MetaphorStar: Image Metaphor Understanding and Reasoning with End-to-End Visual Reinforcement Learning

Anonymous CVPR submission

Paper ID 22012

Abstract

*Metaphorical comprehension in images remains a critical challenge for Nowadays AI systems. While Multimodal Large Language Models (MLLMs) excel at basic Visual Question Answering (VQA), they consistently struggle to grasp the nuanced cultural, emotional, and contextual implications embedded in visual content. This difficulty stems from the task's demand for sophisticated multi-hop reasoning, cultural context, and Theory of Mind (ToM) capabilities, which current models lack. To fill this gap, we propose **MetaphorStar**, the first end-to-end visual reinforcement learning (RL) framework for image implication tasks. Our framework includes three core components: the fine-grained dataset TFQ-Data, the visual RL method TFQ-GRPO, and the well-structured benchmark TFQ-Bench.*

Our fully open-source MetaphorStar family, trained using TFQ-GRPO on TFQ-Data, significantly improves performance by an average of 82.6% on the image implication benchmarks. Compared with 20+ mainstream MLLMs, MetaphorStar-32B achieves state-of-the-art (SOTA) results on True-False Question and Open-Style Question, and significantly outperforms top closed-source models GPT-4.1 and Claude-4.0-Sonnet on Multiple-Choice Question. Crucially, our experiments reveal that learning image implication tasks improves the general understanding ability, especially the complex visual reasoning ability. We further provide a systematic analysis of model parameter scaling, training data scaling, and the impact of different model architectures and training strategies, demonstrating the broad applicability of our method. We will open-source all MetaphorStar model weights, datasets, and method code.

1. Introduction

> We don't see things as they are, we see them as we are.
>
> — Anaïs Nin

This sentiment captures the core challenge of this paper: the profound gap between literal perception and conceptual understanding. The quote presents a dichotomy. "Seeing things as they are" is the realm of literal perception—the ability to identify objects and describe a scene, a task at which modern Multimodal Large Language Models (MLLMs) excel. "Seeing things as we are," however, is the realm of implication. It means interpreting that scene through the lens of human context, culture, and shared knowledge. This is the gap where MLLMs fail.

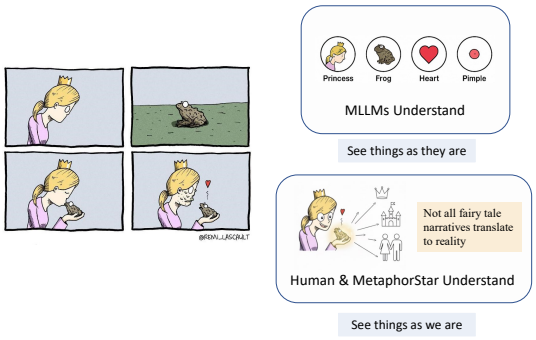

Figure 1. A picture is worth a thousand words: While MLLMs excel at literal object recognition (**"See things as they are"**), they often miss the deeper implication. Humans and our MetaphorStar model interpret the world **"See things as we are"**, grasping complex implications which behind the simple factual descriptions.

This gap is the essence of metaphorical comprehension, as visualized in Figure 1. Metaphors are not just abstract concepts found in literature, such as "time is money" or "life is a journey," but fundamental cognitive tools that allow us to conceptualize our surroundings [16]. In our daily lives, we are surrounded by *visual* metaphors: a political cartoon depicting a government as a "ship of state," an image of a person literally "at a crossroads," or a "wilted plant" on an office desk. An MLLM might see "a person" and "a split road" (seeing as it is), but it fails to infer the implication of a "life-changing decision" (seeing as we are). These images convey complex ideas by mapping one conceptual domain onto another. Just as humans use this abstract thinking to make sense of the world, we aim to enable AI to bridge this

CVPR #22012

gap and truly understand these implications.

In recent years, vision-language reasoning models such as o3 [33], Gemini-2.5-pro [10], and Grok-3-reasoning [45] have achieved outstanding performance. For example, Gemini-2.5-pro has reached a high score on math, code and vision-language reasoning benchmarks [22, 28, 40, 54]. However, these models still struggle with image metaphor questions [26, 56]. They tend to focus on the superficial elements of the image, neglecting the deeper connections and emotional expressions among them. It is important to note that these models excel at logical reasoning tasks, which are based on a different set of cognitive principles compared to image metaphor. Unlike VQA tasks that focus on concrete image comprehension, image metaphors require a stronger emphasis on abstract meaning and higher-order reasoning abilities. It is not a simple logical reasoning task and needs a different method to understand implications. It requires the model to grasp complex and abstract information, such as metaphors, symbols, and emotions in the image, rather than just concrete contents.

Understanding image implication is a more complex and challenging task than conventional VQA tasks. It requires advanced cognitive abilities such as multi-hop reasoning and a sophisticated theory of mind (ToM), which are inherent to human cognition [26, 56].

Existing methods for image metaphor understanding mainly fall into three categories. (1) Explicit mapping, represented by CLOT [60], creates links between metaphor ontologies and visual representations. It struggles with complex many-to-many mappings and dynamic cultural references. (2) Implicit reasoning, exemplified by C4MMD [49], uses training-free CoT reasoning. This passive approach often fails to handle the complex search space of abstract thought. (3) Contextual alignment [55] uses out-of-domain knowledge to align with cultural metaphors. This strategy introduces unpredictability from external knowledge quality and is computationally intensive.

To address these problems, we posit that a new approach is needed. Inspired by human cognitive models like the DIKW pyramid [4], we believe that the implicit reasoning capabilities within MLLMs should be sufficient, but they remain dormant, lacking a method to effectively activate this latent knowledge. Passive, training-free CoT prompting is often too weak to "find" the correct reasoning path. In contrast, Reinforcement Learning (RL) provides an active mechanism to explicitly reward and reinforce the model for exploring and strengthening these complex, non-literal reasoning pathways.

Therefore, we propose **MetaphorStar**, the first end-to-end visual RL framework for image implication. Our framework includes three core components: the fine-grained dataset TFQ-Data, the visual RL method TFQ-GRPO, and the well-structured benchmark TFQ-Bench. Our open-sourced MetaphorStar family, trained using this method, achieves state-of-the-art performance, and experiments consistently verify its superiority across TFQ, MCQ, and OSQ formats. *Our contributions are listed as follows:*

- We systematically analyze the image implication task and find that learning it helps improve general understanding ability, especially the complex visual reasoning ability, as demonstrated through sufficient experiments.
- To the best of our knowledge, we propose the first end-to-end RL framework for image implication tasks, including the fine-grained dataset TFQ-Data, the visual RL method, and the well-structured benchmark TFQ-Bench.
- Our fully open-scoured MetaphorStar family, trained using TFQ-GRPO on TFQ-Data, significantly improves performance by an average of 82.6% on the image implication benchmark. Compared with 20+ mainstream MLLMs, MetaphorStar-32B achieves SOTA on True-False Question and Open-Style Question, significantly outperforms the closed-source models GPT-4.1 and Claude-4.0-Sonnet on Multiple-Choice Question, and generalizes well on general VQA tasks.

2. Related Work

2.1. Image Implication

Image implication encompasses diverse cognitive phenomena, including humor, sarcasm, and broader metaphorical understanding. Early research in this domain often focused on specific aspects, such as humor recognition [12, 13] and sarcasm detection [7]. The rapid development of Large Language Models (LLMs) presents new opportunities for analyzing these implications, necessitating more comprehensive evaluation frameworks. To this end, DeepEval [50] provided a systematic taxonomy of image implications. Subsequently, II-Bench [26] emerged as the first English image implication benchmark, followed by CII-Bench [56], which extended this framework to Chinese images.

Image implication understanding requires sophisticated multi-hop reasoning and theory of mind (ToM) capabilities [26, 56]. Current methods generally fall into three categories. First, explicit metaphor mapping (e.g., CLOT [60]) links visual features to metaphor ontologies. This approach is limited by the complexity of many-to-many metaphorical relationships and the static nature of ontologies, which fail to capture dynamic cultural references. Second, model implicit reasoning (e.g., C4MMD [49]) utilizes techniques like Chain-of-Thought (CoT) prompting. However, it struggles with the non-logical nature of metaphor and the vast search space required for out-of-domain reasoning. Third, contextual alignment (e.g., LAD [55]) iteratively enriches image captions with knowledge from external sources. This strategy is computationally intensive and hindered by the unreliable quality of retrieved external information.

2.2. Vision-language Reasoning

The rapid advancement of LLMs has demonstrated remarkable text reasoning capabilities, as evidenced by models such as o1 [31] and DeepSeek-R1 [6]. However, real-world knowledge often transcends textual representation, with visual information encapsulating world knowledge that pure language models cannot access. For example, images inherently contain rich, multi-layered information that often resists straightforward textual description, including spatial relationships, contextual nuances, and implicit knowledge that humans process intuitively. This limitation has driven research toward integrating visual information into reasoning frameworks. Current research has developed three primary approaches to incorporate visual information into model reasoning: 1) Comprehensive MLLM Description: This approach treats visual content as the text grounding problem, as demonstrated by LLAVA-COT [47] and Mulberry [51]. 2) Multi-turn MLLM Interaction: Models like VoCoT [21], V* [42], o3 [33], Gemini-2.5-pro [10], and DeepEyes [58] employ iterative question-answering to extract fine-grained visual information at various levels of detail. 3) Tool-augmented Reasoning: Frameworks such as Visual Sketchpad [14], Whiteboard-of-Thought [29], o3 [33], Gemini-2.5-pro [10], and DeepEyes [58] leverage tool-based approaches to modify images and augment reasoning with prior knowledge embedded in these tools.

3. Method

3.1. True-False Question (TFQ) For Image Implication Understanding

Previous benchmarks have advanced the evaluation of image implication understanding through diverse question formats. II-Bench [26] introduced the Multiple-Choice Question (MCQ), which offers a balanced assessment of a model's comprehension. Subsequently, CII-Bench [56] proposed the Open-Style Question (OSQ), which represents an upper bound on task difficulty due to its high degree of openness and the sophisticated reasoning it demands.

Our analysis of these formats reveals a clear spectrum of challenges. While MCQ provides a stable, medium-difficulty evaluation and OSQ tests the limits of generative reasoning, there is a need for a more foundational and comprehensive assessment tool. To fill this gap, we introduce the True-False Question (TFQ). The TFQ task is designed as a fine-grained complement to MCQ, establishing a lower bound on difficulty. Unlike formats that target a single inferential conclusion, TFQ probes understanding across multiple dimensions by presenting a series of statements about an image. These statements cover not only the central implication but also essential visual information, akin to basic VQA, thereby ensuring a more holistic evaluation of a model's capabilities from perception to cognition.

As summarized in Table 1, the three formats offer a complementary suite for evaluation. We analyze them across three key dimensions essential for Reinforcement Learning:

- **Knowledge Density:** The breadth of factual and inferential points evaluated per image. TFQ ranks highest as it forces the model to verify multiple distinct propositions per image.
- **Learnability:** The ease with which a model can learn from the signal. TFQ provides a clearer, less noisy gradient signal compared to the complex search space of OSQ.
- **Verifiability:** The objectivity of the ground truth. TFQ offers definitive binary answers, avoiding the subjective ambiguity of open-ended generation.

This makes TFQ the ideal substrate for our visual RL framework, providing a dense and verifiable reward signal.

Ability	TFQ	MCQ	OSQ
Knowledge Density	⋆⋆⋆	⋆⋆	⋆
Learnability	⋆⋆⋆	⋆⋆	⋆
Verifiability	⋆⋆⋆	⋆⋆	⋆

Table 1. Comparison of True-False Question, Multiple-Choice Question, and Open-Style Question across different dimensions. TFQ offers superior properties for training, including higher knowledge density, better learnability and higher verifiability. Relative ranking: ⋆⋆⋆ (Highest) $>$ ⋆⋆ (Medium) $>$ ⋆ (Lowest).

3.2. TFQ-Data & TFQ-Bench

3.2.1. Data Generation

To construct our dataset, we leveraged the 1,434 high-quality metaphorical images from the II-Bench [26]. We utilized the GPT-4.1 model to generate a comprehensive set of TFQs. For each image, the model was provided with its detailed textual description and the ground-truth implication, prompting it to generate an average of 5-10 QA pairs, each with a definitive True/False answer. This process yields a total collection of 14,099 questions.

The question design was guided by several principles to ensure comprehensiveness. First, each TFQ is a proposition that evaluates understanding of key image content related to the central metaphor. Second, the questions are not confined to the implication itself but also probe the model's grasp of primary visual information (akin to basic VQA). Third, the set of questions for each image includes hierarchical difficulty levels; false statements are crafted to be plausible distractors, while true statements are clearly grounded in the visual or contextual evidence.

3.2.2. Dataset and Benchmark Splits

We partition the total collection (1,434 images, 14,099 questions) into dedicated sets for training (TFQ-Data) and evaluation (TFQ-Bench), as summarized in Table 2. The detailed statistic is in Appendix A.

TFQ-Data. The training set is provided in two scales. TFQ-Data-Full is the large-scale training set, containing 1,384 images and 13,607 questions. From this set, we also curate TFQ-Data-Lite, a smaller (100 images, 984 questions) subset hand-picked for its high quality, diversity, and richness, making it ideal for rapid experimentation.

TFQ-Bench. The evaluation component also exists at two scales. TFQ-Bench-Full refers to the entire dataset (1,434 images, 14,099 questions). TFQ-Bench-Lite is the efficient test set, containing 50 representative images and 492 questions, used for concise and standardized evaluation. Crucially, this TFQ-Bench-Lite set is strictly disjoint from the TFQ-Data-Full training set, ensuring a fair and rigorous evaluation of model performance.

Type	Split	Purpose	Images	Questions
TFQ-Data	**Lite**	Efficient Fine-tuning	100	984
	Full	Large-scale Training	1,384	13,607
TFQ-Bench	**Lite**	Efficient Evaluate	50	492
	Full	Full Benchmark	1,434	14,099

Table 2. Statistics of the TFQ-Data and TFQ-Bench splits.

3.3. TFQ-GRPO

Effectively training models for open-style image implication reasoning presents a significant design challenge. Directly training on OSQ is difficult due to its chaotic, high-dimensional search space and sparse reward signals. While MCQ is more structured, it also suffers from lower knowledge density and sparse rewards, making learning inefficient. We posit that our TFQ format is an ideal training mechanism for this task. The TFQ offers high knowledge density, a graduated difficulty spectrum (from easy to hard), and easily verifiable answers, providing a dense and stable learning signal for reinforcement learning.

We therefore propose TFQ-GRPO, a framework that leverages the TFQ-Data to fine-tune the model's reasoning capabilities. For the optimization algorithm, we adopt Group Relative Policy Optimization (GRPO), which has proven effective for diverse tasks.

Reward Design. In multimodal environments, sparse and outcome-driven reward signals are crucial for guiding vision-language models toward effective reasoning and decision-making. Given the open-style thinking process of the image implication question, we adopt a reward formulation that evaluates the reasoning trajectory based on final outcome quality and thinking format. The total reward is composed of two parts: the accuracy reward R_{acc} and the formatting reward R_{format}. The accuracy reward assesses whether the final answer is correct, while the format reward penalizes poorly structured outputs. Formally, given a reasoning trajectory τ, the total reward is defined as:

$$R(\tau) = \alpha R_{acc}(\tau) + (1-\alpha) R_{format}(\tau) \tag{1}$$

where R_{acc} is a binary reward for the correct final answer, R_{format} is a penalty for outputs that do not adhere to the specified tag structure, and $\alpha \in [0,1]$ is a hyperparameter balancing their importance.

GRPO. GRPO is an on-policy reinforcement learning algorithm. For each input x, the old policy model $\pi_{\theta_{old}}$ from previous step generate a group of rollouts $\{o_i\}_{i=1}^{G}$. Then, our reward function is used to calculate rewards for each o_i, getting $\{r_i\}_{i=1}^{G}$. We design a unified reward mechanism and the relative advantage is calculated as:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{2}$$

GRPO maximizes the following objective to optimize the model π_θ:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\text{Train Batch},\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|x)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{\pi_\theta(o_i \mid x)}{\pi_{\theta_{\text{old}}}(o_i \mid x)} A_i,\ \text{clip}\left(\frac{\pi_\theta(o_i \mid x)}{\pi_{\theta_{\text{old}}}(o_i \mid x)}, 1-\varepsilon, 1+\varepsilon \right) A_i \right) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \right]. \tag{3}$$

The core component of TFQ-GRPO is the structured reasoning prompt that guides the model through the desired inferential logic: *Image Description* → *Implication Analysis* → *Final Answer*. We instruct the model to first describe the image, then analyze its implications, and finally reason to get the answer. Our training template is shown in Table 3.

SYSTEM: Please according to the image, and try to answer the following true-false questions with the option T (True) or F (False). First, describe the image, then analyze the image implication, and finally reason to get the answer. Output the thinking process in `<think></think>` and the final correct answer in `<answer></answer>` tags. The output format should be as follows: `<think>...</think>` `<answer>...</answer>`.
USER: True-false questions: {}

Table 3. Training Template of TFQ-GRPO.

4. MetaphorStar Family

We introduce the MetaphorStar family, which comprises three sizes: 3B, 7B, and 32B. We utilize the QwenVL-2.5 series as the base model. We provide a detailed analysis of these models in the following sections.

4.1. Training Setup

We train all MetaphorStar models using an end-to-end TFQ-GRPO. We initially investigate a conventional two-stage pipeline, which involves a Supervised Fine-Tuning (SFT) warmup stage before RL. However, we find this method suboptimal, as the SFT warmup tends to constrain the model's intrinsic reasoning capabilities. The details are

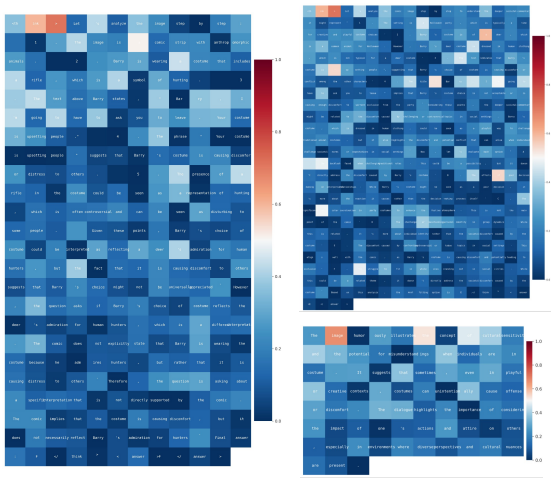

Figure 2. The visualization of token entropy for MetaphorStar-7B on TFQ, MCQ, and OSQ. High-entropy (red) indicates high uncertainty, while low-entropy (blue) indicates high confidence.

in Section 6 and Appendix C. In contrast, training directly with end-to-end TFQ-GRPO yields superior performance and better generalization. Therefore, we adopt the direct end-to-end RL for all experiments. The training process leverages the TFQ-Data-Lite. For the TFQ-GRPO algorithm, we set the group size for rollouts to $G = 5$. In our reward formulation, the hyperparameter α that balances the accuracy (R_{acc}) and format reward (R_{format}) is set to 0.5.

4.2. Analyzing Token Entropy in Reasoning

To gain insight into the internal reasoning mechanisms of our model, we analyze its token-level generation entropy. Figure 2 provides a visualization of this entropy as MetaphorStar-7B generates responses for the TFQ, MCQ, and OSQ tasks. Our analysis reveals that high-entropy tokens, representing points of highest uncertainty for the model, are not randomly distributed. This aligns with recent findings that "high-entropy minority" of tokens is critical for complex reasoning [41]. In the context of image implication, we observe that these spikes in uncertainty consistently occur at crucial semantic and logical junctions.

Specifically, the model exhibits high entropy when generating logical connectors (e.g., "therefore", "thus", "but") that pivot the argument or establish a causal link. We also note high entropy for key function words (e.g., "the", "is"), quantifiers, and pronouns, suggesting that the model's core cognitive effort is concentrated on making definitive logical leaps and structuring the relationship between concepts. Conversely, low-entropy (high-confidence) tokens are typically associated with reproducing factual details from the image or completing deterministic phrasal structures.

5. Experiment

5.1. Main Experiment

We carefully select a diverse range of MLLMs. Our evaluation utilizes the TFQ-Bench-Lite. For the comprehensive evaluation on image implication tasks, we also test on the high-level bench (EN) [55], featuring Multiple-Choice Question (MCQ) and Open-Style Question (OSQ). The details are in Appendix B.

5.1.1. True-False Question

Table 4 presents comprehensive results of TFQ across different MLLMs on the TFQ-Bench-Lite. Our MetaphorStar family achieves SOTA performance. MetaphorStar-32B (74%) and MetaphorStar-7B (70%) secure the first and second ranks, both outperforming the strongest closed-source model Gemini-2.5-pro (68%). Surprisingly, MetaphorStar-3B (62%) also surpasses powerful models Claude-4.0-Sonnet (52%) and GPT-4.1 (40%), indicating a severe deficiency in existing top-tier MLLMs for this task.

The effectiveness of our training is stark. MetaphorStar-7B (70%) shows a 150% relative improvement over its QwenVL-2.5-7B base (28%), and MetaphorStar-3B (62%) achieves a 210% relative gain over its QwenVL-2.5-3B base (20%). This demonstrates the potent efficacy of our TFQ-Data and the TFQ-GRPO method. We also observe that reasoning models generally perform better than general models on TFQ, which we attribute to the task's inclusion of basic VQA-style questions that probe primary visual information.

5.1.2. Multiple-Choice Question

Table 4 presents comprehensive results of MCQ across different MLLMs on the high-level bench (EN). Our models demonstrate strong generalization. MetaphorStar-32B is the top-performing open-source model, and MetaphorStar-7B (74%) is second, outperforming closed-source top models GPT-4.1 (74%). The generalization from our TFQ-centric training is evident. MetaphorStar-7B (74%) achieves a 60% relative improvement over its base model (46%), and MetaphorStar-3B (64%) achieves a 34% relative improvement over its base model (48%). Notably, on this task, the distinction between "reasoning" and "general" models is minimal. This suggests that the RL-based training in many existing reasoning models (often focused on math or code) has limited generalization to the abstract domain of image implication, which again highlights the unique effectiveness of our TFQ-GRPO.

5.1.3. Open-Style Question

Table 4 presents results of OSQ across different MLLMs on the high-level bench (EN). On the highly challenging OSQ task, MetaphorStar-32B (3.94) achieves the best score, significantly outperforming all other models, including Gemini-2.5-pro (3.38), Claude-4.0-Sonnet (3.46). This

Model	True-False Question	Multiple-Choice Question	Open-Style Question
General Models			
QwenVL-2.5-3B [3]	20%	48%	2.44
LLaVA-1.5-7B [23]	0%	16%	2.06
QwenVL-2.5-7B [3]	28%	46%	2.34
DeepSeek-VL2 [43]	20%	46%	2.82
GLM-4.1V-8B [59]	38%	60%	2.60
QwenVL-2.5-32B [3]	56%	62%	3.08
GPT-4o-mini [30]	36%	44%	2.98
Gemini-2.5-flash [10]	56%	76%	3.34
QwenVL-2.5-72B [3]	50%	72%	1.56
InternVL3-78B [61]	36%	70%	3.42
GLM-4V-plus [59]	42%	64%	3.01
Grok-3 [45]	36%	66%	3.24
Claude-3.5-Sonnet [1]	38%	68%	3.22
Claude-4.0-Sonnet [2]	52%	60%	<u>3.46</u>
GPT-4o [30]	50%	74%	2.94
GPT-4.1 [32]	40%	74%	3.30
Vision-language Reasoning Models			
Gemini-2.5-flash-thinking [11]	54%	78%	3.42
QVQ-72B [38]	28%	62%	3.10
o4-mini [33]	42%	58%	3.26
Doubao-1.5-thinking-vision-pro [36]	62%	66%	3.16
Grok-3-reasoning [45]	36%	74%	3.06
Gemini-2.5-pro [10]	68%	**82%**	3.38
Our MetaphorStar Family			
MetaphorStar-3B	62%	64%	3.06
MetaphorStar-7B	<u>70%</u>	74%	3.22
MetaphorStar-32B	**74%**	<u>78%</u>	**3.94**

Table 4. Overall results of different models on True-False Question, Multiple-Choice Question and Open-Style Question. The best-performing model in each category is **in-bold**, and the second best is <u>underlined</u>.

further proves the robust generalization of our method. MetaphorStar-7B shows a 38% relative gain over its base. Interestingly, unlike the MCQ results, we see significant performance disparities between reasoning and general models on OSQ. We also note that some models (e.g., QwenVL-2.5-72B) perform well on MCQ but poorly on OSQ. We attribute this to potential overfitting to multiple-choice formats and insufficient exposure to open-style generation. In addition, LLMs or even MLLMs may not genuinely understand the questions but rather predict options as answers, having evaluation bias and demonstrating sensitivity to option positioning, with similar findings in [55].

5.2. Generalization Experiment

5.2.1. Benchmarks and baselines

We evaluate generalization across two benchmark categories: (1) Reasoning, which is critical for complex decision-making, and (2) Understanding, which is crucial for real-world robustness. Appendix B lists the specific benchmarks. For a high-level overview, we report an average score (normalized 0–100, higher is better) across all benchmarks. We compare our MetaphorStar models against the QwenVL-2.5 series [3] baselines. To ensure a fair comparison, all evaluations employ VLMEvalKit [8].

5.2.2. Evaluation Results

Table 5 details the generalization performance of the MetaphorStar family against their respective base models. The results verify that our training on the image implication task provides a significant boost to visual reasoning, while simultaneously maintaining or even slightly improving performance on general visual understanding tasks, demonstrating robust and targeted generalization. Please see the detailed analysis in Appendix C.

Reasoning. The MetaphorStar family shows substantial and consistent reasoning improvements. On average, MetaphorStar-7B improves by 3.2 points and MetaphorStar-32B by 2.9 points over their baselines. The gains are most pronounced on challenging benchmarks: MetaphorStar-32B achieves a +16.2 point increase on MMMU (with 7B at +6.8 and 3B at +2.8). Strong gains also appear on MathVerse (+6.2 for 7B) and V* (+5.2 for 7B). This suggests our task's complex, multi-hop inference enhances underlying logical and visual reasoning faculties.

Understanding. In this domain, our specialized training does not harm, and often slightly improves general visual understanding. The MetaphorStar family maintains stable performance, with slight average improvements (e.g., +0.3 points for MetaphorStar-7B) across the 14 benchmarks. We

Benchmark	MetaphorStar Family			Base Model		
	MetaphorStar-32B	MetaphorStar-7B	MetaphorStar-3B	QwenVL-2.5-32B	QwenVL-2.5-7B	QwenVL-2.5-3B
Reasoning						
$MMMU_{test}$	**49.8**$_{\uparrow 16.2}$	48.8$_{\uparrow 6.8}$	45.1$_{\uparrow 2.8}$	33.6	42.0	42.3
VisualPuzzles	**39.7**$_{\uparrow 2.5}$	35.9$_{\uparrow 2.2}$	33.8$_{\uparrow 2.8}$	37.2	33.7	31.0
LogicVista	**56.6**$_{\uparrow 1.6}$	47.2$_{\uparrow 3.1}$	39.4	55.0	44.1	39.4
VisuLogic	25.5$_{\downarrow 0.8}$	**26.9**$_{\uparrow 2.2}$	18.8$_{\downarrow 0.3}$	26.3	24.7	19.1
V*	**81.2**$_{\uparrow 0.1}$	76.4$_{\uparrow 5.2}$	34.0	81.1	71.2	34.0
$ZeroBench_{main}$	**1.0**$_{\uparrow 1.0}$	**1.0**$_{\uparrow 1.0}$	0.0	0.0	0.0	0.0
$ZeroBench_{sub}$	**18.0**$_{\uparrow 2.4}$	15.3$_{\uparrow 1.2}$	6.6$_{\uparrow 1.2}$	15.6	14.1	5.4
MathVision	**38.1**$_{\uparrow 0.7}$	25.3$_{\uparrow 0.2}$	22.2$_{\uparrow 1.0}$	37.4	25.1	21.2
$MathVerse_{Vision}$	**50.8**$_{\uparrow 2.4}$	41.4$_{\uparrow 6.2}$	30.0$_{\uparrow 0.8}$	48.4	35.2	29.2
WeMath	**48.6**$_{\uparrow 2.5}$	36.7$_{\uparrow 2.4}$	21.7$_{\downarrow 1.2}$	46.1	34.3	22.9
Avg.	**41.0**$_{\uparrow 2.9}$	35.5$_{\uparrow 3.2}$	25.4$_{\uparrow 1.0}$	38.1	32.3	24.4
Understanding						
SEEDBench	**77.6**$_{\uparrow 0.2}$	77.1$_{\uparrow 0.1}$	74.0	77.4	77.0	74.0
SEEDBench2 Plus	**73.2**$_{\uparrow 0.8}$	70.8$_{\uparrow 0.1}$	63.6$_{\uparrow 0.3}$	72.4	70.7	63.3
$MMBench\text{-}V1.0\text{-}EN_{test}$	85.8$_{\downarrow 0.6}$	83.5	79.7$_{\uparrow 0.6}$	**86.4**	83.5	79.1
$MMBench\text{-}V1.1\text{-}EN_{test}$	**84.4**$_{\uparrow 0.4}$	82.5$_{\uparrow 0.3}$	77.6$_{\uparrow 0.8}$	84.0	82.2	76.8
MMStar	**68.5**$_{\uparrow 2.2}$	64.1$_{\uparrow 0.2}$	55.5$_{\downarrow 0.4}$	66.3	63.9	55.9
OCRBench	86.1$_{\uparrow 0.5}$	**88.6**$_{\uparrow 2.2}$	81.8$_{\uparrow 2.1}$	85.6	86.4	79.7
$AI2D_{test}$	83.3$_{\uparrow 1.0}$	**84.4**$_{\uparrow 0.5}$	81.2$_{\uparrow 0.4}$	82.3	83.9	80.8
$ScienceQA_{test}$	**91.3**$_{\uparrow 0.4}$	89.0	81.8$_{\uparrow 0.4}$	90.9	89.0	81.4
POPE	**86.3**$_{\uparrow 0.6}$	86.0$_{\uparrow 0.1}$	86.2$_{\uparrow 0.3}$	85.7	85.9	85.9
$MMT\text{-}Bench_{val}$	**65.7**$_{\uparrow 0.4}$	62.3$_{\uparrow 0.2}$	56.6$_{\uparrow 0.3}$	65.3	62.1	56.3
$RealworldQA_{avg}$	**71.1**$_{\uparrow 1.0}$	68.1$_{\downarrow 0.4}$	62.4$_{\downarrow 3.0}$	70.1	68.5	65.4
$BLINK_{val}$	62.9$_{\downarrow 0.8}$	56.6$_{\uparrow 1.3}$	44.5$_{\downarrow 0.2}$	**63.7**	55.3	44.7
$HallusionBench_{avg}$	55.3$_{\downarrow 1.4}$	49.9$_{\downarrow 1.8}$	46.5$_{\uparrow 0.2}$	**56.7**	51.7	46.3
MMVet Hard	59.6$_{\downarrow 4.4}$	54.6$_{\uparrow 1.7}$	50.2$_{\downarrow 0.5}$	**64.0**	52.9	50.7
Avg.	**75.1**$_{\uparrow 0.1}$	72.7$_{\uparrow 0.3}$	67.3$_{\uparrow 0.1}$	75.0	72.4	67.2
Overall Avg.	**60.9**$_{\uparrow 1.3}$	57.2$_{\uparrow 1.5}$	49.9$_{\uparrow 0.5}$	59.6	55.7	49.4

Table 5. Results on different visual question answering tasks. The best-performing model in each category is **in-bold**. Performance differences relative to base models are shown as colodarkred subscripts: ↑ for improvements, ↓ for declines.

note positive gains on challenging benchmarks like MMStar (+2.2 for 32B) and OCRBench (+2.2 for 7B). The overall performance confirms our method enhances reasoning without sacrificing foundational understanding.

6. Ablation Study

6.1. Model Parameter Scaling

We analyze the impact of model parameter scaling, with results in Figure 3 and Table 4. Our analysis reveals that the TFQ-GRPO training is crucial for unlocking the benefits of model scaling. Base models (w/o TFQ-GRPO) exhibit inconsistent or weak scaling; for instance, on OSQ, the 7B base model (2.34) underperforms the 3B base model (2.44). In sharp contrast, our trained MetaphorStar models demonstrate a clean and monotonic performance increase with scale (3.06 → 3.22 → 3.94). With our method enabling effective scaling, we observe that performance systematically improves with parameter count. This effect is most pronounced on OSQ, which shows accelerating marginal returns (3B→7B: +0.16 vs. 7B→32B: +0.72). This suggests that OSQ's open-ended reasoning disproportionately benefits from larger model capacity. Conversely, the closed-ended TFQ and MCQ tasks show more linear, though still consistent, performance gains as model size increases.

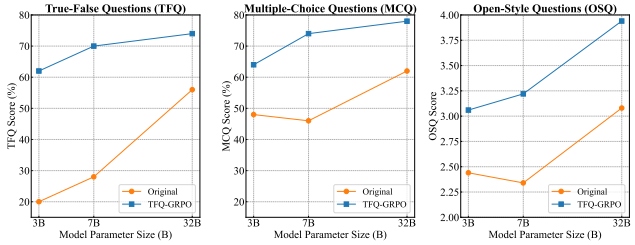

Figure 3. The model parameter scaling law.

6.2. Training Data Scaling

We investigate the impact of training data volume on model performance. We create three high-quality data subsets from TFQ-Data at different scales: Small (0.1k images), Lite (1k images), and Full (1.4k images). We train three distinct MetaphorStar-7B models on these datasets using identical TFQ-GRPO training parameters. As shown in Table 6, the results demonstrate two trends. First, performance scales positively and significantly with data quantity across all three tasks. Second, all three models substantially outperform the QwenVL-2.5-7B base model, confirming the powerful effect of our TFQ-Data. It is particularly noteworthy that even the MetaphorStar-7B-Small model, trained on only 0.1k images, improving 48% on TFQ and 64% on MCQ than base model. This highlights the high quality and data efficiency of our dataset. Furthermore, MetaphorStar-

Data	TFQ	MCQ	OSQ
Small (0.1k)	48%	64%	3.04
Lite (1k)	70%	**74%**	3.22
Full (1.4k)	**84%**	**74%**	**3.48**

Table 6. Results of scaling training data. The best-performing model in each category is **in-bold**.

Model	TFQ	MCQ	OSQ
LLaVA-1.5-7B			
w/o TFQ-GRPO	0%	16%	2.06
w/ TFQ-GRPO	**6%**	**34%**	**2.78**

Table 7. Results of different base models. The best-performing model in each category is **in-bold**.

Model	TFQ	MCQ	OSQ
QwenVL-2.5-7B	28%	46%	2.34
+ TFQ-SFT	42%	28%	3.34
+ TFQ-GRPO	**70%**	**74%**	3.22
+ TFQ-SFT & TFQ-GRPO	56%	28%	**3.66**

Table 8. Results of different training strategies. The best-performing model in each category is **in-bold**.

7B-Full, trained on the complete 1.4k dataset, achieves SOTA performance on the TFQ task at 84%. This result not only leads the 7B scale but also surpasses the 74% score of the MetaphorStar-32B model (trained on 1k images), underscoring that for this task, data scale can also be critical.

6.3. Different Model Architecture

To validate the generalizability of our TFQ-GRPO training framework, we test its effectiveness on a different model architecture. We select LLaVA-1.5-7B [23], which is based on the Vicuna (LLaMA-based), presenting a distinct architecture from the QwenVL series. We train this model using the same TFQ-GRPO training parameters and dataset TFQ-Data-Lite as our MetaphorStar-7B model and use identical evaluation protocols. The results are presented in Table 7. The base LLaVA-1.5-7B model struggles significantly with the image implication task, scoring 0% on TFQ, 16% on MCQ, and 2.06 on OSQ. After applying TFQ-GRPO, the model's performance improves dramatically across all three tasks: TFQ score increases to 6%, MCQ score more than doubles to 34% (+18%), and the OSQ score rises to 2.78 (+0.72). These substantial gains demonstrate that our training method is not a specialized fit for QwenVL but is a robust framework capable of enhancing the reasoning capabilities of diverse MLLM architectures.

6.4. Different Training Strategy

We explore the impact of different training strategies by comparing three distinct strategies: 1) TFQ-SFT: Supervised Fine-Tuning only. 2) TFQ-SFT & TFQ-GRPO: SFT as the warmup, followed by RL. 3) TFQ-GRPO: End-to-end RL, which is the main strategy used for MetaphorStar. For SFT, we utilize TFQ-Data-Lite-SFT, a dataset of 984 expert reasoning trajectories generated by Claude-3.7-thinking.

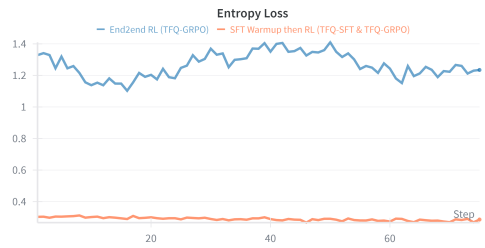

Figure 4. Entropy loss of models with different strategies.

The results in Table 8 lead to a critical finding. Our end-to-end RL method (+ TFQ-GRPO) yields the strongest performance on TFQ and MCQ. Conversely, both strategies involving SFT cause a catastrophic drop in MCQ performance (from 46% to 28%), indicating SFT severely damages model generalization. This exposes an important paradox: SFT-based methods score highest on the MLLM-judged OSQ task (3.66). We find the high score is an artifact. SFT models learn to be overly verbose, and this verbosity—which often includes contradictory viewpoints—is misinterpreted as comprehensive by the MLLM judge. Our RL model provides more concise and accurate answers, which are unfairly penalized. We term this phenomenon the "SFT Curse", technically explained by token entropy (Figure 4). The base model (1.33) and our end-to-end RL model (1.23) maintain high entropy, allowing for a broad exploration of the solution space. SFT, however, acts as an "entropy bottleneck," collapsing the model's policy to a low-entropy state (0.30) as it imitates a narrow data distribution. This low-entropy state persists even after RL (0.29), trapping the model in a local optimum focused on stylistic imitation rather than robust reasoning. In contrast, the end-to-end TFQ-GRPO leverages the model's high initial entropy to conduct a broader, more effective search for a global optimum. More details are in Appendix C.

7. Conclusion

We address the critical challenge of image implication, a form of sophisticated, non-literal reasoning where MLLMs currently struggle. We propose MetaphorStar, the visual reinforcement learning (RL) framework designed to bridge this gap. Our contributions include the True-False Question (TFQ) format for image implication tasks, along with the corresponding TFQ-Data and TFQ-Bench. We also develop TFQ-GRPO, the end-to-end RL training method, and release the MetaphorStar family of models, which achieve SOTA performance. Our experiments also reveal two crucial insights. First, image implication tasks can significantly enhance model performance on complex visual reasoning. Second, we identify the "SFT Curse", demonstrating that traditional SFT warmup creates the "entropy bottleneck" that harms generalization, and show that end-to-end RL methods are more suitable for image implication tasks, even visual reasoning tasks. We open-source all models, datasets, and code to help advance MLLMs beyond literal perception toward deeper, conceptual understanding.

References

[1] Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet, 2024. 6

[2] Anthropic. Introducing claude 4, 2025. 6

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[4] Sasa Baskarada and Andy Koronios. Data, information, knowledge, wisdom (dikw): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension. *Australasian Journal of Information Systems*, 2013. 2

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024. 12

[6] DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[7] Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *AAAI*, 2022. 2

[8] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACMMM*, 2024. 6

[9] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024. 12

[10] Gemini. Introducing gemini 2.0: our new ai model for the agentic era, 2024. 2, 3, 6

[11] Gemini. Gemini 2.0 model updates: 2.0 flash, flash-lite, pro experimental, 2025. 6

[12] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *ACL*, 2023. 2

[13] Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *ACL*, 2024. 2

[14] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024. 3

[15] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 12

[16] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008. 1

[17] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 12

[18] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 12

[19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*, 2023. 13

[20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 12

[21] Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024. 3

[22] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. 2

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023. 6, 8

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 12

[25] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895*, 2023. 12

[26] Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, et al. Ii-bench: An image implication understanding benchmark for multimodal large language models. In *NeurIPS*, 2024. 2, 3, 12

[27] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 12

[28] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024. 2

[29] Sachit Menon, Richard Zemel, and Carl Vondrick. Whiteboard-of-thought: Thinking step-by-step across modalities. *arXiv*, 2024. 3

[30] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[31] OpenAI. Learning to reason with llms, 2024. 3

[32] OpenAI. Introducing gpt-4.1 in the api, 2025. 6

[33] OpenAI. Openai o3 and o4-mini system card, 2025. 2, 3, 6

[34] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe

Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024. 12

[35] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, Vatsal Raina, Hanyi Xiong, Vishaal Udandarao, Jingyi Lu, Shiyang Chen, Sam Purkis, Tianshuo Yan, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025. 12

[36] ByteDance Seed. Doubao-1.5-thinking-vision-pro, 2025. 6

[37] Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025. 12

[38] Qwen Team. Qvq: To see the world with wisdom, 2024. 6

[39] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv:2402.14804*, 2024. 12

[40] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024. 2

[41] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025. 5

[42] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024. 3, 12

[43] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6

[44] xAI. Grok-1.5 vision preview, 2024. 12

[45] xAI. Grok 3 beta — the age of reasoning agents, 2025. 2, 6

[46] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. 12

[47] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 3

[48] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025. 12

[49] Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. Exploring chain-of-thought for multi-modal metaphor detection. In *ACL*, 2024. 2

[50] Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images? In *ACL*, 2024. 2

[51] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024. 3

[52] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 12

[53] Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024. 13

[54] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 2, 12

[55] Chenhao Zhang and Yazhe Niu. Let androids dream of electric sheep: A human-like image implication understanding and reasoning framework. *arXiv preprint arXiv:2505.17019*, 2025. 2, 5, 6, 12

[56] Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, et al. Can mllms understand the deep implication behind chinese images? *arXiv preprint arXiv:2410.13854*, 2024. 2, 3

[57] 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 *ECCV*, 2024. 12

[58] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 3

[59] Zhipu.ai. Glm-4v, 2024. 6

[60] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. *arXiv preprint arXiv:2312.02439*, 2024. 2

[61] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao,

et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6

A. Dataset Statistics

To construct our dataset, we leveraged the 1,434 high-quality metaphorical images from II-Bench [26]. II-Bench encompasses images from six distinct domains: Life, Art, Society, Psychology, Environment and Others. It features a diverse array of image types, including Illustrations, Memes, Posters, Multi-panel Comics, Single-panel Comics, Logos and Paintings. We manually construct the TFQ-Data-Lite and TFQ-Bench-Lite by selecting 50-100 high-quality, diverse and representative images. The general statistic is in Table 9 10 11.

Statistics of TFQ-Data & TFQ-Bench Images	
Life	516 (42.2%)
Art	70 (5.7%)
Society	408 (33.4%)
Psychology	127 (10.4%)
Environment	44 (3.6%)
Other	57 (4.7%)
Positive	169 (13.8%)
Neutral	702 (57.5%)
Negative	351 (28.7%)
Illustration	374 (28.7%)
Meme	269 (20.6%)
Poster	111 (8.5%)
Multi-panel Comic	311 (23.9%)
Single-panel Comic	90 (6.9%)
Logo	59 (4.5%)
Painting	89 (6.8%)

Table 9. General statistics of the TFQ-Data and TFQ-Bench.

Statistics of TFQ-Data-Lite Images	
Life	39 (39%)
Society	23 (23%)
Psychology	19 (19%)
Art	12 (12%)
Environment	6 (6%)
Others	1 (1%)
Multi-panel Comic	33 (28.7%)
Meme	22 (19.1%)
Illustration	20 (17.4%)
Poster	17 (14.8%)
Logo	15 (13.0%)
Single-panel Comic	7 (6.1%)
Painting	1 (0.9%)

Table 10. General statistics of the TFQ-Data-Lite.

B. Experiment Setup

Parameter Details. We set the model temperature as 0.5 and top_p as 0.9 in TFQ and MCQ experiments, and temperature as 0.7 and top_p as 0.9 in OSQ experiments. Additionally, we set the evaluation model GPT-4o temperature as 0 and evaluate more than three times to get the average score in OSQ experiments. And the average human-model scoring consistency reached 96.5% on OSQ [55]. All experiments are conducted on NVIDIA A800 and H200 GPUs.

Statistics of TFQ-Bench-Lite Images	
Society	21 (42%)
Life	16 (32%)
Art	6 (2%)
Psychology	4 (8%)
Others	3 (6%)
Multi-panel Comic	16 (32%)
Single-panel Comic	9 (18%)
Illustration	5 (10%)
Meme	5 (10%)
Poster	5 (10%)
Painting	5 (10%)
Logo	5 (10%)

Table 11. General statistics of the TFQ-Bench-Lite.

Main Experiment. To comprehensively compare with the MetaphorStar family, we carefully select a diverse range of MLLMs, encompassing both open-source and closed-source models, with the aim of covering a wide spectrum of model characteristics and scales. These models span parameter sizes from 7B to 300B, ensuring that models of varying complexity and capability are thoroughly assessed. In selecting the models, we focus on the following key aspects: 1) General and Reasoning models, 2) Open-Source and Closed-Source models, and 3) model parameter scaling law.

The high-level bench (EN) [55], which is manually constructed by 50 high-quality, diverse, and representative English images from varied image types like illustrations and comics, featuring Multiple-Choice Question (MCQ) and Open-Style Question (OSQ). And the average human-model scoring consistency reached 96.5% on OSQ [55].

Generalization Experiment. We mainly select two categories of benchmarks — Reasoning and Understanding.

We provide a comprehensive review of benchmarks specifically designed to assess various facets of MLLM reasoning capabilities, which are critical for their deployment in environments requiring complex decision-making. Therefore, we select MathVision [39], MathVerse [57], We-Math [34], LogicVista [46], VisuLogic [48], VisualPuzzles [37], V* [42], ZeroBench [35], and MMMU [54] to verify the model's reasoning ability.

We revisit multimodal understanding benchmarks designed to assess MLLMs' ability to perceive and comprehend information presented in various formats, such as text and images. These benchmarks are crucial for fine-tuning MLLMs, ensuring their robustness and generalization in real-world applications. These benchmarks include SEEDBench [17], SEED-2-Plus [18], MMBench (English) [24], MMBench v1.1 (English) [24], MMStar [5], OCRBench [25], AI2D [15], ScienceQA [27], POPE [20], MMT-Bench [52], RealWorld QA [44], BLINK [9], Hallu-

sionBench [19], and MMVet Hard [53].

C. Discussion

C.1. Why SFT Warmup Lose?

The ablation in Section 6 demonstrates that a conventional SFT warmup stage is not only unnecessary but is actively detrimental to performance on image implication tasks. This phenomenon, which we term the "SFT Curse," stems from a fundamental mismatch between the SFT objective and the nature of the task, as we analyze from three perspectives.

Task Nature: Creative Generalization. Image implication is not a simple pattern recognition task; it demands creative generalization—the ability to connect semantically distant concepts and generate novel, low-probability insights. Supervised Fine-Tuning, as a maximum likelihood objective, directly penalizes this. It trains the model to reproduce the "safe," high-probability sequences from the training data, acting as an "entropy bottleneck" (see Figure 4). This behavioral cloning teaches form over function, trapping the model in a "cognitive straitjacket." In contrast, end-to-end RL is driven purely by the reward signal. It is free to explore and reinforce these creative, low-probability reasoning paths as long as they lead to a correct answer, fostering the robust, abstract reasoning required for metaphors.

Question Format: The Talker vs. The Thinker. This "form over function" problem is most evident in the MCQ results. TFQ and MCQ are not purely generative tasks; they are highly discriminative. SFT trains the model to be a "talker"—to generate text that sounds plausible and adheres to the structural format (e.g., '¡think¿...¡/think¿'). It does not, however, train the model to be a "thinker"—to perform the underlying logical discrimination needed to identify and reject incorrect options. This explains the catastrophic collapse in MCQ performance (28% accuracy) for SFT-warmed models. The end-to-end RL model, by optimizing directly for the accuracy reward (R_{acc}), is forced to learn this crucial discriminative capability.

The OSQ Paradox: Evaluation Bias. This analysis also explains the "OSQ Paradox" in Table 8, where the objectively worse SFT+RL model achieves the highest subjective OSQ score. This is an artifact of the LLM-as-a-judge evaluation. The SFT-trained model produces verbose, well-structured outputs that often mix multiple (and sometimes contradictory) viewpoints. The LLM judge, relying on heuristics, misinterprets this stylistic adherence and verbosity as "deeper thought." The end-to-end RL model, which produces more concise and accurate answers, is unfairly penalized by this bias.

In summary, SFT warmup fails because it creates a low-entropy policy focused on imitation. The subsequent on-policy RL algorithm (GRPO) starts from this skewed distribution and is unable to escape this local optimum. End-to-end RL, by leveraging the high initial entropy of the base model, allows for a true, global search for the optimal reasoning policy.

C.2. Why Image Implication Tasks Can Help with Visual Reasoning?

Our generalization experiment (Table 5) confirms that training on image implication provides significant gains to downstream visual reasoning tasks and even benefits general VQA. We attribute this powerful generalization effect to key properties of the task itself and our training methodology.

Cultivating Multi-Hop Abstract Reasoning. At its core, image implication is a form of sophisticated, multi-hop abstract reasoning. Unlike standard VQA, which often requires literal, single-hop answers, implication tasks force the model to move from literal perception (e.g., "see a person") to abstract conceptualization (e.g., "understand the person represents a concept") and then to a final conclusion (e.g., "infer the relationship between concepts"). This process of connecting disparate concepts and performing non-literal inference trains the same underlying cognitive faculties required for formal logic, mathematical reasoning, and other complex visual reasoning benchmarks.

The Efficacy of the TFQ Format as a Reasoning Trainer. The benefits are not just from the what (metaphors) but the how (our TFQ format). As discussed in Section 3, TFQ has a high knowledge density, presenting the model with multiple fine-grained propositions to verify for a single image. This transforms the model from a simple "answer generator" into a "propositional verifier." This learned skill of methodically evaluating the truth value of specific claims is a core component that is highly transferable to all logical, mathematical, and sequential reasoning domains.

Simultaneous Grounding of Abstraction in Factual Perception. Our `TFQ-Data` design deliberately includes statements that probe basic visual facts alongside the central implication. This dual-objective training ensures that the model does not "drift" into ungrounded abstraction. It learns to simultaneously maintain its core perceptual accuracy (which benefits general VQA) while also building the new scaffolding for abstract inference. This forces the model to learn how to connect concrete visual evidence to abstract logical conclusions, a skill that is central to all robust reasoning.