METAPHORSTAR: IMAGE METAPHOR UNDERSTANDING AND REASONING WITH END-TO-END VISUAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding metaphors in images remains difficult for current AI systems. Multimodal Large Language Models (MLLMs) answer basic Visual Question Answering (VQA) questions well but often miss the cultural, emotional, and contextual implications of visual content, which require multi-hop reasoning, cultural context, and Theory of Mind (ToM). Reinforcement learning (RL) is a natural tool for eliciting such reasoning, but the target output of the task, an interpretation, cannot be verified automatically. We propose **MetaphorStar**, an end-to-end visual RL framework for image implication that addresses this by decomposing the reference interpretation of each image into a set of True-False propositions, each of which yields a binary, verifiable reward. The framework includes three components: the fine-grained dataset MetaphorQA, the visual RL method MetaphorGRPO, and the benchmark MetaphorBench. Trained with rewards computed only from these propositions and a format check, the MetaphorStar family (3B/7B/32B) raises the item-level True-False accuracy of the 7B base model from 28% to 70% and transfers to Multiple-Choice and Open-Style questions that are never seen during training; among 32 open and proprietary MLLMs, including the newest GPT-5.6-Sol, Gemini-3.1-Pro, and Qwen3.8-Max, MetaphorStar-32B obtains the highest True-False and Open-Style scores, while the strongest proprietary models remain ahead on Multiple-Choice questions. We further find that, in our setting, supervised warmup on distilled rationales lowers performance and collapses policy entropy, that the LLM judge commonly used for open-style evaluation ranks verbose SFT-trained models above more accurate RL models, and that the trained models also score above their base models on general visual reasoning benchmarks. We analyze parameter scaling, data scaling, reward weighting, and different backbones including Qwen3-VL and LLaVA. All models, data, and code will be released.

1 INTRODUCTION

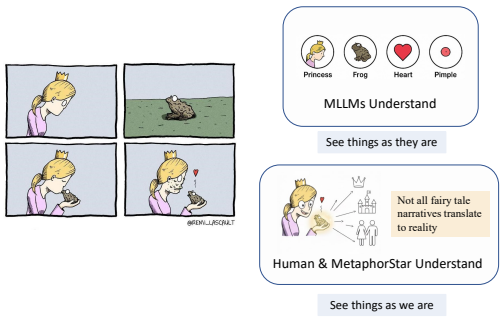
> *We don't see things as they are, we see them as we are.* — Anaïs Nin

This sentiment captures the gap between literal perception and conceptual understanding. "Seeing things as they are" is literal perception, identifying objects and describing a scene, at which Multimodal Large Language Models (MLLMs) excel. "Seeing things as we are" is implication, interpreting the scene through human context, culture, and shared knowledge, where MLLMs fail.

This gap is the essence of metaphorical comprehension (Figure 1). Metaphors are fundamental cognitive tools for conceptualizing our surroundings (Lakoff & Johnson, 2008), and visual metaphors are everywhere: an MLLM may see "a person" and "a split road" but fail to infer "a life-changing decision."

Vision-language reasoning models such as o3/o4-mini (OpenAI, 2025b), Gemini (Gemini, 2024; Team, 2025), and GPT-5.4 (OpenAI, 2026a) achieve outstanding results on math, code, and visual reasoning benchmarks (Lu et al., 2024; Yue et al., 2024), yet they still struggle with image metaphors (Liu et al., 2024; Zhang et al., 2024a): on our benchmark, the seven newest proprietary

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Implication: fairy-tale narratives do not always translate into reality.

Type	True-False Question (MetaphorBench)	Label
P	The princess's crown and action of kissing the frog are clear references to the classic fairy tale of "The Frog Prince."	T
P	The final panel shows the frog turning into a prince, which confirms the fairy tale expectation of transformation through love.	F
I	The comic humorously critiques the unrealistic expectations set by fairy tales, using the unchanged frog as a metaphor for reality.	T
D	The princess appears disappointed after kissing the frog, as the heart floating above them implies dissatisfaction.	F

Reward per question: $r_k = \mathbf{1}[\hat{\ell}_k = \ell_k]$, plus a format term.

Figure 1: **A picture is worth a thousand words.** Left: MLLMs recognize literal objects ("see things as they are") but miss the implication that humans and MetaphorStar grasp ("see things as we are"). Right: MetaphorStar turns the implication into verifiable True-False Questions on perception (P), implication (I), and plausible distractors (D), each a binary RL reward (4 of the 18 MetaphorBench questions of this comic, verbatim; all in Appendix Table 16).

models answer all True-False questions of an item correctly for only 38–70% of the items (Table 5). Image implication requires multi-hop reasoning and Theory of Mind (ToM) beyond logical deduction alone.

Existing methods use (1) explicit mapping between metaphor ontologies and visual representations (Zhong et al., 2024), (2) training-free chain-of-thought (Xu et al., 2024), or (3) contextual alignment with retrieved knowledge (Zhang & Niu, 2025), which struggle with many-to-many mappings, large abstract search spaces, or computational cost, respectively. Inspired by the DIKW pyramid (Baskarada & Koronios, 2013), we posit that the reasoning abilities needed for implication are latent in MLLMs and can be activated by reinforcement learning (RL). RL with verifiable rewards (RLVR) has driven recent progress in reasoning (OpenAI, 2024b; DeepSeek-AI et al., 2025; Shao et al., 2024), but it requires a checkable answer, and an implication is an interpretation. MCQ rewards give one signal per image, and open-style rewards need an LLM judge, which in our inspection favors verbose outputs (Section 6.4).

Our key idea is to decompose each image's reference implication into True-False Questions (TFQ): about ten binary propositions per image entry (9.8 on average) that cover its visual content, its implication, and plausible wrong readings. Each TFQ is checked by exact match against a reference label,[1] so metaphor understanding becomes an RLVR problem with about ten verifiable signals per image and no judge during training. Building on this, we propose **MetaphorStar**, an end-to-end visual RL framework for image implication, with three components: the fine-grained dataset **MetaphorQA**, the visual RL method **MetaphorGRPO**, and the benchmark **MetaphorBench**. Our contributions are:

- We propose MetaphorStar, to our knowledge the first end-to-end visual RL framework for image implication, including the dataset MetaphorQA, the method MetaphorGRPO, and the benchmark MetaphorBench, together with an analysis of the "SFT Curse": in our setting, SFT warmup on distilled rationales collapses policy entropy and lowers accuracy relative to end-to-end RL.
- Our MetaphorStar family, trained on only 984 TFQs, raises the 7B base model's TFQ accuracy from 28% to 70%, transfers to Multiple-Choice and Open-Style Questions unseen in training across Qwen2.5-VL, Qwen3-VL, and LLaVA, and at 32B obtains the highest TFQ and OSQ scores among 32 MLLMs, including the newest GPT-5.6-Sol, Gemini-3.1-Pro, and Qwen3.8-Max.
- MetaphorGRPO training is accompanied by higher general visual reasoning scores: the reasoning average (ten scores from nine benchmarks) rises by +0.7 to +3.1 while general understanding changes by 0.0 to +0.3.

[1] As in RLVR, *verifiable* means that the reward is computed deterministically against a reference label; it does not mean that an interpretation is objectively true.

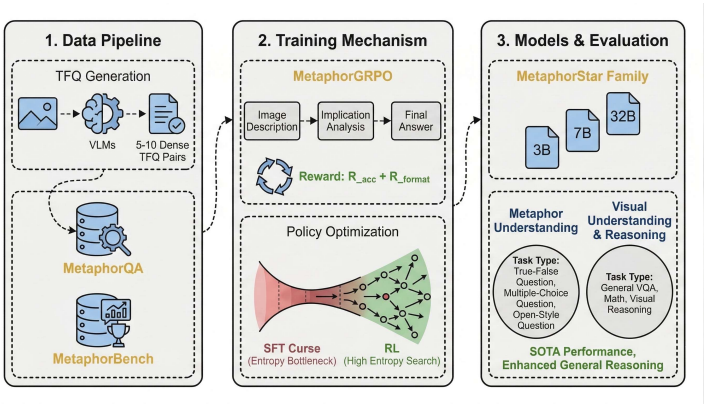


Figure 2: Overall framework of MetaphorStar: TFQ generation from II-Bench images and reference implications, MetaphorQA and MetaphorBench construction, end-to-end MetaphorGRPO training, and evaluation on TFQ, MCQ, OSQ, and general benchmarks, with analyses of training dynamics such as the SFT Curse.

2 Related Work

2.1 Image Implication

Prior work studied humor (Hessel et al., 2023; Horvitz et al., 2024), sarcasm (Desai et al., 2022), multimodal metaphor detection (Zhang et al., 2021; Xu et al., 2022; Zhang et al., 2023), visual metaphor generation and understanding (Akula et al., 2023; Chakrabarty et al., 2023; Kalarani et al., 2024), and meme and satire interpretation (Hwang & Shwartz, 2023; Nandy et al., 2024). DeepEval (Yang et al., 2024) proposed a taxonomy of deep semantics, and II-Bench (Liu et al., 2024) and CII-Bench (Zhang et al., 2024a) benchmark English and Chinese image implication. Existing methods, explicit mapping (CLOT; Zhong et al., 2024), prompting (C4MMD; Xu et al., 2024), and retrieval-based alignment (LAD; Zhang & Niu, 2025), are training-free; MetaphorStar instead trains the policy with RL.

2.2 Vision-language Reasoning

GRPO-based RLVR (DeepSeek-AI et al., 2025; Shao et al., 2024) has been extended to MLLMs, e.g., Visual-RFT (Liu et al., 2025), VLM-R1 (Shen et al., 2025), Vision-R1 (Huang et al., 2026), MM-Eureka (Meng et al., 2025), Perception-R1 (Yu et al., 2025), VL-Rethinker (Wang et al., 2025a), and DeepEyes (Zheng et al., 2025b). These target rule-checkable answers (math, counting, grounding). Without checkable answers, RL relies on reward models or LLM judges, which prefer longer responses (Zheng et al., 2023; Dubois et al., 2024), or on rubrics and checklists scored by a judge (Gunjal et al., 2025; Viswanathan et al., 2025); FActScore (Min et al., 2023) evaluates generation through atomic claims. Our TFQs follow this decomposition idea but carry reference labels, so the reward is computed by exact match without a judge. On training pipelines, SFT tends to memorize while RL generalizes (Chu et al., 2025), SFT on distilled rationales can hinder later RL in LVLMs (Chen et al., 2025), and policy entropy governs exploration (Cui et al., 2025; Wang et al., 2025b); our SFT Curse extends these findings to abstract visual understanding.

3 Method

3.1 True-False Question (TFQ) for Image Implication Understanding

II-Bench (Liu et al., 2024) introduced the Multiple-Choice Question (MCQ, six options), a balanced assessment of comprehension, and CII-Bench (Zhang et al., 2024a) the Open-Style Question (OSQ), scored by an LLM judge against a human reference, the upper bound of difficulty. We introduce the **True-False Question (TFQ)** as a fine-grained complement. Each image comes with statements $\{(s_k, \ell_k)\}_{k=1}^{K}$, $\ell_k \in \{\mathrm{T}, \mathrm{F}\}$, covering the central implication and the visual information it depends on (Figures 1 and 2). Table 1 compares the formats on three dimensions that matter for RL:

Table 1: Comparison of True-False, Multiple-Choice, and Open-Style Questions as RL training signals. Signals per entry for TFQ are the mean over MetaphorQA-Full; chance is per signal.

Format	Knowledge density	Learnability	Verifiability	
	signals / entry	chance / signal	checked by	judge in training
TFQ (ours)	9.8	1/2	exact match	no
MCQ	1	1/6	exact match	no
OSQ	1	–	LLM judge	yes

- **Knowledge density**: points checked per entry; TFQ is highest, with 9.8 per entry.
- **Learnability**: each TFQ gives a clean binary signal on one proposition rather than one signal for a whole interpretation.
- **Verifiability**: TFQ and MCQ are checked by exact match against reference labels, while OSQ requires a judge.

This makes TFQ a dense and verifiable reward substrate for visual RL. Random guessing obtains an expected reward of 0.5 per TFQ, and false statements are plausible distractors about the same image, so reward above chance requires separating true from false statements. Wording cues remain: marking statements with absolute words (e.g., "always") as false gives 56.5% question-level but 0% item-level accuracy on MetaphorBench (Appendix B). At evaluation, each TFQ is queried in its own prompt (so the alternating label order of the released files is never exposed), and an item counts as correct only if *all* of its TFQs are answered correctly, whose chance level 0.5^K is 0.1–0.4% for the test items ($K \in [8, 10]$).

3.2 METAPHORQA & METAPHORBENCH

Data generation. We build on II-Bench (Liu et al., 2024), which contains 1,222 unique high-quality metaphorical images with 1,434 image–question entries, each with a human-written description and implication. For each entry, multiple VLMs receive the description and the ground-truth implication and generate 5–10 TFQs with definitive True/False answers (4–10 retained per entry); propositions from different generators are cross-checked, and we keep those with consistent labels. Conditioned on the reference, the generators only rephrase and perturb a known interpretation, which is much easier than inferring it. We also manually verify the generated data, which yields 14,099 TFQs (examples in Table 3). Each set probes both the central metaphor and primary visual information (akin to basic VQA) with graded difficulty, with false statements crafted as plausible distractors.

Dataset and benchmark constructs. Table 2 lists the splits (statistics in Appendix Tables 9 and 10). **MetaphorQA-Full** (1,384 entries, 13,607 TFQs) is the large-scale training set, and **MetaphorQA-Lite** (100 entries, 984 TFQs) is a subset hand-picked for quality, diversity, and coverage of all domains and image types, used to train all MetaphorStar models. **MetaphorBench** contains 50 held-out entries (44 unique images, 492 TFQs); the High-Level Bench of Zhang & Niu (2025) poses its MCQ and OSQ on the same 50 entries, so all three formats are evaluated on the same items. Exact (MD5) and perceptual hashing (pHash) find no overlap between the 44 test images and the 1,178 unique training images. Labels are balanced (50.6–53.0% true), and every test item has both true and false TFQs, so a constant answer scores 0% at item level. For the training-strategy study (Section 6.4), **MetaphorQA-Lite-SFT** contains 984 reasoning trajectories generated by Claude-3.7-Sonnet (thinking) for the MetaphorQA-Lite TFQs.

3.3 METAPHORGRPO: END-TO-END VISUAL RL FOR ABSTRACT IMPLICATION

RL for abstract visual reasoning faces three challenges: (1) verifiable, dense *reward design* for a subjective task; (2) *training stability* without the entropy collapse of supervised warmup (Section 6.4); and (3) *cross-format generalization*. MetaphorGRPO addresses them with TFQ rewards, a composite reward over a structured template, and end-to-end RL without SFT warmup.

Reward formulation. Whereas MCQ/OSQ provide one sparse signal per image, TFQ acts as a high-density reward scaffold: about ten verifiable training signals per entry, although each rollout still receives one binary correctness reward. We instruct the model to first describe the image, then

Table 2: Statistics of the splits (Images: unique images by MD5). MetaphorBench-Full is the complete pool and is not a test set for models trained on MetaphorQA.

Type	Split	Entries	Images	TFQs
MetaphorQA	Lite	100	100	984
	Full	1,384	1,178	13,607
MetaphorBench	Lite (test)	50	44	492
	Full (pool)	1,434	1,222	14,099

Table 3: MetaphorQA TFQs for three images, quoted verbatim; object recognition alone is not sufficient to answer them.

Dim.	Question (paired with the image)	A
Cultural	The comic uses wordplay on the phrase "right to bear arms" to humorously explore multiple literal and figurative interpretations.	T
Emotional	The comic implies that emotions always lead to positive life decisions, because the heart is smiling.	F
Context	The smartphone screen acting as a doorway metaphorically represents how technology serves as a gateway between real life and the digital world.	T

Table 4: Training template of MetaphorGRPO (verbatim).

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: {}

analyze its implication, and finally reason to the answer (template in Table 4); each training input is one image with one TFQ. For a trajectory τ, the reward is

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

where R_{acc} is a binary reward for the correct T/F answer, R_{format} checks the *Image Description* $\rightarrow$ *Implication Analysis* $\rightarrow$ *Final Answer* tag structure, and $\alpha = 0.5$ (equal weighting).

GRPO. For each input x, the old policy $\pi_{\theta_{old}}$ samples a group of G rollouts $\{o_i\}_{i=1}^{G}$ with rewards $\{r_i\}$, and the advantage is $A_i = (r_i - \text{mean}(\{r_j\}))/\text{std}(\{r_j\})$. GRPO (Shao et al., 2024) maximizes

$$\mathcal{J}(\theta) = \mathbb{E}_{x,\{o_i\}\sim\pi_{\theta_{old}}}\Big[\frac{1}{G}\sum_{i=1}^{G}\min\big(\rho_i A_i,\ \text{clip}(\rho_i, 1-\varepsilon, 1+\varepsilon)A_i\big)\Big] - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big), \tag{2}$$

with $\rho_i = \pi_\theta(o_i \mid x)/\pi_{\theta_{old}}(o_i \mid x)$. MetaphorGRPO does not modify the GRPO objective; its contributions are the TFQ reward and the end-to-end recipe.

4 METAPHORSTAR FAMILY

The MetaphorStar family has three sizes (3B, 7B, 32B) built on Qwen2.5-VL (Bai et al., 2025b); we also apply the recipe to Qwen3-VL-8B (Bai et al., 2025a) and LLaVA-1.5-7B (Liu et al., 2023a).

Training setup. A two-stage pipeline with SFT warmup proved suboptimal because the warmup constrains the model's intrinsic reasoning (Section 6.4), so all models are trained with end-to-end MetaphorGRPO on MetaphorQA-Lite (984 TFQs), using EasyR1 (Zheng et al., 2025a), a verl-based framework (Sheng et al., 2024), with $G = 5$ and $\alpha = 0.5$ (all hyperparameters in Appendix Table 13).

Analyzing token entropy in reasoning. Figure 3 visualizes the token-level entropy of MetaphorStar-7B on TFQ, MCQ, and OSQ. High-entropy tokens are not randomly distributed, in line with the "high-entropy minority" that drives reasoning (Wang et al., 2025b). In these examples, entropy peaks at logical connectors that pivot the argument or establish a causal link (e.g., "therefore", "thus", "but"), and at key function words, quantifiers, and pronouns, suggesting that the model's effort concentrates on logical steps and on relating concepts. Low-entropy tokens typically reproduce factual details from the image or complete deterministic phrases.

5 EXPERIMENT

5.1 MAIN EXPERIMENT

We compare with 32 open and proprietary MLLMs, including seven of the newest proprietary models, all evaluated with the same protocols (Table 5; all in Appendix Tables 14 and 15). TFQ uses

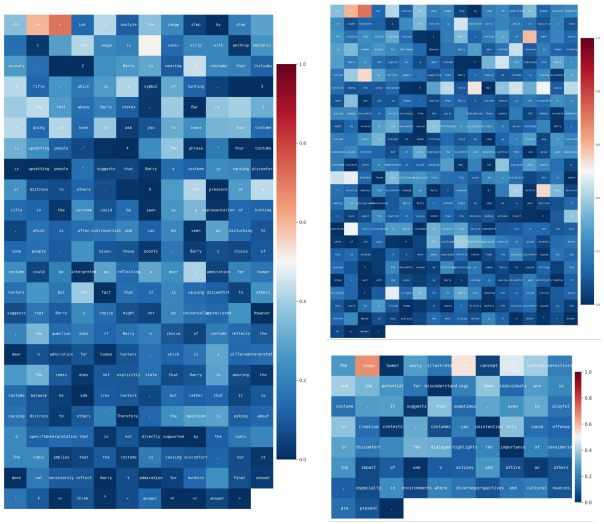


Figure 3: Token entropy of MetaphorStar-7B on TFQ (left), MCQ (top right), and OSQ (bottom right); red is high entropy (uncertainty), blue low entropy (confidence). Enlarged in Appendix Figure 8.

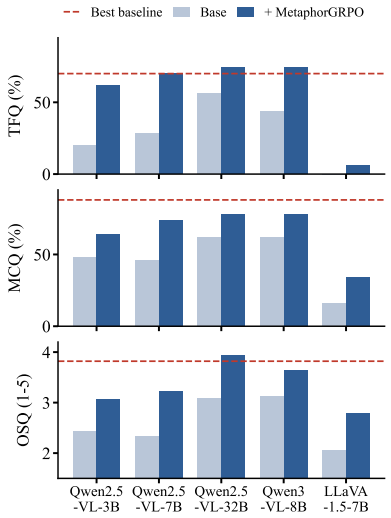


Figure 4: Base vs. MetaphorGRPO on five backbones; dashed: best of the 32 baselines per format (Table 5).

Table 5: Overall results on MetaphorBench (True-False Question) and High-Level Bench-EN (Multiple-Choice Question, Open-Style Question). Open-source models are listed by size, followed by proprietary models. The best result in each column is in **bold** and the second best is underlined. Subscripts give the change relative to the base model. Appendix Table 15 lists all 32 baselines, including GLM-4.1V-8B, QVQ-72B, and Doubao-1.5-Thinking-Vision-Pro.

Model	True-False Question	Multiple-Choice Question	Open-Style Question
Qwen2.5-VL-3B (Bai et al., 2025b)	20%	48%	2.44
LLaVA-1.5-7B (Liu et al., 2023a)	0%	16%	2.06
Qwen2.5-VL-7B (Bai et al., 2025b)	28%	46%	2.34
Qwen3-VL-8B (Bai et al., 2025a)	44%	62%	3.12
Qwen2.5-VL-32B (Bai et al., 2025b)	56%	62%	3.08
Qwen2.5-VL-72B (Bai et al., 2025b)	50%	72%	1.56
InternVL3-78B (Zhu et al., 2025)	36%	70%	3.42
Claude-Sonnet-5 (Anthropic, 2026c)	38%	58%	2.86
Claude-Opus-4.7 (Anthropic, 2026b)	42%	64%	3.02
Grok-4.6 (xAI, 2026)	56%	76%	3.56
Gemini-3.1-Pro (Google DeepMind, 2026a)	60%	**88%**	3.06
Qwen3.8-Max (Qwen Team, 2026)	70%	84%	3.80
Gemini-3.8-Flash (Google DeepMind, 2026b)	56%	78%	3.46
GPT-5.6-Sol (OpenAI, 2026b)	54%	78%	3.12
MetaphorGRPO models (trained on 984 propositions)			
LLaVA-1.5-7B + MetaphorGRPO	6%↑6	34%↑18	2.78↑0.72
Qwen3-VL-8B + MetaphorGRPO	**74%**↑30	78%↑16	3.64↑0.52
MetaphorStar-3B	62%↑42	64%↑16	3.06↑0.62
MetaphorStar-7B	70%↑42	74%↑28	3.22↑0.88
MetaphorStar-32B	**74%**↑18	78%↑16	**3.94**↑0.86

MetaphorBench (item-level accuracy), and the English High-Level Bench (Zhang & Niu, 2025) poses Multiple-Choice (MCQ, six options) and Open-Style Questions (OSQ, scored 1–5 by GPT-4o against the human reference with a fixed rubric, Appendix Figure 7) on the same 50 items (details in Appendix D). Bounding the exact McNemar p-value over all per-item pairings (Appendix Table 11), the TFQ gains of 3B, 7B, and Qwen3-VL-8B, the 7B MCQ gain, and the MCQ gap between end-to-end RL and SFT warmup are significant ($p < 0.05$) under every pairing. Three of the newest proprietary models were also evaluated on 200 further held-out items (Appendix Table 12).

Cross-format generalization. Trained only on binary TFQs, MetaphorStar improves MCQ (comparing distractors) and OSQ (free-form explanation) for every backbone (Figure 4). We hypothesize

Table 6: Results on general visual benchmarks (VLMEvalKit, identical prompts and decoding for base and trained models). Subscripts give the change relative to the base model: ↑ improvement, ↓ decline. Average changes are computed before rounding.

Benchmark	MetaphorStar-32B	MetaphorStar-7B	MetaphorStar-3B	Qwen2.5-VL-32B	Qwen2.5-VL-7B	Qwen2.5-VL-3B
Reasoning						
$\text{MMMU}_{\text{test}}$	49.8↑16.2	48.8↑6.8	45.1↑2.8	33.6	42.0	42.3
VisualPuzzles	39.7↑2.5	35.9↑2.2	33.8↑2.8	37.2	33.7	31.0
LogicVista	56.6↑1.6	47.2↑3.1	39.4	55.0	44.1	39.4
VisuLogic	25.5↓0.8	26.9↑2.2	18.8↓0.3	26.3	24.7	19.1
V*	81.2↑0.1	76.4↑5.2	34.0	81.1	71.2	34.0
$\text{ZeroBench}_{\text{main}}$	1.0↑1.0	1.0↑1.0	0.0	0.0	0.0	0.0
$\text{ZeroBench}_{\text{sub}}$	18.0↑2.4	15.3↑1.2	6.6↑1.2	15.6	14.1	5.4
MathVision	38.1↑0.7	25.3↑0.2	22.2↑1.0	37.4	25.1	21.2
$\text{MathVerse}_{\text{Vision}}$	50.8↑2.4	41.4↑6.2	30.0↑0.8	48.4	35.2	29.2
WeMath	48.6↑2.5	36.7↑2.4	21.7↓1.2	46.1	34.3	22.9
Avg.	40.9↑2.9	35.5↑3.1	25.2↑0.7	38.1	32.4	24.5
Understanding						
SEEDBench	77.6↑0.2	77.1↑0.1	74.0	77.4	77.0	74.0
SEEDBench2-Plus	73.2↑0.8	70.8↑0.1	63.6↑0.3	72.4	70.7	63.3
$\text{MMBench-V1.0-EN}_{\text{test}}$	85.8↓0.6	83.5	79.7↑0.6	86.4	83.5	79.1
$\text{MMBench-V1.1-EN}_{\text{test}}$	84.4↑0.4	82.5↑0.3	77.6↑0.8	84.0	82.2	76.8
MMStar	68.5↑2.2	64.1↑0.2	55.5↓0.4	66.3	63.9	55.9
OCRBench	86.1↑0.5	88.6↑2.2	81.8↑2.1	85.6	86.4	79.7
$\text{AI2D}_{\text{test}}$	83.3↑1.0	84.4↑0.5	81.2↑0.4	82.3	83.9	80.8
$\text{ScienceQA}_{\text{test}}$	91.3↑0.4	89.0	81.8↑0.4	90.9	89.0	81.4
POPE	86.3↑0.6	86.0↑0.1	86.2↑0.3	85.7	85.9	85.9
$\text{MMT-Bench}_{\text{val}}$	65.7↑0.4	62.3↑0.2	56.6↑0.3	65.3	62.1	56.3
RealWorldQA	71.1↑1.0	68.1↓0.4	62.4↓3.0	70.1	68.5	65.4
$\text{BLINK}_{\text{val}}$	62.9↓0.8	56.6↑1.3	44.5↓0.2	63.7	55.3	44.7
HallusionBench	55.3↓1.4	49.9↓1.8	46.5↑0.2	56.7	51.7	46.3
MMVet-Hard	59.6↓4.4	54.6↑1.7	50.2↓0.5	64.0	52.9	50.7
Avg.	75.1	72.7↑0.3	67.3↑0.1	75.1	72.4	67.2
Overall Avg.	60.9↑1.2	57.2↑1.5	49.7↑0.4	59.6	55.7	49.4

that judging all propositions of an image requires a complete reading of its metaphorical content rather than format-specific shortcuts.

True-False Question. MetaphorStar has the highest observed TFQ scores: 7B improves its base from 28% to 70%, 3B from 20% to 62% (on par with Gemini-3.1-Pro at 60%), and MetaphorStar-32B and Qwen3-VL-8B + MetaphorGRPO reach the highest score, 74%. The strongest proprietary model, Qwen3.8-Max (70%), is two items behind, a difference within noise (McNemar $p \geq 0.50$); the other newest models reach 38–60%, and MetaphorStar-32B's lead over GPT-5.6-Sol, Gemini-3.8-Flash, Grok-4.6, and Claude-Opus-4.7 is 18–32 points. Over the 25 earlier baselines (Appendix Table 15), reasoning models score higher than general models on average (TFQ 48.8% vs. 37.8%, MCQ 69.8% vs. 59.3%, OSQ 3.38 vs. 2.88).

Multiple-Choice Question. MetaphorStar-32B and Qwen3-VL-8B + MetaphorGRPO (78%) match or exceed all earlier baselines (e.g., Gemini-3.0-Pro 76%, GPT-5.4 74%; Appendix Table 15) and the newest GPT-5.6-Sol and Gemini-3.8-Flash (78%), but trail Gemini-3.1-Pro (88%) and Qwen3.8-Max (84%); MetaphorStar-7B improves its base from 46% to 74%. Within model families, however, reasoning variants do not consistently beat their general counterparts (Qwen2.5-VL-72B 72% vs. QVQ-72B 62%; Grok-3 66% vs. Grok-3-Reasoning 74% in Appendix Table 15), suggesting that RL on math or code does not reliably transfer to image implication.

Open-Style Question. MetaphorStar-32B obtains the best OSQ score (3.94), above all 32 baselines, including Gemini-3.0-Pro (3.82), Qwen3.8-Max (3.80), and Grok-4.6 (3.56), and MetaphorStar-7B improves its base by 0.88. Some models strong on MCQ are weak on OSQ (Qwen2.5-VL-72B: 72%, 1.56), consistent with option-selection bias (Zheng et al., 2024).

5.2 Generalization Experiment

Table 6 evaluates reasoning and understanding benchmarks with VLMEvalKit (Duan et al., 2024) (identical settings for base and trained models; Appendix D), although MetaphorQA contains no math, charts, or documents.

Reasoning. The reasoning average improves by +0.7, +3.1, and +2.9 (3B/7B/32B), e.g., +6.2 on MathVerse for 7B and +16.2 on $\text{MMMU}_{\text{test}}$ for 32B.

Understanding. The understanding average changes by 0.0 to +0.3 points (largest decrease: MMVet-Hard, −4.4 for 32B); the source of the reasoning gains is discussed in Section 6.5.

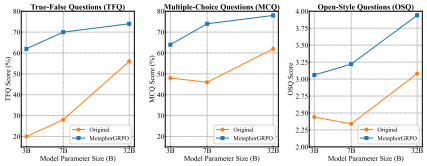


Figure 5: TFQ, MCQ, and OSQ of base and MetaphorGRPO models at 3B, 7B, and 32B.

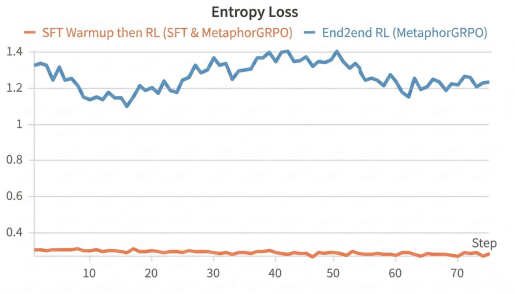


Figure 6: Entropy loss over RL steps for end-to-end MetaphorGRPO and for SFT warmup then MetaphorGRPO (Qwen2.5-VL-7B); base and SFT-only values are given in the text.

Table 7: Training data scaling and reward weight α (MetaphorStar-7B).

Data	**TFQ**	**MCQ**	**OSQ**
Small (0.1k)	48%	64%	3.04
Lite (1k)	70%	**74%**	3.22
Full (13.6k)	**84%**	**74%**	**3.48**
Reward weight	**TFQ**	**MCQ**	**OSQ**
$\alpha = 0.3$	66%	64%	3.22
$\alpha = 0.5$	**70%**	**74%**	3.22
$\alpha = 1.0$	not parseable		

Table 8: Results of different training strategies. LAD is from Zhang & Niu (2025).

Model	**TFQ**	**MCQ**	**OSQ**
Qwen2.5-VL-7B	28%	46%	2.34
+ Standard CoT	26%	44%	3.04
+ Few-shot CoT	18%	48%	2.94
+ LAD	–	64%	3.64
+ SFT	42%	28%	3.34
+ SFT & MetaphorGRPO	56%	28%	**3.66**
+ MetaphorGRPO	**70%**	**74%**	3.22
Qwen2.5-VL-32B	56%	62%	3.08
+ Standard CoT	50%	62%	3.32
+ Few-shot CoT	24%	40%	3.54
+ MetaphorGRPO	**74%**	**78%**	**3.94**

6 Ablation Study

6.1 Model Parameter Scaling

The base models scale inconsistently (OSQ: 7B 2.34 below 3B 2.44), whereas MetaphorStar improves monotonically on all formats (Figure 5), most on OSQ (3B→7B +0.16, 7B→32B +0.72).

6.2 Training Data Scaling

Training MetaphorStar-7B on 0.1k, 1k, and 13.6k TFQs (Table 7), even 0.1k TFQs raise TFQ from 28% to 48% and MCQ from 46% to 64%, and the full set reaches 84% TFQ, above MetaphorStar-32B trained on Lite (74%). For the reward weight, $\alpha = 0.3$ lowers TFQ and MCQ, and $\alpha = 1.0$ (no format reward) yields outputs that cannot be parsed; the default $\alpha = 0.5$ was fixed beforehand. A format-only reward ($\alpha = 0$) carries no label information and was not trained; eliciting the same describe-analyze-answer format by prompting (CoT) lowers TFQ at both 7B and 32B (Table 8).

6.3 Different Model Architecture

With the same data and parameters, MetaphorGRPO improves Qwen3-VL-8B, a newer and stronger backbone (TFQ 44%→74%, MCQ 62%→78%, OSQ 3.12→3.64), and LLaVA-1.5-7B, whose Vicuna (LLaMA-based) architecture differs from Qwen and whose base scores 0% on TFQ (TFQ 6%, MCQ 16%→34%, OSQ +0.72; bottom of Table 5). MetaphorGRPO thus also helps a non-Qwen model: its MCQ/OSQ gains (+18, +0.72) are comparable to Qwen's, while its TFQ gain is not significant ($p = 0.25$).

6.4 Different Training Strategy

We compare SFT, SFT warmup then RL, end-to-end MetaphorGRPO, and training-free methods (Table 8). CoT prompting raises OSQ but lowers TFQ (also at 32B: 50% and 24% for standard and few-shot CoT vs. 56% base and 74% MetaphorGRPO), as prompted models over-interpret literal details. LAD (with retrieval) has higher OSQ (3.64 vs. 3.22) but lower MCQ (64% vs. 74%).

The SFT Curse. SFT warmup, standard in RLVR pipelines, is harmful in our setting: both SFT-based strategies reduce MCQ below the base (46%→28%), and SFT&GRPO trails end-to-end GRPO on TFQ (56% vs. 70%). Policy entropy (Figure 6) stays high for the base model (1.33) and end-to-end RL (1.23) but collapses to 0.30 after SFT and is not recovered by RL (0.29). This extends Chen et al. (2025); Chu et al. (2025) to abstract visual reasoning.

The OSQ paradox. SFT&MetaphorGRPO obtains the highest 7B OSQ score (3.66) despite 28% MCQ accuracy. In our inspection, SFT models list several, sometimes contradictory, readings, one of which usually matches the reference, and the rubric rewards "detailed" and "comprehensive" answers (Appendix Figure 7), consistent with length preferences of LLM judges (Zheng et al., 2023; Dubois et al., 2024); a judge reward could thus favor verbosity, unlike TFQ rewards.

6.5 Discussion

Why does SFT warmup lose? A correct implication judgment can be reached through many reasoning paths, but SFT concentrates probability mass on one teacher trajectory, acting as an entropy bottleneck (Figure 6); GRPO then likely samples near-identical rollouts with uninformative advantages. SFT teaches the *form* of reasoning (a "talker"), not the discrimination TFQ and MCQ require (a "thinker").

Why might implication training help visual reasoning? Implication requires multi-hop steps from perception to abstraction, and TFQs train the model to verify claims, including visual facts, which keeps reasoning grounded; both skills serve visual reasoning (Table 6). Appendix H gives the full discussion.

7 Conclusion

MetaphorStar turns image implication, a task whose answers are open-ended interpretations, into a problem that reinforcement learning can optimize directly. It contributes at three levels. (1) *Reward design*: decomposing each reference interpretation into True-False propositions yields dense, verifiable rewards for a subjective visual task. (2) *Data*: MetaphorQA and MetaphorBench provide image-disjoint training and evaluation sets built on these propositions. (3) *Training paradigm*: end-to-end MetaphorGRPO preserves policy entropy, whereas SFT warmup on distilled rationales collapses it (the SFT Curse). Trained on only 984 propositions, MetaphorStar transfers to Multiple-Choice and Open-Style Questions on Qwen2.5-VL, Qwen3-VL, and LLaVA, obtains the highest True-False and Open-Style scores among 32 MLLMs at 32B, and raises general visual reasoning scores. More broadly, these results suggest that tasks judged by interpretation rather than by a single answer can still be trained with verifiable rewards once the interpretation is decomposed into checkable claims.

Ethics Statement

MetaphorQA and MetaphorBench are derived from the publicly released II-Bench (Liu et al., 2024) and follow its license. Some images contain satire or negative sentiment, and their interpretations mainly reflect English-language cultural conventions, which trained models may reproduce. The data contain no personal information beyond what is visible in the published images.

Reproducibility Statement

Section 3.1 and Section 3.3 define the proposition format, the reward (Eq. 1), and the GRPO objective; Section 3.2 and Table 2 define all data splits and the leakage control; Appendix C gives

the training template; Appendix D lists decoding parameters, the judge protocol, and all evaluation benchmarks. Training uses EasyR1 (Zheng et al., 2025a), built on verl (Sheng et al., 2024); all hyperparameters are listed in Appendix Table 13. All model weights, data splits, SFT trajectories, and training and evaluation code will be released upon publication.

AI USE STATEMENT

Large language models were used only to polish the language of this paper. The use of VLMs to generate propositions and of GPT-4o as the Open-Style Question judge is part of the method and is described in Section 3.2 and Appendix D.

REFERENCES

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. Metaclue: Towards comprehensive visual metaphors research. In *CVPR*, 2023.

Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet, 2024. URL https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf.

Anthropic. Introducing claude 4, 2025. URL https://www.anthropic.com/news/claude-4.

Anthropic. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026a.

Anthropic. Claude Opus 4.7. API model `claude-opus-4-7`, 2026b.

Anthropic. Claude Sonnet 5. API model `claude-sonnet-5`, 2026c.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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*, 2025b.

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.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.

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

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *ICML*, 2025.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

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

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

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.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

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.

Gemini. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Gemini. Gemini 2.0 model updates: 2.0 flash, flash-lite, pro experimental, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/.

Google DeepMind. Gemini 3.1 Pro. API model `gemini-3.1-pro-preview`, 2026a.

Google DeepMind. Gemini 3.8 Flash. API model `gemini-3.8-flash`, 2026b.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

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.

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.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *ICLR*, 2026.

EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes. In *EMNLP*, 2023.

Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, and Sumit Shekhar. Unveiling the invisible: Captioning videos with metaphors. *arXiv preprint arXiv:2406.04886*, 2024.

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

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago Press, 2008.

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*, 2023a.

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.

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*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023a.

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*, 2023b.

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*, 2023c.

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.

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

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.

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.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, 2023.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. In *EMNLP*, 2024.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024a.

OpenAI. Learning to reason with llms, 2024b. URL https://openai.com/index/learning-to-reason-with-llms/.

OpenAI. Introducing gpt-4.1 in the api, 2025a. URL https://openai.com/index/gpt-4-1/.

OpenAI. Openai o3 and o4-mini system card, 2025b. URL https://openai.com/index/o3-o4-mini-system-card/.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026a.

OpenAI. GPT-5.6-Sol. API model `gpt-5.6-sol`, 2026b.

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.

Qwen Team. Qwen3.8-Max. API model `qwen3.8-max`, 2026.

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.

ByteDance Seed. Doubao-1.5-thinking-vision-pro, 2025. URL https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-vision-pro.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.

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.

Gemini Team. Gemini 3 pro: Best for complex tasks and bringing creative concepts to life, 2025. URL https://deepmind.google/models/gemini/pro/.

Qwen Team. Qvq: To see the world with wisdom, 2024. URL https://qwenlm.github.io/blog/qvq-72b-preview/.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *NeurIPS*, 2025.

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

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.

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*, 2025b.

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

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.

xAI. Grok-1.5 vision preview, 2024. URL https://x.ai/news/grok-1.5v.

xAI. Grok 3 beta — the age of reasoning agents, 2025. URL https://x.ai/news/grok-3.

xAI. Grok 4.6. API model `grok-4.6`, 2026.

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

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. Met-meme: A multimodal meme dataset rich in metaphors. In *SIGIR*, 2022.

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.

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

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

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.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

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.

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.

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.

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

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *ACL*, 2021.

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. Multicmet: A novel chinese benchmark for understanding multimodal metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

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*, 2024b.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025a.

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*, 2025b.

Zhipu.ai. Glm-4v, 2024. URL https://www.bigmodel.cn/dev/api/normal-model/glm-4v.

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.

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

A Dataset statistics

II-Bench spans six domains and seven image types. Table 9 reports the distribution over the 1,222 unique images (with sentiment labels), MetaphorQA-Lite, and MetaphorBench. For all images and MetaphorQA-Lite, domain and sentiment counts sum to the number of images; MetaphorBench counts are per entry (50 entries, 44 unique images). Image-type counts are taken from the II-Bench annotation and sum to 1,303 for all images and 115 for MetaphorQA-Lite. MetaphorQA-Lite was selected manually to cover all domains and image types; MetaphorBench covers all image types and five of the six domains (no Environment image).

Table 9: Image statistics. Percentages are relative to each split.

All images (1,222)		**MetaphorQA-Lite (100)**		**MetaphorBench (50 entries)**	
Life	516 (42.2%)	Life	39 (39%)	Society	21 (42%)
Society	408 (33.4%)	Society	23 (23%)	Life	16 (32%)
Psychology	127 (10.4%)	Psychology	19 (19%)	Art	6 (12%)
Art	70 (5.7%)	Art	12 (12%)	Psychology	4 (8%)
Environment	44 (3.6%)	Environment	6 (6%)	Other	3 (6%)
Other	57 (4.7%)	Other	1 (1%)		
Illustration	374	Multi-panel comic	33	Multi-panel comic	16 (32%)
Meme	269	Meme	22	Single-panel comic	9 (18%)
Multi-panel comic	311	Illustration	20	Illustration	5 (10%)
Poster	111	Poster	17	Meme	5 (10%)
Single-panel comic	90	Logo	15	Poster	5 (10%)
Painting	89	Single-panel comic	7	Painting	5 (10%)
Logo	59	Painting	1	Logo	5 (10%)
Positive / Neutral / Negative	169 / 702 / 351				

B Proposition statistics, baselines, and paired significance

Proposition statistics. Table 10 summarizes the propositions of each split. Labels are close to balanced, and almost all entries have 8–10 propositions. In MetaphorQA-Full, 160 images are shared by several II-Bench entries (366 entries); MetaphorQA-Lite has no shared images, and MetaphorBench has six images with two entries each.

Table 10: Proposition statistics. Length is the number of words per proposition.

Split	**Entries**	**Images**	**Propositions**	**Per entry (min–max)**	**True**	**Length**
MetaphorQA-Lite	100	100	984	9.84 (8–10)	51.8%	21.1
MetaphorQA-Full	1,384	1,178	13,607	9.83 (4–10)	53.0%	21.0
MetaphorBench	50	44	492	9.84 (8–10)	50.6%	19.7

Constant and random baselines. On MetaphorBench, answering True for every proposition gives 50.6% question-level accuracy and answering False gives 49.4%; both give 0% item-level strict accuracy, because every item contains both true and false propositions. Uniform random guessing has an expected item-level strict accuracy of 0.12%. The item-level metric in Table 5 is therefore not attainable through label priors. Wording cues exist as well: 15.6% of the false but only 3.6% of the true MetaphorBench TFQs contain an absolute word (always, never, all, only, every, none, completely, entirely, solely, exclusively, must, nothing, no one, any). A rule that marks such statements false and all others true reaches 56.5% question-level accuracy (60.7% on MetaphorQA-Lite) but 0% item-level accuracy on MetaphorBench, so the item-level metric is also robust to this cue. **Order of TFQs.** In the released files, the TFQs of each entry are stored with alternating labels (T, F, T, F, ...). All evaluations in this paper query each TFQ in a separate prompt, so this order is never visible to a model; for evaluators who present several TFQs together, we will release a shuffled version.

Paired significance bounds. For two models evaluated on the same $N{=}50$ items with a and b correct items, the numbers of discordant items are not determined by the accuracies alone. We enumerate every joint outcome consistent with a and b and report the smallest and largest two-sided exact McNemar p-value (Table 11). An upper bound below 0.05 means the difference is significant whatever the per-item pairing is.

Table 11: Range of exact McNemar p-values over all per-item pairings consistent with the reported accuracies (N=50 items).

Comparison	Format	Accuracy (%)	p range
MetaphorStar-3B vs. Qwen2.5-VL-3B	TFQ	62 vs. 20	$[9.5\times10^{-7}, 0.002]$
MetaphorStar-7B vs. Qwen2.5-VL-7B	TFQ	70 vs. 28	$[9.5\times10^{-7}, 0.004]$
MetaphorStar-32B vs. Qwen2.5-VL-32B	TFQ	74 vs. 56	[0.004, 0.18]
Qwen3-VL-8B + MetaphorGRPO vs. base	TFQ	74 vs. 44	$[6.1\times10^{-5}, 0.028]$
LLaVA-1.5-7B + MetaphorGRPO vs. base	TFQ	6 vs. 0	[0.25, 0.25]
MetaphorStar-3B vs. Qwen2.5-VL-3B	MCQ	64 vs. 48	[0.008, 0.29]
MetaphorStar-7B vs. Qwen2.5-VL-7B	MCQ	74 vs. 46	$[1.2\times10^{-4}, 0.038]$
MetaphorStar-32B vs. Qwen2.5-VL-32B	MCQ	78 vs. 62	[0.008, 0.20]
Qwen3-VL-8B + MetaphorGRPO vs. base	MCQ	78 vs. 62	[0.008, 0.20]
LLaVA-1.5-7B + MetaphorGRPO vs. base	MCQ	34 vs. 16	[0.004, 0.11]
MetaphorGRPO vs. SFT→MetaphorGRPO (7B)	MCQ	74 vs. 28	$[2.4\times10^{-7}, 0.001]$
MetaphorGRPO vs. SFT→MetaphorGRPO (7B)	TFQ	70 vs. 56	[0.016, 0.32]
MetaphorStar-32B vs. Doubao-1.5-Thinking-Vision-Pro	TFQ	74 vs. 62	[0.031, 0.38]
MetaphorStar-32B vs. GPT-5.4	TFQ	74 vs. 58	[0.008, 0.23]
MetaphorStar-32B vs. Grok-4.6 / Gemini-3.8-Flash	TFQ	74 vs. 56	[0.004, 0.18]
MetaphorStar-32B vs. GPT-5.6-Sol	TFQ	74 vs. 54	[0.002, 0.13]
MetaphorStar-32B vs. Qwen3.8-Max	TFQ	74 vs. 70	[0.50, 0.85]
MetaphorStar-32B vs. Gemini-3.1-Pro	TFQ	74 vs. 60	[0.016, 0.30]
MetaphorStar-32B vs. Claude-Opus-4.7	TFQ	74 vs. 42	$[3\times10^{-5}, 0.020]$
MetaphorStar-32B vs. Claude-Sonnet-5	TFQ	74 vs. 38	$[7.6\times10^{-6}, 0.010]$

A larger held-out sample for API baselines. To measure how much the 50-item estimates change with more items, we sampled 200 further entries (1,962 TFQs, fixed seed) from the 1,236 MetaphorBench-Full entries whose images appear in neither MetaphorQA-Lite, the training set of the MetaphorStar models in Table 5, nor MetaphorBench. Three of the newest API models were evaluated with the same protocol (one TFQ per prompt, item-level strict accuracy; Table 12). All three score higher on the 200 items than on MetaphorBench, by 9.0 (GPT-5.6-Sol), 2.0 (Gemini-3.8-Flash), and 16.5 (Claude-Sonnet-5) points, and for each model the two 95% intervals overlap. GPT-5.6-Sol and Gemini-3.8-Flash swap order, and Claude-Sonnet-5 is the lowest on both samples. The MetaphorStar models were not evaluated on these 200 items, so the table compares the baselines across the two samples and does not compare them with MetaphorStar.

Table 12: Three API baselines on MetaphorBench (50 items) and on 200 further held-out MetaphorBench-Full items. Item-level strict accuracy with Wilson 95% intervals; question-level accuracy in the last column.

Model	MetaphorBench (50)	200 further items	Question-level (50 / 200)
GPT-5.6-Sol	54.0 [40.4, 67.0]	63.0 [56.1, 69.4]	92.5 / 94.0
Gemini-3.8-Flash	56.0 [42.3, 68.8]	58.0 [51.1, 64.6]	93.7 / 93.3
Claude-Sonnet-5	38.0 [25.9, 51.8]	54.5 [47.6, 61.3]	90.0 / 92.8

C Prompts and evaluation rubric

The training template (Table 4) is reproduced verbatim; the placeholder is filled with one proposition per training input. The format reward R_{format} checks that the output follows the `<think>...</think> <answer>...</answer>` structure, and the accuracy reward compares the T/F answer inside `<answer>` with the reference label.

Open-style judge. OSQ answers are scored by GPT-4o with the rubric and prompt of Zhang & Niu (2025), shown in Figure 7. The rubric assesses surface-level information, emotional expression, domain and context, rhetorical skills, and deep implications, and maps them to a 1–5 score. The two highest levels require "detailed and precise" and "comprehensive and precise" descriptions, which is relevant to the length preference discussed in Section 6.4.

Evaluation Metric	Evaluation Standard
1. Surface-level Information: • Identification of primary entities within the image • Analysis of color composition and application • Recognition of intricate details and their significance **2. Emotional Expression:** • Identification of conveyed emotions (e.g., tranquility, intensity, melancholy) • Depth of emotional resonance and its alignment with the image's theme • Consistency of emotional expression across the image's elements **3. Domain and Context:** • Recognition of the image's domain (e.g., art, commerce, social commentary) • Contextualization within its cultural, historical, or societal background • Evaluation of the image's innovation within its domain **4. Rhetorical Skills:** • Identification of rhetorical devices (e.g., symbolism, contrast, personification) • Analysis of how rhetorical techniques enhance the image's expression • Integration of rhetorical devices with metaphorical implications to create a cohesive interpretation **5. Deep Implications:** • Recognition of metaphorical elements and their layered meanings • Depth of interpretation of philosophical, cultural, or social values embedded in the image • Evaluation of the originality and creativity in metaphorical interpretation	**[1 point]:** Fails to capture key elements within the image (such as text, and important entities). Does not identify emotions, domain, or rhetorical devices. Only provides a superficial description of surface-level information, lacking depth and creativity, with a significant gap from the standard answer. **[2 points]:** Captures some key elements within the image, but the identification of emotions, domain, and rhetorical devices is vague. The description of surface-level information is relatively complete, but there is a clear deficiency in exploring deeper meanings, showing a noticeable gap from the standard answer. **[3 points]:** Effectively captures key elements within the image and initially identifies emotions, domain, and rhetorical devices. The description of surface-level information is relatively accurate, and there is some relevant expression of deep meanings. However, there is still room for improvement in depth and creativity, and it is generally close to the standard answer. **[4 points]:** Accurately captures key elements within the image and clearly identifies emotions, domain, and rhetorical devices. The description of surface-level information is detailed and precise, with a relatively deep exploration of deep meanings, demonstrating a certain level of creativity and depth. It is largely consistent with the standard answer but may have minor deficiencies in some details or depth. **[5 points]:** Accurately and precisely captures key elements within the image and profoundly identifies emotions, domain, and rhetorical devices. The description of surface-level information is comprehensive and precise, with unique insights into deep meanings, skillfully integrating image elements with metaphorical implications. It demonstrates exceptional creativity and depth, is highly consistent with the standard answer, and shows a profound grasp of metaphor creation and cultural understanding.

(a) Evaluation dimensions (left) and 1–5 scoring standard (right).

```
# Role
You are an impartial judge who is familiar with Internet culture and memes, and is good at digging out and analyzing the deep meaning of Internet memes.

## Attention
You are responsible for evaluating the quality of the answer provided by the model for Internet culture and memes. Your evaluation should refer to the human answer and image, and score based on the Evaluation Standard.

## Evaluation Standard
- [1 point]:
Fails to capture key elements within the image (such as text, and important entities). Does not identify emotions, domain, or rhetorical devices. Only provides a superficial description of surface-level information, lacking depth and creativity, with a significant gap from the standard answer.
- [2 points]:
Captures some key elements within the image, but the identification of emotions, domain, and rhetorical devices is vague. The description of surface-level information is relatively complete, but there is a clear deficiency in exploring deeper meanings, showing a noticeable gap from the standard answer.
- [3 points]:
Effectively captures key elements within the image and initially identifies emotions, domain, and rhetorical devices. The description of surface-level information is relatively accurate, and there is some relevant expression of deep meanings. However, there is still room for improvement in depth and creativity, and it is generally close to the standard answer.
- [4 points]:
Accurately captures key elements within the image and clearly identifies emotions, domain, and rhetorical devices. The description of surface-level information is detailed and precise, with a relatively deep exploration of deep meanings, demonstrating a certain level of creativity and depth. It is largely consistent with the standard answer but may have minor deficiencies in some details or depth.
- [5 points]:
Accurately and precisely captures key elements within the image and profoundly identifies emotions, domain, and rhetorical devices. The description of surface-level information is comprehensive and precise, with unique insights into deep meanings, skillfully integrating image elements with metaphorical implications. It demonstrates exceptional creativity and depth, is highly consistent with the standard answer, and shows a profound grasp of metaphor creation and cultural understanding.

## Standrad Answer:
Human answer: {}

## Constraints
- Avoid any position biases and be as objective as possible
- Do not allow the length of the descriptions to influence your evaluation
- Output your final verdict by strictly following this format: "[ratings]"

## Solve:
Model answer: {}
```

(b) Judge prompt (GPT-4o, temperature 0, three runs averaged).

Figure 7: Open-Style Question evaluation of Zhang & Niu (2025), used without modification.

D EXPERIMENTAL DETAILS

Decoding and judging. We use temperature 0.5 and top-p 0.9 for TFQ and MCQ, and temperature 0.7 and top-p 0.9 for OSQ. The OSQ judge is GPT-4o with temperature 0, scoring each answer against the human-written reference with the rubric of Zhang & Niu (2025); each answer is judged three times and the scores are averaged. Zhang & Niu (2025) report 95.7% consistency between this judge and 16 human raters on English and Chinese OSQ responses of general MLLMs. That study validates the judge on typical model outputs; it does not cover the verbose outputs of SFT-trained models discussed in Section 6.4. Experiments are run on NVIDIA A800 and H200 GPUs.

Training. MetaphorGRPO is implemented with EasyR1 (Zheng et al., 2025a), built on verl (Sheng et al., 2024), with the settings of Table 13 and the reward of Eq. 1 with α=0.5. Each training input is one image and one proposition. The data-scaling runs (Table 7) use subsets of 0.1k, 1k (MetaphorQA-Lite), and 13.6k (MetaphorQA-Full) propositions with otherwise identical settings. The SFT runs (Table 8) fine-tune on MetaphorQA-Lite-SFT; SFT→MetaphorGRPO then applies MetaphorGRPO on MetaphorQA-Lite from the SFT checkpoint.

Training hyperparameters. Table 13 lists the MetaphorGRPO settings, identical for the 3B, 7B, and 32B models (only the number of GPUs and the rollout tensor parallelism differ). The accuracy and format rewards follow Eq. 1: the format reward is 1 if the output fully matches `<think>...</think>...<answer>...</answer>`, and the accuracy reward is 1 if the content of `<answer>` matches the label.

Table 13: MetaphorGRPO training hyperparameters (EasyR1).

Hyperparameter	Value	Hyperparameter	Value
Training data	MetaphorQA-Lite (984 TFQs)	Rollouts per prompt G	5
Epochs	5	Rollout temperature / top-p	1.0 / 0.99
Rollout batch size	64	Max prompt / response length	5500 / 2048
Global (update) batch size	32	Advantage estimator	GRPO
Optimizer	AdamW	KL loss (low-variance) coefficient β	0.01
Learning rate	1×10^{-6} (no warmup)	Reward weight α	0.5
Weight decay	0.01	Max gradient norm	1.0
Image pixels (min / max)	262,144 / 4,194,304	Vision tower	trained (not frozen)
GPUs	2 (3B), 8 (7B, 32B)	Validation decoding	temperature 0.5

Evaluation of the newest models. The seven newest models (Table 15, marked ‡) were evaluated later than the other baselines, through an OpenAI-compatible API gateway, with the baseline prompt of the other models, one TFQ per request, temperature 0.5, and top-p 0.9. Served model versions and sampling at this temperature can shift item-level accuracy by several items between runs on 50 items. **Baselines.** The 25 baselines cover general and reasoning models, open-source models from 3B to 78B parameters, and proprietary models (Table 14).

Table 14: Baseline models.

General models	Reasoning models
Qwen2.5-VL-3B/7B/32B/72B (Bai et al., 2025b)	QVQ-72B (Team, 2024)
Qwen3-VL-8B (Bai et al., 2025a)	o4-mini (OpenAI, 2025b)
LLaVA-1.5-7B (Liu et al., 2023a)	Grok-3-Reasoning (xAI, 2025)
DeepSeek-VL2 (Wu et al., 2024)	Gemini-2.5-Flash-Thinking (Gemini, 2025)
GLM-4.1V-8B, GLM-4V-Plus (Zhipu.ai, 2024)	Doubao-1.5-Thinking-Vision-Pro (Seed, 2025)
InternVL3-78B (Zhu et al., 2025)	Claude-4.6-Opus (Anthropic, 2026a)
GPT-4o-mini, GPT-4o (OpenAI, 2024a); GPT-4.1 (OpenAI, 2025a)	GPT-5.4 (OpenAI, 2026a)
Gemini-2.5-Flash (Gemini, 2024); Grok-3 (xAI, 2025)	Gemini-3.0-Pro (Team, 2025)
Claude-3.5-Sonnet (Anthropic, 2024); Claude-4.0-Sonnet (Anthropic, 2025)	

General benchmarks. Reasoning: MathVision (Wang et al., 2024), MathVerse (Zhang et al., 2024b), WeMath (Qiao et al., 2024), LogicVista (Xiao et al., 2024), VisuLogic (Xu et al., 2025),

Table 15: Complete version of Table 5 with all 32 baselines and five MetaphorGRPO models. Overall results on MetaphorBench (True-False Question, item-level strict accuracy) and High-Level Bench-EN (Multiple-Choice Question accuracy, Open-Style Question score 1–5), all on the same 50 held-out items. The best result in each column is in **bold** and the second best is underlined. Subscripts give the change of each MetaphorGRPO model relative to its base model. ‡Newest models, evaluated later with the same protocols as the other baselines (one TFQ per prompt).

Model	True-False Question	Multiple-Choice Question	Open-Style Question
General models			
Qwen2.5-VL-3B (Bai et al., 2025b)	20%	48%	2.44
LLaVA-1.5-7B (Liu et al., 2023a)	0%	16%	2.06
Qwen2.5-VL-7B (Bai et al., 2025b)	28%	46%	2.34
DeepSeek-VL2 (Wu et al., 2024)	20%	46%	2.82
GLM-4.1V-8B (Zhipu.ai, 2024)	38%	60%	2.60
Qwen3-VL-8B (Bai et al., 2025a)	44%	62%	3.12
Qwen2.5-VL-32B (Bai et al., 2025b)	56%	62%	3.08
GPT-4o-mini (OpenAI, 2024a)	36%	44%	2.98
Gemini-2.5-Flash (Gemini, 2024)	56%	76%	3.34
Qwen2.5-VL-72B (Bai et al., 2025b)	50%	72%	1.56
InternVL3-78B (Zhu et al., 2025)	36%	70%	3.42
GLM-4V-Plus (Zhipu.ai, 2024)	42%	64%	3.01
Grok-3 (xAI, 2025)	36%	66%	3.24
Claude-3.5-Sonnet (Anthropic, 2024)	38%	68%	3.22
Claude-4.0-Sonnet (Anthropic, 2025)	52%	60%	3.46
GPT-4o (OpenAI, 2024a)	50%	74%	2.94
GPT-4.1 (OpenAI, 2025a)	40%	74%	3.30
Vision-language reasoning models			
Gemini-2.5-Flash-Thinking (Gemini, 2025)	54%	78%	3.42
QVQ-72B (Team, 2024)	28%	62%	3.10
o4-mini (OpenAI, 2025b)	42%	58%	3.26
Doubao-1.5-Thinking-Vision-Pro (Seed, 2025)	62%	66%	3.16
Grok-3-Reasoning (xAI, 2025)	36%	74%	3.06
Claude-4.6-Opus (Anthropic, 2026a)	52%	70%	3.60
GPT-5.4 (OpenAI, 2026a)	58%	74%	3.64
Gemini-3.0-Pro (Team, 2025)	58%	76%	3.82
Claude-Sonnet-5‡ (Anthropic, 2026c)	38%	58%	2.86
Claude-Opus-4.7‡ (Anthropic, 2026b)	42%	64%	3.02
GPT-5.6-Sol‡ (OpenAI, 2026b)	54%	78%	3.12
Gemini-3.8-Flash‡ (Google DeepMind, 2026b)	56%	78%	3.46
Grok-4.6‡ (xAI, 2026)	56%	76%	3.56
Gemini-3.1-Pro‡ (Google DeepMind, 2026a)	60%	**88%**	3.06
Qwen3.8-Max‡ (Qwen Team, 2026)	70%	84%	3.80
MetaphorGRPO models (trained on 984 propositions)			
LLaVA-1.5-7B + MetaphorGRPO	6%↑6	34%↑18	2.78↑0.72
MetaphorStar-3B	62%↑42	64%↑16	3.06↑0.62
MetaphorStar-7B	70%↑42	74%↑28	3.22↑0.88
Qwen3-VL-8B + MetaphorGRPO	**74%**↑30	78%↑16	3.64↑0.52
MetaphorStar-32B	**74%**↑18	78%↑16	**3.94**↑0.86

VisualPuzzles (Song et al., 2025), V* (Wu & Xie, 2024), ZeroBench (Roberts et al., 2025) (main and sub-questions), and MMMU (Yue et al., 2024). Understanding: SEEDBench (Li et al., 2023a), SEED-2-Plus (Li et al., 2024), MMBench v1.0 and v1.1 (English) (Liu et al., 2023b), MMStar (Chen et al., 2024), OCRBench (Liu et al., 2023c), AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), POPE (Li et al., 2023b), MMT-Bench (Ying et al., 2024), RealWorldQA (xAI, 2024), BLINK (Fu et al., 2024), HallusionBench (Guan et al., 2024), and MMVet-Hard (Yu et al., 2024). All are run with VLMEvalKit (Duan et al., 2024) using the same prompts and decoding settings for base and trained models. Absolute numbers can therefore differ from officially reported ones, and comparisons should be made within Table 6 in Section 5.2.

E Enlarged token-entropy visualization

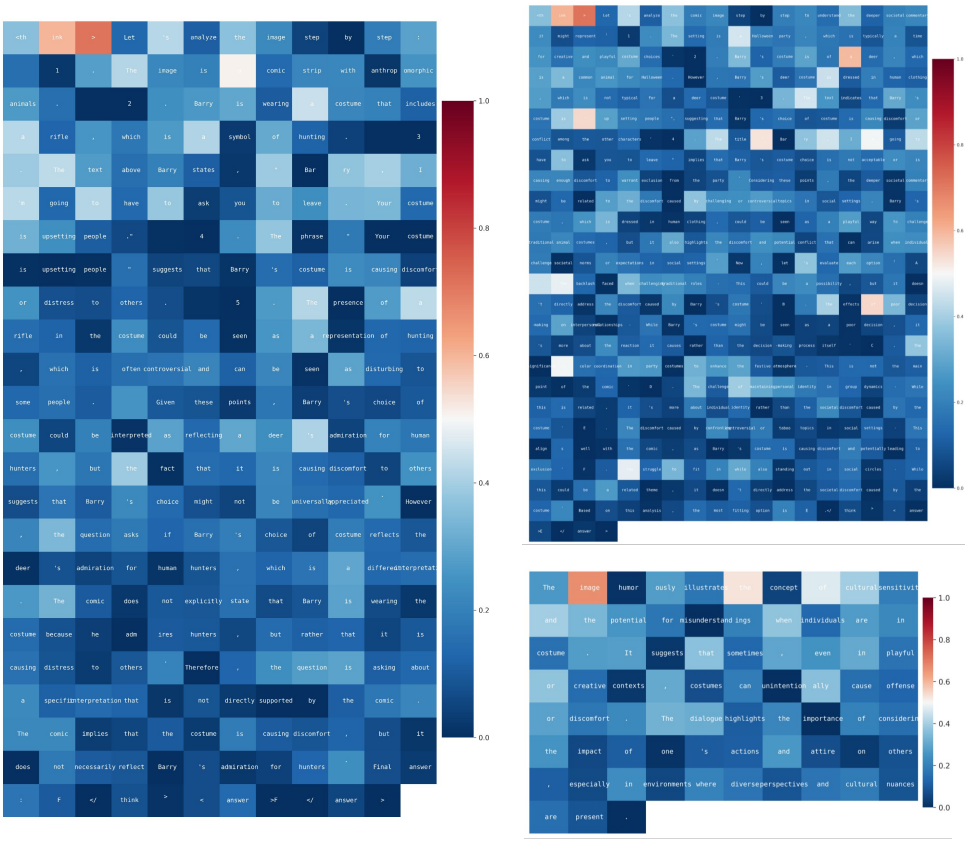


Figure 8: Enlarged version of Figure 3: token entropy of MetaphorStar-7B when answering a TFQ (left), an MCQ (top right), and an OSQ (bottom right). Red tokens have high entropy (high uncertainty) and blue tokens have low entropy. High-entropy tokens concentrate on logical connectives (e.g., "therefore", "but"), function words, quantifiers, and pronouns; tokens that restate visual facts or complete fixed phrases have low entropy.

F Complete proposition set of the Figure 1 item

The comic in Figure 1 appears in MetaphorBench as two II-Bench entries (E5 and E6) that share the same image. Table 16 lists all 18 propositions verbatim. True propositions describe the visual setup and the subverted fairy-tale expectation; false propositions state plausible but wrong readings, such as a transformation that does not happen or an emotion that the comic does not show.

Table 16: All MetaphorBench propositions for the comic in Figure 1.

Entry	Proposition	Label
E5	The comic strip uses the fairy tale motif of a princess kissing a frog, but subverts expectations by not having the frog transform into a prince.	T
E5	The final panel shows the frog turning into a prince, which confirms the fairy tale expectation of transformation through love.	F
E5	The presence of a heart above the princess and frog in the last panel suggests affection, despite the lack of transformation.	T
E5	The comic strip implies that true love always results in magical transformations, as depicted by the unchanged frog.	F
E5	The lack of transformation in the comic strip suggests a metaphor that reality may not always align with fairy-tale expectations.	T
E5	The princess appears disappointed in the last panel because the frog did not transform into a prince.	F
E5	The comic strip plays on the idea that love or affection does not necessarily lead to drastic changes in reality.	T
E5	The comic strip ends with the princess rejecting the frog, indicating a lack of affection or love.	F
E6	The comic subverts the classic fairy tale trope by showing no transformation of the frog after the kiss, suggesting a twist on traditional expectations.	T
E6	The princess appears disappointed after kissing the frog, as the heart floating above them implies dissatisfaction.	F
E6	The lack of transformation in the comic could metaphorically suggest that love does not always lead to dramatic changes or "fairy-tale" outcomes.	T
E6	The comic suggests that the princess was expecting the frog to remain unchanged, as her actions imply she anticipated a transformation.	F
E6	The heart floating above the princess and frog in the final panel implies that affection or love can exist without the need for transformation or idealized outcomes.	T
E6	The comic implies that the frog was never meant to transform into a prince, as it directly aligns with traditional fairy tale logic.	F
E6	The princess's crown and action of kissing the frog are clear references to the classic fairy tale of "The Frog Prince."	T
E6	The comic's message is that fairy tale logic always applies to real-life scenarios, as the lack of transformation contradicts this idea.	F
E6	The comic humorously critiques the unrealistic expectations set by fairy tales, using the unchanged frog as a metaphor for reality.	T
E6	The princess's affection for the frog is shown to be conditional on its transformation into a prince, as the heart suggests otherwise.	F

G Limitations and Future Work

MetaphorBench is built from II-Bench and mainly covers English-language images. Its 50 held-out items are complemented by exact paired tests (Appendix Table 11) and a 200-item evaluation of API baselines (Appendix Table 12). Future work includes extending the benchmark to other languages and cultures, larger held-out evaluations of all models, and reward controls that isolate the contribution of the structured template.

H Discussion

This section expands Section 6.5. The explanations below are hypotheses consistent with our results (Tables 8 and 6, Figure 6); isolating each factor requires additional controls.

H.1 Why does SFT warmup lose?

The ablation in Section 6.4 shows that, in our setting, a conventional SFT warmup stage is unnecessary and harmful for image implication. We see three contributing factors.

Task nature: creative generalization. Image implication is not simple pattern recognition; it requires connecting semantically distant concepts and producing low-probability insights. SFT, as a maximum-likelihood objective, trains the model to reproduce the high-probability sequences of the training data, acting as an entropy bottleneck (Figure 6: 1.33 for the base model, 0.30 after SFT). End-to-end RL is driven only by the reward and can explore and reinforce less likely reasoning paths as long as they lead to correct answers.

Question format: the talker versus the thinker. TFQ and MCQ are discriminative: the model must reject wrong statements or options. SFT trains the model to generate text that looks plausible and follows the `<think>`/`<answer>` structure, but not to perform the underlying discrimination. This is consistent with the MCQ accuracy of 28% for both SFT-based models, below the base model (46%), whereas end-to-end RL optimizes the accuracy reward R_{acc} directly and reaches 74%.

Evaluation: the OSQ paradox. The SFT→RL model obtains the highest OSQ score (3.66) although it is weakest on MCQ. In our inspection, SFT-trained models write verbose, structured answers that mix several, sometimes contradictory, viewpoints, one of which usually matches the reference, while the end-to-end RL model gives a shorter answer committed to one reading. The rubric's top levels reward "detailed" and "comprehensive" answers (Figure 7), and LLM judges are known to prefer longer responses (Zheng et al., 2023; Dubois et al., 2024). We did not measure response lengths, so this remains an explanation to be tested.

In summary, SFT warmup produces a low-entropy policy focused on imitation, and on-policy GRPO starting from this distribution likely samples near-identical rollouts that yield uninformative group-relative advantages; we infer this from the entropy curves and did not measure rollout diversity directly. End-to-end RL starts from the high-entropy base policy and can search more broadly.

H.2 Why can implication training help visual reasoning?

Table 6 shows that training on image implication is associated with higher reasoning scores and stable understanding scores, although MetaphorQA contains no math, charts, or documents. We see three candidate factors.

Multi-hop abstract reasoning. Implication tasks require moving from literal perception ("a person") to abstract conceptualization ("the person represents a concept") and to a conclusion ("the relationship between the concepts"). This non-literal, multi-step inference uses faculties that are also needed in logical, mathematical, and visual reasoning benchmarks.

TFQ as a reasoning trainer. Each image comes with about ten propositions to verify, which turns the model from an answer generator into a propositional verifier. Checking the truth of specific claims is a component of many reasoning tasks.

Grounding abstraction in perception. MetaphorQA includes propositions about basic visual facts alongside the central implication, so the model must connect concrete visual evidence to abstract conclusions instead of drifting into ungrounded interpretation. This is consistent with the stable scores on perception-oriented benchmarks such as POPE and SEEDBench.

Separating these factors from the general effect of RL with a structured template is left to future work.

Scope of the recipe. The construction requires only a human-written reference per item and a generator that turns it into true and plausible false statements; it needs neither a rule-checkable answer nor a judge during training. Tasks whose targets are interpretations, such as explaining humor, satire, or advertisements, meet these requirements; contested references must first be agreed on.