much by abstract reasoning ( $P_R^{(t)}$ ). To gauge the *disagreement* between these signals, we  
 710 compute two Jensen–Shannon divergences  
 711

$$713 D_R^{(t)} = \text{JS}(P_R^{(t)} \parallel P^{(t)}), \quad D_P^{(t)} = \text{JS}(P_O^{(t)} \parallel P^{(t)}).$$

714  $D_R^{(t)}$  is large when the *reasoner* (LRM) wants a different token than the perceiver (OLLM text-only);  
 715  $D_P^{(t)}$  is large when masking the multi-modal content changes the OLLM’s belief is crucial, i.e. when  
 716 *perception*.  
 717

### 718 Why does this work?

- 720 1. If the current step mainly requires **perception** (e.g. reading a label in the image or detecting  
 721 a sound), masking the modalities hurts the OLLM, so  $D_P^{(t)}$  is large while  $D_R^{(t)}$  stays small.  
 722 Consequently  $\alpha_t^r \approx 0$  and the model trusts the *perception* logits.
- 724 2. If the step calls for **reasoning** (e.g. algebraic manipulation after the visual information is  
 725 extracted), the text-only LRM disagrees with  $P^{(t)}$ , so  $D_R^{(t)}$  dominates and  $\alpha_t^r \approx 1$ . The  
 726 generation is therefore guided by the stronger logical signal from the LRM.
- 727 3. For mixed cases,  $\alpha_t$  smoothly interpolates between the two extremes, allowing the decoder  
 728 to blend perception and reasoning in real time.

## 730 B PSEUDO CODE

732 To clearly illustrate the decoding procedure of THINKOMNI, we provide the overall pseudocode in  
 733 Algorithm 1. This algorithm describes how the Omni-modal Large Language Model (OLLM) and  
 734 the Large Reasoning Model (LRM) collaborate at inference time to achieve adaptive, training-free  
 735 omni-modal reasoning. At each decoding step, the framework dynamically fuses the perception and  
 736 reasoning signals via stepwise contrastive scaling, ensuring optimal balance between multi-modal  
 737 understanding and deep reasoning.

### 739 B.1 EXPERIMENT DETAILS

741 **Hyperparameter Settings** During our experiments, we used the following inference parameters  
 742 to ensure standard model outputs and to prevent performance degradation and endless rep-  
 743 etitions, following (Yang et al., 2025a): a temperature of 0.6, a `top_p` value of 0.95, a  
 744 `repetition_penalty` of 1.03, and a `max_new_tokens` of 4096. In addition, all LRMs  
 745 append the `<think>` tag to the end of the prompt during inference to ensure the reasoning state is  
 746 activated (Guo et al., 2025).

747 **Benchmarks Details** The datasets used in our experiments include MathVista (Lu et al., 2023),  
 748 MathVision (Wang et al., 2024), MathVerse (Zhang et al., 2024), MMAU (Sakshi et al., 2024),  
 749 Daily-Omni (Zhou et al., 2025) and OmniBench (Li et al., 2024).

- 751 • **MathVista** (Lu et al., 2023): We evaluate on the test-mini split of MathVista (1,000  
 752 samples), a unified benchmark for mathematical reasoning in visual contexts, which includes  
 753 three newly introduced datasets (IQTest, FunctionQA, and PaperQA), as well as 9 MathQA  
 754 and 19 VQA datasets from previous work.
- 755 • **MathVision** (Wang et al., 2024): We evaluate on the MathVision dataset (3,040 samples),  
 a curated collection of high-quality mathematical problems with visual contexts. Covering

---

**Algorithm 1:** THINKOMNI Decoding Framework

---

**Input:** OLLM  $M_O$ ; LRM  $M_R$ ; omni-modal input  $O$ ; initial tokens  $x_{<t}$   
**Output:** Generated sequence  $Y$

$Y \leftarrow x_{<t};$

**for**  $t = 1$  **to**  $T$  **do**

$z^{\text{base}} \leftarrow M_O(x_{<t}, O); \quad z^- \leftarrow M_O(x_{<t}); \quad z^+ \leftarrow M_R(x_{<t}); \quad // \text{ Compute logits}$

$P_O \leftarrow \text{Softmax}(z^{\text{base}}); \quad P_R \leftarrow \text{Softmax}(z^+); \quad P \leftarrow \text{Softmax}(z^-); \quad // \text{ Get distributions}$

$D_R \leftarrow \text{JS}(P_R \parallel P); \quad D_P \leftarrow \text{JS}(P_O \parallel P);$

$\alpha_t^r \leftarrow D_R / (D_R + D_P); \quad \alpha_t^p \leftarrow D_P / (D_R + D_P); \quad // \text{ Adaptive weights}$

$\hat{z}_t \leftarrow z^{\text{base}} + \alpha_t^r(z^+ - z^-) + \alpha_t^p(z^{\text{base}} - z^-);$

$\hat{P}_t \leftarrow \text{Softmax}(\hat{z}_t); \quad // \text{ Fuse and normalize}$

Sample  $x_t$  from  $\hat{P}_t$ ; Append  $x_t$  to  $Y$ ;  $x_{<t+1} \leftarrow (x_1, \dots, x_t);$

**if**  $x_t$  is end-of-sequence **then**

**break**

**end**

**end**

**return**  $Y$ ;

---

16 mathematical disciplines and five difficulty levels, MATHVision offers a comprehensive and diverse benchmark for assessing the mathematical reasoning abilities of LMMs.

- **MathVerse** (Zhang et al., 2024): We evaluate on the test-mini split of MathVerse (3,940 samples), a visual math benchmark with 2,612 multi-subject problems and diagrams. Each issue is annotated into six multi-modal versions, totaling 15K samples, to assess MLLMs' understanding of visual information in math reasoning.
- **MMAU** (Sakshi et al., 2024): We evaluate on the test-mini split of MMAU (1,000 samples), which consists of 10K audio clips with natural language questions and answers across speech, sounds, and music. Covering 12 retrieval and 15 reasoning types, MMAU challenges models with expert-level, domain-specific audio understanding and reasoning.
- **Daily-Omni** (Zhou et al., 2025): We evaluate on Daily-Omni (1,197 samples), which contains 684 videos from 11 YouTube categories and questions requiring integration of audio, visual, and textual information. The benchmark covers 30-second and 60-second videos to assess multi-modal reasoning abilities.
- **OmniBench** (Li et al., 2024): We evaluate on OmniBench (1,142 samples), which covers a wide range of reasoning and cognitive skills, from perception to complex reasoning. Tasks include object recognition, temporal and spatial reasoning, symbolic and quantitative processing, and various audio types, including speech, sound events, and music.

## C FUTURE WORK

THINKOMNI represents an attempt to introduce omni-modal capabilities based on textual reasoning abilities. We plan to explore additional modalities, such as 3D point clouds, protein structures, and reasoning applications in image/video generation scenarios. Moreover, we are also curious about "what truly works during the reasoning process", which is of great significance for understanding why reasoning abilities in the textual domain can generalize to a wider range of modalities.

810  
811  
812  
D MORE CASES  
813  
814  
815  
816  
817  
818

819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

**Please answer the following question based on the given image and audio:**  
*where this person in white robe might be.*

**Please choose only one answer from the following options:**

*A. woman in jail  
 B. woman in shelter  
 C. man in court  
 D. man in police station*



**Audio content:**  
 (Key to open the door + conversation + close the door)  
 I'll come back when the paperwork is complete.

**Answer:**

*The dim lighting, bare walls, and the presence of a simple bed suggest a confined space typical of a jail cell. The sound effects of metal clanging and a door slamming reinforce the idea of a prison environment. The person's attire is modest, which could indicate a prisoner.*

*While shelters can have similar settings, the specific sound effects and the presence of a bed rather than a cot or sleeping bag make it less likely. Shelters typically have more varied furnishings and may not have the same level of confinement.*

*The setting does not resemble a courtroom, which usually has a bench, desks, and more formal attire. The sounds heard do not match those of a court proceeding.*

*A police station might have similar elements, but the presence of sound effects and the specific description of the room as a cell point more towards a jail rather than a police station. Police stations often have more varied spaces and different types of furniture.*

*The answer is: A.*

Figure 11: Omni-modal Reasoning.

**Question: What is the most likely environment and time of day for these sounds?**

**Choices:**

*(A) Urban street at night  
 (B) Quiet home interior in the morning  
 (C) Busy office during lunch hour  
 (D) Forest trail at dusk*



**Audio content:**  
 The audio includes the sounds of cats and birds, and the ambient sound is relatively open.

**Answer:** The cat's meowing, combined with bird chirping, suggests a quiet setting where both animals are active, which is more common in a natural environment like a forest trail during dusk. The presence of bird chirping also indicates daylight, eliminating the possibility of it being at night or during a busy office hour. The quietness of the environment points towards a home interior, but the presence of birds suggests an outdoor setting, making the forest trail at dusk a less likely choice. Therefore, the most probable environment is a quiet home interior in the morning.

**The answer is: B.**

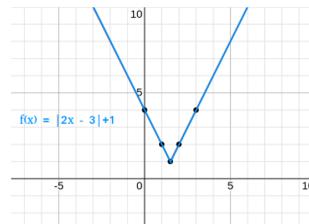
Figure 12: Audio Reasoning.

864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879

**Question:** The derivative of  $f(x)$  at  $x=2$  is \_\_\_\_ that at  $x=5$

**Choices:**

- (A) larger than
- (B) equal to
- (C) smaller than



**Answer:** To determine the relationship between the derivative of  $f(x)$  at  $x=2$  and  $x=5$ , we need to analyze the graph of the function  $f(x) = |2x - 3| + 1$ .

**1. Identify the critical point:** The function  $f(x) = |2x - 3| + 1$  has a V-shape, which means it has a vertex where the expression inside the absolute value changes sign. This occurs at  $2x - 3 = 0$ , which gives  $x = 3/2$ .

**2. Analyze the slope on either side of the vertex:**

- For  $x < 3/2$ , the expression inside the absolute value is negative, so  $f(x) = -(2x - 3) + 1 = -2x + 3 + 1 = -2x + 4$ . The slope here is  $-2$ .
- For  $x > 3/2$ , the expression inside the absolute value is positive, so  $f(x) = (2x - 3) + 1 = 2x - 3 + 1 = 2x - 2$ . The slope here is  $2$ .

**3. Evaluate the slopes at  $x=2$  and  $x=5$ :**

- At  $x=2$ , which is greater than  $3/2$ , the slope is  $2$ .
- At  $x=5$ , which is also greater than  $3/2$ , the slope is still  $2$ .

Since the slope (derivative) is the same for both  $x=2$  and  $x=5$ , the derivative of  $f(x)$  at  $x=2$  is equal to that at  $x=5$ .

Therefore, the correct answer is B.

905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

Figure 13: Visual Reasoning.