IMAGENET-THINK-250K: A LARGE-SCALE SYNTHETIC DATASET FOR MULTIMODAL REASONING FOR VISION LANGUAGE MODELS

Krishna Teja Chitty-Venkata *
Argonne National Laboratory
Lemont, IL, USA
schittyv@anl.gov

Murali Emani
Argonne National Laboratory
Lemont, IL, USA
memani@anl.gov

ABSTRACT

We develop ImageNet-Think-250K, a multimodal reasoning dataset designed to aid the development of Vision Language Models (VLMs) with explicit reasoning capabilities. Our dataset is built on 250,000 images from ImageNet-21k dataset, providing structured thinking tokens and corresponding answers. Our synthetic dataset is generated by two state-of-the-art VLMs: GLM-4.1V-9B-Thinking and Kimi-VL-A3B-Thinking-2506. Each image is accompanied by two pairs of thinking-answer sequences, creating a resource for training and evaluating multimodal reasoning models. We capture the step-by-step reasoning process of VLMs and the final descriptive answers. Our goal with this dataset is to enable the development of more robust VLMs while contributing to the broader understanding of multimodal reasoning mechanisms. The dataset and evaluation benchmarks will be publicly available to aid research in reasoning/thinking multimodal VLMs. The dataset is available here on HuggingFace.

1 INTRODUCTION

The advancement of Vision Language Models (VLMs) (Wu et al. (2024); Wang et al. (2024b); Lu et al. (2024); Liu et al. (2023); Chen et al. (2024)) has demonstrated remarkable capabilities in understanding and reasoning visual content (Radford et al. (2021); Li et al. (2022)). Recent developments in large language models have shown that incorporating explicit reasoning steps, such as chain-of-thought (CoT) prompting (Wei et al. (2022)), significantly improves performance on complex tasks. This success motivated the computer vision community to explore similar approaches for multimodal reasoning. The emergence of thinking models such as OpenAI's o1 series (Jaech et al. (2024)), DeepSeek-R1 (Guo et al. (2025)), GLM-4.1V-Thinking (Hong et al. (2025)), Kimi-VL-Thinking (Team et al. (2025)), and R1-Onevision (Yang et al. (2025b)) represents a paradigm shift toward more systematic reasoning in VLMs. These models explicitly generate intermediate reasoning steps before producing final answers, offering unprecedented insights into the cognitive processes underlying multimodal understanding. However, the datasets used to train these models are often proprietary or limited in scope, creating a significant barrier for broader research community.

Current multimodal datasets, while extensive in scale and diversity, primarily focus on input-output mappings without capturing the intermediate reasoning steps that lead to final answers (Lin et al. (2014); Goyal et al. (2017)). This limitation hinders the development of reasoning models and makes it challenging to diagnose model failures or understand decision-making processes. Furthermore, existing reasoning-focused datasets are often domain-specific or limited in scale, restricting their utility for training robust, general-purpose VLMs.

To address these challenges, we introduce **ImageNet-Think-250K**, a large-scale multimodal reasoning dataset that captures explicit thinking processes from state-of-the-art Vision Language Models. Our dataset consists of 250,000 images, sampled from ImageNet-21k dataset (Ridnik et al.

*Work done at Argonne National Laboratory. Now at Red Hat AI.

(2021)), each annotated with structured thinking tokens and corresponding answers. The reasoning and answer tokens are generated by GLM-4.1V-Thinking (Hong et al. (2025)) and Kimi-VL-Thinking (Team et al. (2025)) models. This multi-model dataset generation approach ensures diversity in reasoning patterns and provides comprehensive coverage of different analytical perspectives. ImageNet-Think enables researchers to train more interpretable and reliable VLMs while advancing our understanding of multimodal reasoning models.

Contributions: Our contributions are as follows:

- **Large-scale Reasoning Dataset**: ImageNet-Think provides 250,000 images with 500,000 thinking-answer pairs (two per image), representing one of the largest publicly available datasets with explicit reasoning annotations.
- **Multi-model reasoning diversity**: Our dataset captures diverse visual content and reasoning tokens by incorporating reasoning patterns from two models (GLM-4.1V-Thinking (Hong et al. (2025)) and Kimi-VL-Thinking (Team et al. (2025))), enabling more robust VLM evaluation.
- **Comprehensive evaluation benchmarks**: We benchmark several reasoning VLMs on ImageNet-Think and present results across multiple evaluation metrics. The models include InternVL3.5-8B (Wang et al. (2025c)), VL-Rethinker-7B (Wang et al. (2025a)), VisionThink-Efficient (Yang et al. (2025a)), OpenVLThinker-7B (Deng et al. (2025)), and R1-OneVision-7B (Yang et al. (2025b)).

2 Related Work

2.1 Multimodal Datasets

MME benchmark (Yin et al. (2024)) provides a comprehensive evaluation framework with 14 subtasks focusing on perception and cognition, such as object recognition and commonsense reasoning, utilizing manually constructed instruction-answer pairs to mitigate data leakage. TextVQA (Singh et al. (2019)) is designed for visual question answering that requires reading and reasoning about text in images, comprising 45,336 questions across 28,408 images from the OpenImages dataset. AI2D (Kembhavi et al. (2016)) consists of over 5,000 science diagrams from elementary school curricula, annotated with more than 15,000 elements and multiple-choice questions to test diagram interpretation skills. DocVQA (Mathew et al. (2021)) offers 50,000 questions on 12,767 document images, including forms, tables, and figures, framing document understanding as an extractive question-answering task emphasizing layout and content reasoning. MMMU (Yue et al. (2024)) is a massive multi-discipline benchmark with over 11,500 college-level problems spanning six disciplines, incorporating diverse image types like diagrams and charts to evaluate expert-level knowledge. InfoVQA (Mathew et al. (2022)) includes 30,035 questions on 5,485 infographics, demanding joint reasoning over complex layouts, textual elements, and data visualizations. RealWorldQA (Zhang et al. (2024)) features over 700 high-resolution real-world images, often captured from vehicles, with verifiable questions to assess practical applicability and challenge models on intricate scenarios. MMBench (Liu et al. (2024)) evaluates a broad range of abilities through 3,217 multiple-choice questions across 20 dimensions, including a dedicated test set to prevent overfitting. GQA (Hudson & Manning (2019)) focuses on compositional questions for structured visual reasoning on real-world images. POPE (Li et al. (2023)) probes object hallucination in vision-language models via evaluative tasks. ChartQA (Masry et al. (2022)) targets question answering on charts and plots to test data interpretation proficiency. TextCaps (Sidorov et al. (2020)) involves generating captions that incorporate scene text reading for image description tasks.

2.2 Multimodal Reasoning Datasets

The development of multimodal reasoning capabilities has been supported by various datasets, each contributing unique perspectives to the field. Early datasets such as VQA (Antol et al. (2015)) established foundational benchmarks for visual question answering and scene understanding. However, these datasets primarily focus on immediate responses without capturing intermediate reasoning steps. Recent efforts have introduced more sophisticated reasoning challenges. ScienceQA (Lu et al. (2022)) provides 21,208 multimodal multiple-choice questions from science topics, covering natural, language, and social sciences with integrated image and text contexts, along with detailed annotations including lectures and explanations. MATH-Vision (Wang et al. (2024a)) specifically targets mathematical reasoning with visual contexts, while MV-MATH (Wang et al. (2025b)) extends evaluation to multi-visual scenarios. These datasets represent significant progress toward more complex reasoning tasks but remain limited in scale and domain coverage. The Visual CoT dataset (Shao et al. (2024)) introduced 438,000 question-answer pairs with intermediate bounding boxes and reasoning steps, marking a significant advancement in capturing visual reasoning processes.

2.3 Chain-of-Thought and Thinking Models

Chain-of-thought (CoT) reasoning refers to the process by which a model explicitly generates intermediate reasoning steps before producing the final answer. Instead of mapping inputs directly to outputs, the model reveals its step-by-step logical progression, often resembling human problem-solving through scratchpad calculations or explanations. This approach has been shown to improve performance on tasks involving arithmetic, commonsense reasoning, and multi-step logical inference, as the intermediate reasoning helps the model decompose complex problems into smaller, more manageable parts. The success of chain-of-thought prompting in Large Language Models (Wei et al. (2022)) has inspired similar approaches in multimodal contexts. Multimodal reasoning capability refers to the ability of a model to jointly process and integrate information from multiple modalities, such as text, images, audio, and video, in order to perform coherent reasoning and generate contextually grounded outputs. Recent work has demonstrated that structured reasoning significantly improves performance on complex visual tasks (Xu et al. (2024)). LLaVA-CoT (Xu et al. (2024)) introduced autonomous multistage reasoning for vision-language models, showing marked improvements on reasoning-intensive benchmarks. The emergence of thinking models represents a paradigm shift toward more transparent AI systems. OpenAI's o1 series demonstrated the potential of extended reasoning phases, while GLM-4.1V-Thinking (Hong et al. (2025)) and Kimi-VL-Thinking (Team et al. (2025)) brought similar capabilities to open-source multimodal models. R1-Onevision (Yang et al. (2025b)) further advanced this direction by incorporating formal language-driven reasoning approaches.

2.4 Vision-Language Model Evaluation

Evaluating multimodal reasoning capabilities requires sophisticated benchmarks that go beyond simple accuracy metrics. Recent work has introduced various evaluation frameworks, including MMMU (Yue et al. (2024)) for massive multimodal understanding, MathVista (Lu et al. (2023)) for mathematical visual reasoning, and MM-Vet (Yu et al. (2023)) for comprehensive multimodal evaluation. However, existing evaluation approaches often fail to assess the quality of intermediate reasoning steps, focusing primarily on final answer accuracy. This limitation makes it difficult to understand model capabilities and failure modes, highlighting the need for datasets that explicitly capture reasoning processes.

2.5 Limitations of Existing Approaches

The existing multimodal reasoning datasets face several limitations that ImageNet-Think addresses:

- **Scale limitations**: Most reasoning datasets that are currently available contain fewer than 100,000 samples in total, which significantly limits their utility and effectiveness for training large-scale models that typically require substantially larger amounts of data to achieve optimal performance.
- **Domain specificity**: Many existing datasets tend to focus narrowly on specific domains, such as science or mathematics, which inherently limits their generalizability and applicability to broader reasoning tasks across different fields and contexts.
- **Lack of reasoning transparency**: Only a relatively small number of datasets successfully capture and document explicit reasoning steps within their samples, which makes it considerably more difficult for researchers to develop and improve better CoT models.
- **Single-model perspectives**: Most datasets currently in use rely on single annotation sources or individual annotators for their labeling and verification, which limits the diversity in reasoning approaches and perspectives that are represented within the data.
- **Evaluation gaps**: Existing benchmarks and evaluation frameworks often lack comprehensive and thorough evaluation metrics for assessing reasoning quality, making it challenging to accurately measure and compare the true reasoning capabilities of different models.

Dataset	#Images	Minimum Tokens	Maximum Tokens	Average Tokens	Total Tokens	**Reasoning**	**Models**	**Domains**
Visual CoT Shao et al. (2024)	438K	218	4871	1373	733k	Partial	Single	Localization
LLaVA-CoT-100K Xu et al. (2024)	98.5K	19	855	261	25.7M	Yes	Single	General
Emma Hao et al. (2025)	1.8k	10	4050	209	397k	Yes	Single	Coding, Chemistry, Physics
MAC Jiang et al. (2025)	9k	23	2086	239	2.1M	Yes	Single	General
ImageNet-Think	**250K**	521	196388	1542	0.3B	**Yes**	**Multi**	**General**

Table 1: Comparison of ImageNet-Think with existing multimodal reasoning datasets. The table reports several key statistics: **#Images** denotes the number of images in each dataset. **Minimum Tokens** and **Maximum Tokens** correspond to the smallest and largest number of tokens per sample across all data points, respectively. **Average Tokens** reflects the mean number of tokens per sample, while **Total Tokens** represents the cumulative token count for the entire dataset. Our ImageNet-Think dataset provides the largest scale with comprehensive reasoning annotations from multiple models, making it broadly applicable across general domains.

ImageNet-Think addresses these limitations by providing a large-scale, domain-diverse dataset with explicit reasoning annotations from multiple state-of-the-art models, enabling more comprehensive evaluation of multimodal reasoning capabilities.

3 IMAGENET-THINK-250K DATASET

3.1 IMAGE SOURCE

We construct our ImageNet-Think dataset by sampling 250k samples across several classes from ImageNet-21k (Ridnik et al. (2021)) dataset, a comprehensive image classification dataset containing over 14 million images across 21,841 categories. ImageNet-21k was selected as our base dataset for several reasons: (1) its extensive diversity across object categories and visual concepts, (2) its widespread adoption in the computer vision community, ensuring familiarity and comparability, (3) the high quality of images and annotations, and (4) its comprehensive coverage of real-world visual scenarios. From the complete ImageNet-21k dataset, we sample 250,000 images across all classes, ensuring balanced representation across different categories while maintaining the original distribution characteristics. Our sampling strategy considered category frequency, visual complexity, and semantic diversity to create a representative subset suitable for multimodal reasoning tasks. The dataset consists of a total of 250K images distributed across 10,450 classes. Each class contains between 22 and 31 samples, with an average of 23.9 samples per class, which is summarized in Table 2. The inclusion criteria for image selection were: (1) image resolution of at least 224×224 pixels to ensure sufficient visual detail, (2) clear visual content without significant occlusion or corruption, (3) diverse semantic categories to promote generalization, and (4) balanced representation across different visual complexity levels. Images with excessive noise, copyright restrictions, or potentially harmful content were excluded from the dataset.

Table 2: Summary of Images

Total Images	Total Classes	Min. Samples Per Class	Max. Samples Per Class	Avg. Samples Per Class
250K	10450	22	31	23.9

3.2 COLLECTION PROTOCOL

We annotate the selected images using two state-of-the-art VLMs with explicit reasoning capabilities: GLM-4.1V-Thinking (Hong et al. (2025)) and Kimi-VL-Thinking (Team et al. (2025)). These models were selected based on their demonstrated performance on multimodal reasoning benchmarks and their ability to generate structured thinking tokens. Our collection protocol followed a systematic approach:

1. **Image Preprocessing**: Each selected image was standardized to common formats and resolutions suitable for both the models.
2. **Prompt design**: We developed a unified prompt template that elicits comprehensive reasoning from each model: ”Please analyze this image step by step. First, describe what you

observe, then explain your reasoning process, and finally provide your conclusion about the main content or concept depicted."

3. **Model inference**: Each image and prompt pair is processed by both models independently, generating thinking tokens and final answers.
4. **Output formatting**: The generated responses were parsed to separate thinking tokens from final answers, ensuring consistent structure across all models.
5. **Quality control**: A subset of outputs underwent manual review to verify parsing accuracy and content quality.

The collection process was distributed across multiple computing nodes to ensure efficiency and scalability. Each model processed the entire image set, resulting in three distinct reasoning perspectives for each image. We consumed over 6000 A100 GPU hours to curate the dataset.

3.3 Annotation Process

The annotation process centered on capturing the explicit reasoning processes of two advanced Vision Language Models. Unlike traditional annotation approaches that rely on human annotators, our methodology leverages the sophisticated reasoning capabilities of SOTA VLMs to generate high-quality thinking tokens and answers. This comprehensive annotation process ensures that ImageNet-Think captures diverse reasoning patterns while maintaining high quality and consistency across all samples.

Figures 1 and 2 illustrate the structure and content of our ImageNet-Think dataset. Figure 1 shows how each sample begins with an input question ("Please analyze this image step by step..."), followed by intermediate *thinking tokens* that capture reasoning steps (e.g., identifying objects, context, or relationships), which are then paired with corresponding *answer tokens* that refine or finalize the interpretation of the image. Two representative examples are shown: (a) a traditional stone-milling setup and (b) a large reptile (turtle/tortoise). Figure 2 provides a detailed instance where a subject engaged in archery is analyzed step by step, with multiple rounds of thinking tokens elaborating fine-grained elements (attire, equipment, setting) and final answers synthesizing these observations into a coherent explanation. Together, these figures highlight how our dataset explicitly separates intermediate reasoning from outcome answers, enabling evaluation of both reasoning quality and answer correctness.

3.4 Model-Specific Protocols:

Kimi-VL-A3B-Thinking-2506: This model employs a Mixture-of-Experts architecture with 2.8B activated parameters and generates structured thinking tokens with clear delineation between reasoning steps and final conclusions. The model's thinking process typically follows a pattern of observation, analysis, and synthesis.

GLM-4.1V-9B-Thinking: Utilizing Reinforcement Learning with Curriculum Sampling (RLCS), this model generates comprehensive reasoning chains that demonstrate deep understanding of visual content and contextual relationships.

4 Dataset Characteristics

4.1 Statistics

As the name suggests, ImageNet-Think-250K comprises 250,000 images with comprehensive multimodal reasoning annotations, resulting in 500,000 thinking-answer pairs across two VLMs. The dataset exhibits substantial scale and diversity, making it suitable for training/finetuning/evaluating robust reasoning VLMs.

4.2 Scale and Coverage:

Our dataset spans 10,450 unique ImageNet-21k classes, providing extensive coverage of visual concepts from everyday objects to specialized scientific subjects. The average number of images per

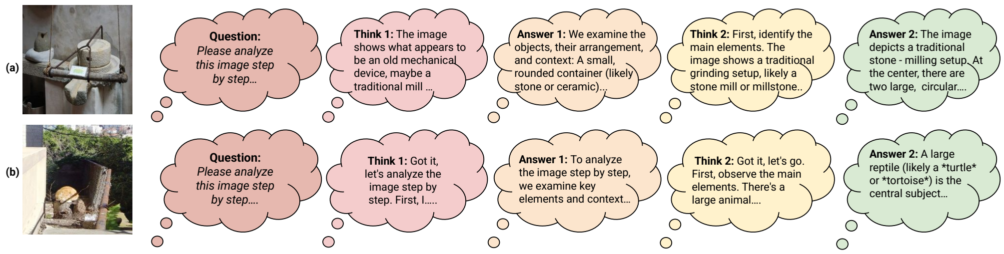

Figure 1: Organization of our ImageNet-Think-250K dataset. Each sample consists of an input **question** (“Please analyze this image step by step...”), followed by multiple rounds of **thinking tokens** (*Think 1, Think 2, ...*), where the model produces intermediate reasoning steps describing objects, context, and relationships. These are then paired with corresponding **answer tokens** (*Answer 1, Answer 2, ...*), which provide refined explanations or final interpretations of the image. The figure illustrates two examples: (a) a traditional stone-milling setup, and (b) a large reptile (turtle/tortoise). This organization highlights how our dataset explicitly separates reasoning traces from final answers, enabling the evaluation of both reasoning quality and outcome accuracy.

category is 16.4, with a standard deviation of 12.8, reflecting the natural distribution of ImageNet-21k while maintaining balanced representation across different domains.

4.3 Language Characteristics:

All thinking tokens and answers are generated in English, with vocabulary analysis showing an average of 1,247 unique words per 1,000 reasoning samples. The language complexity spans from simple descriptive statements to sophisticated analytical discourse, reflecting the diverse reasoning requirements across different visual contexts.

4.4 Comparison with Existing Datasets

Table 1 compares ImageNet-Think with existing multimodal reasoning datasets. While prior benchmarks such as Visual CoT (438K images, 733k tokens), LLaVA-CoT-100K (98.5K images, 25.7M tokens), Emma (1.8K images, 397k tokens), and MAC (9K images, 2.1M tokens) provide valuable reasoning traces, they are either limited in scale, token coverage, or domain diversity. In contrast, ImageNet-Think offers the largest dataset with 250K images, over 0.3B annotated tokens, and reasoning sequences reaching up to 196K tokens per sample. Moreover, unlike single-model annotations in previous datasets, ImageNet-Think integrates reasoning from multiple models, ensuring broader generalization and comprehensive multimodal reasoning coverage. Our dataset significantly exceeds previous efforts in terms of scale, reasoning annotation coverage, and model diversity.

Scale Advantages: ImageNet-Think provides 2.5× more reasoning-annotated images than the largest existing dataset (Visual CoT) and 25× more than specialized reasoning datasets like ScienceQA. This scale advantage enables training of more robust models and supports comprehensive evaluation across diverse scenarios.

Reasoning Depth: Unlike datasets that provide limited reasoning annotations (e.g., ScienceQA with brief explanations), ImageNet-Think captures complete thinking processes from initial observation to final conclusion. The average reasoning chain length of 1.5k tokens significantly exceeds typical explanation lengths in existing datasets.

Multi-Model Perspective: ImageNet-Think is the first large-scale dataset to capture reasoning patterns from multiple state-of-the-art VLMs, providing 2× the reasoning diversity of single-model datasets. This multi-perspective approach enables analysis of reasoning consistency and development of more robust evaluation metrics.

Domain Generality: While specialized datasets like MATH-Vision focus on specific domains, ImageNet-Think provides broad coverage across visual concepts, making it suitable for general-purpose reasoning model evaluation.

4.5 Bias and Limitations

Model-Specific Biases: Our dataset inherits certain biases from the source models used for annotation generation. Each model brings distinct architectural biases and training data influences that may affect reasoning and final answer patterns. Kimi-VL-A3B-Thinking-2506 may exhibit biases toward systematic, step-by-step analysis due to its MoE architecture and training methodology. GLM-4.1V-9B-Thinking might demonstrate biases toward comprehensive analysis influenced by its RLCS training approach.

Representational Limitations: Despite our best efforts to maintain balanced sampling, certain limitations persist as we rely on ImageNet-21K images. ImageNet-21k's original geographic distribution biases may influence visual content representation. The reasoning patterns reflect primarily English-language analytical frameworks and may not capture diverse cultural reasoning approaches.

5 Benchmarking ImageNet-Think

In this section, we conduct a comprehensive evaluation of several state-of-the-art VLMs on our proposed dataset. Our analysis covers a wide range of quantitative metrics to assess model performance, and an ablation study on thinking accuracy and answer accuracy.

5.1 Task Definitions

We evaluate multiple reasoning models by assessing different aspects of multimodal understanding and analytical capabilities. We evaluate a model's ability to generate coherent, step-by-step reasoning processes when analyzing visual content. Given an input image, models must produce structured thinking tokens that demonstrate systematic analysis from initial observation to logical conclusion. The success is measured by reasoning coherence, logical progression, and alignment with human-like analytical patterns. Models must not only answer questions about visual content but also justify their responses through clear analytical processes. This task assesses both accuracy and reasoning quality, promoting more interpretability.

5.2 Baseline Models

We evaluate a set of state-of-the-art open-source reasoning VLMs, namely InternVL3.5-8B (Wang et al. (2025c)), VL-Rethinker-7B (Wang et al. (2025a)), VisionThink-Efficient (Yang et al. (2025a)), OpenVLThinker-7B (Deng et al. (2025)), and R1-OneVision-7B (Yang et al. (2025b)). Each model is assessed using a custom prompt (identical to the one used to generate the dataset with minor model-specific instructions) that jointly measures reasoning quality and final answer accuracy across varying levels of visual complexity. To ensure reproducibility and transparency, we restrict our study to open-source models and deliberately exclude proprietary systems such as GPT-4V (OpenAI (2023)) and Gemini 2.5 Pro (Google DeepMind (2025)).

5.3 Evaluation Metrics and Protocol

To comprehensively evaluate the quality of generated text, we employ a diverse suite of metrics designed to capture complementary aspects of text similarity and generation quality. Our evaluation framework is organized into four families of metrics, each targeting distinct linguistic properties such as lexical overlap, semantic similarity, and structural alignment.

For a given input, let s_1 and s_2 denote the similarity scores of a model obtained by comparing the generated text against the reference output in our dataset. We aggregate these pairwise scores using two complementary strategies. First, we compute the average score:

$$m_{\text{avg}} = \frac{s_1 + s_2}{2} \tag{1}$$

which reflects the mean similarity across both references. Second, we compute the maximum score:

$$m_{\text{max}} = \max(s_1, s_2) \tag{2}$$

which captures the best-case alignment with either reference.

We aggregate these quantities across all M examples in the evaluation set. Specifically, we define:

$$\text{Avg of Avg} = \frac{1}{M}\sum_{m=1}^{M} m_{\text{avg}}^{(m)}, \quad \text{Avg of Max} = \frac{1}{M}\sum_{m=1}^{M} m_{\text{max}}^{(m)} \tag{3}$$

where $m_{\text{avg}}^{(m)}$ and $m_{\text{max}}^{(m)}$ denote the average and maximum scores for the m-th example, respectively. The *Avg of Max* statistic reflects peak performance by averaging the best score obtained across paired comparisons, while *Avg of Avg* emphasizes stability and consistency by averaging mean scores. This dual perspective provides a more nuanced understanding of model behavior, balancing both occasional high-quality generations and overall robustness.

All metric values are normalized to the range $[0, 1]$, where higher values indicate stronger similarity between model-generated outputs and reference texts. This evaluation paradigm aligns with the concept of *LLM-as-Judge* (Gu et al., 2024), as it involves assessing model outputs by directly comparing them against the outputs of other models.

5.3.1 Semantic Similarity Metrics

We evaluate semantic similarity using embedding-based metrics that capture meaning beyond surface-level word overlap. **BERTScore** (Zhang et al., 2019) computes similarity using contextual embeddings. We report both the F1 score and the overall BERTScore, which provide robust measures of semantic similarity that correlate well with human judgments. **Sentence-BERT Cosine Similarity** (Reimers & Gurevych, 2019) measures semantic similarity in a unified sentence embedding space, enabling direct comparison of sentence-level representations.

Table 3 reports semantic similarity results across five VLMs using BERTScore and Sentence-BERT Cosine Similarity. Overall, all models achieve high BERTScore values (0.858-0.872), with VisionThink-Efficient and VL-Rethinker-7B attaining the highest scores (0.867/0.872), indicating strong contextual alignment. In contrast, Sentence-BERT similarity reveals greater differentiation: OpenVLThinker-7B leads (0.807/0.832), followed by InternVL-3.5-8B and R1-Onevision-7B (above 0.78/0.81), while VisionThink-Efficient and VL-Rethinker-7B underperform (0.726-0.727/0.750-0.752). These results highlight a trade-off, where some models excel in token-level contextual similarity (BERTScore) but lag in sentence-level semantic alignment, whereas OpenVLThinker-7B demonstrates a balanced performance across both metrics, suggesting stronger overall semantic preservation.

Table 3: Semantic Similarity Metrics

Model	BERTScore		Sentence-BERT Cosine Similarity	
	Avg.	Max	Avg.	Max
InternVL-3.5-8B	0.860	0.865	0.784	0.809
R1-Onevision-7B	0.858	0.863	0.789	0.811
OpenVLThinker-7B	0.864	0.869	0.807	0.832
VisionThink-Efficient	0.867	0.872	0.727	0.752
VL-Rethinker-7B	0.867	0.872	0.726	0.750

5.3.2 N-gram Overlap Metrics

To assess lexical overlap patterns, we employ ROUGE metrics (Lin, 2004), which are standard in text generation evaluation. **ROUGE-1** measures unigram overlap between generated and reference texts, capturing vocabulary similarity. **ROUGE-L** computes the longest common subsequence, accounting for word order while allowing for flexible matching. Table 4 presents lexical overlap results using ROUGE-1 and ROUGE-L, which measure unigram overlap and longest common subsequence alignment, respectively. InternVL-3.5-8B records the lowest ROUGE scores (0.428/0.477 for ROUGE-1 and 0.205/0.224 for ROUGE-L), suggesting weaker vocabulary and sequence-level overlap. In contrast, VisionThink-Efficient and OpenVLThinker-7B achieve the strongest performance, both reaching 0.529/0.559 on ROUGE-1, with OpenVLThinker-7B slightly leading on

ROUGE-L (0.262/0.274). VL-Rethinker-7B and R1-Onevision-7B fall in between, with consistent improvements over InternVL-3.5-8B but trailing the top performers. These results show that while all models capture lexical overlap to some degree, VisionThink-Efficient and OpenVLThinker-7B stand out for their stronger alignment in both unigram and subsequence matching.

Table 4: N-gram Overlap Metrics

Model	ROUGE-1		ROUGE-L	
	Avg.	Max	Avg.	Max
InternVL-3.5-8B	0.428	0.477	0.205	0.224
R1-Onevision-7B	0.476	0.512	0.232	0.247
OpenVLThinker-7B	0.527	0.559	0.262	0.274
VisionThink-Efficient	0.529	0.559	0.253	0.266
VL-Rethinker-7B	0.518	0.547	0.247	0.261

5.3.3 Lexical Set-Overlap Metrics

We evaluate token-level overlap using set-based metrics that treat texts as collections of unique tokens. **Jaccard Index** measures the ratio of intersection to union of token sets, providing a normalized overlap score robust to text length variations. **Overlap Coefficient** computes the ratio of common tokens to the smaller set size, which is particularly useful when comparing texts of different lengths.

Table 5 reports results for lexical set-overlap metrics using the Jaccard Index and Overlap Coefficient, which capture token-level overlap while accounting for text length variations. InternVL-3.5-8B records the lowest scores across both metrics (0.271/0.287 Jaccard, 0.534/0.586 Overlap), indicating weaker token-level alignment. In contrast, VisionThink-Efficient achieves the highest performance (0.308/0.324 Jaccard, 0.557/0.608 Overlap), followed closely by VL-Rethinker-7B (0.301/0.317 Jaccard, 0.542/0.589 Overlap). R1-Onevision-7B and OpenVLThinker-7B yield intermediate results, showing improvements over InternVL-3.5-8B but not matching the top performers. Overall, VisionThink-Efficient demonstrates the strongest ability to capture token-level overlap, while InternVL-3.5-8B lags behind across both set-based metrics.

Table 5: Lexical Set-Overlap Metrics

Model	Jaccard Index		Overlap Coefficient	
	Avg.	Max	Avg.	Max
InternVL-3.5-8B	0.271	0.287	0.534	0.586
R1-Onevision-7B	0.288	0.305	0.540	0.580
OpenVLThinker-7B	0.285	0.300	0.525	0.568
VisionThink-Efficient	0.308	0.324	0.557	0.608
VL-Rethinker-7B	0.301	0.317	0.542	0.589

5.3.4 Vector-Space Metrics

Finally, we assess similarity in high-dimensional lexical feature spaces. **TF-IDF Cosine Similarity** measures similarity using term frequency-inverse document frequency weighting, emphasizing distinctive vocabulary while down-weighting common terms. **Cosine Similarity** provides a baseline vector-space comparison using raw term frequencies. Table 6 illustrates the results for TF-IDF Cosine Similarity and raw Cosine Similarity. InternVL-3.5-8B shows the weakest performance overall (0.754/0.778 TF-IDF, 0.382/0.403 Cosine), while VisionThink-Efficient achieves the strongest results across both metrics (0.805/0.832 TF-IDF, 0.430/0.453 Cosine), highlighting its ability to capture distinctive lexical patterns and maintain robust vector-space alignment. OpenVLThinker-7B and VL-Rethinker-7B also perform competitively, with scores around 0.795–0.820 (TF-IDF) and 0.399–0.439 (Cosine). R1-Onevision-7B yields mid-tier results, outperforming InternVL-3.5-8B but lagging behind the top models. Overall, VisionThink-Efficient emerges as the most effective model in high-dimensional lexical similarity, while InternVL-3.5-8B consistently underperforms.

Table 6: Vector Space Metrics

Model	TF-IDF Cosine Similarity		Cosine Similarity	
	Avg.	Max	Avg.	Max
InternVL-3.5-8B	0.754	0.778	0.382	0.403
R1-Onevision-7B	0.765	0.789	0.397	0.418
OpenVLThinker-7B	0.795	0.820	0.399	0.421
VisionThink-Efficient	0.805	0.832	0.430	0.453
VL-Rethinker-7B	0.792	0.819	0.418	0.439

5.4 Ablation Studies: Thinking vs Answer tokens.

Table 7 compares the performance of OpenVLThinker across the full output of the model, thinking tokens, and answer tokens using multiple lexical and semantic similarity metrics. Overall, *thinking tokens* consistently yield the highest semantic similarity scores, with the best BERTScore (0.872 Avg., 0.877 Max.) and Sentence-BERT cosine similarity (0.810 Avg., 0.834 Max.). In contrast, *answer tokens* achieve relatively lower values across most metrics, particularly in ROUGE-1 (0.433 Avg.) and TF-IDF (0.671 Avg.), though they perform competitively in overlap coefficient (0.585 Avg., 0.647 Max.). The *full output* generally provides balanced performance between the two extremes, showing stable lexical scores (e.g., Jaccard 0.285 Avg., 0.300 Max.) and strong TF-IDF cosine similarity (0.399 Avg., 0.421 Max.). These results highlight that reasoning-oriented thinking tokens capture richer semantic alignment, whereas concise answer tokens excel more in exact lexical overlap.

Table 7: Thinking Tokens vs Answer Tokens of OpenVLThinker Model

Metric	Full Output		Thinking Tokens		Answer Tokens	
	Avg.	Max.	Avg.	Max.	Avg.	Max.
BERTScore	0.864	0.869	0.872	0.877	0.847	0.860
Sentence-BERT Cosine Similarity	0.807	0.832	0.810	0.834	0.772	0.808
ROUGE-1	0.527	0.559	0.497	0.543	0.433	0.476
ROUGE-L	0.262	0.274	0.260	0.277	0.248	0.277
Jaccard Index	0.285	0.300	0.283	0.304	0.227	0.253
Overlap Coefficient	0.525	0.568	0.511	0.542	0.585	0.647
TF-IDF	0.795	0.820	0.763	0.798	0.671	0.699
TF-IDF Cosine Similarity	0.399	0.421	0.395	0.418	0.340	0.367

6 Conclusion and Future Work

In this paper, we present the ImageNet-Think-250K dataset, a large-scale reasoning tokens from state-of-the-art Vision Language Models. Our dataset offers explicit multimodal reasoning processes to enable the development of more interpretable and robust reasoning VLMs. The main contributions of this work include the creation of the largest publicly available dataset with explicit multimodal reasoning annotations, comprising 250,000 images with 500,000 thinking-answer pairs and benchmarking them across five open source reasoning VLMs.

Future Work: In future, we plan to expand the dataset to include additional image modalities (video, audio) and incorporate more recent thinking models as they become available. We also plan to include multi-lingual reasoning tokens and domain-specific applications. We also plan to develop more sophisticated metrics for assessing reasoning quality and output accuracy. Leveraging ImageNet-Think to develop novel architectures specifically designed for explicit reasoning generation, including hierarchical reasoning models, multi-step verification systems, and interactive reasoning frameworks.

ACKNOWLEDGEMENTS

This research used resources of the Argonne Leadership Computing Facility, which is a U.S. Department of Energy Office of Science User Facility operated under contract DE-AC02-06CH11357.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement, 2025. URL https://arxiv.org/abs/2503.17352.

Google DeepMind. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, March 2025. Blog post / announcement.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

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

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pp. arXiv–2507, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Mohan Jiang, Jin Gao, Jiahao Zhan, and Dequan Wang. Mac: A live benchmark for multimodal large language models in scientific understanding. *arXiv preprint arXiv:2508.15802*, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521, 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. *arXiv preprint arXiv:2310.02255*, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

OpenAI. Gpt-4 with vision (gpt-4v) system card. System card, OpenAI, September 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf. OpenAI system card.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pp. 742–758. Springer, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 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, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024a.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19541–19551, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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, 2024. *URL https://arxiv. org/abs/2412.10302*, 2024.

Guowei Xu, Peng Jin, Ziang Wu, 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.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Visionthink: Smart and efficient vision language model via reinforcement learning. *arXiv preprint arXiv:2507.13348*, 2025a.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

A ImageNet-Think-250K Dataset Example

Figure 2 illustrates a data point from our ImageNet-Think-250K dataset. Each instance in the dataset consists of three components: a Question that prompts detailed image analysis, a Think section that captures the intermediate reasoning process (showing how the model breaks down visual elements, context, and details), and multiple Answer fields that provide the final summarized conclusions. Using this archery scene as an example, the dataset demonstrates how models can be taught to explicitly articulate their reasoning about visual details, such as identifying the person's attire, the equipment being used, and the environmental context—before arriving at coherent answers, enabling researchers to study and improve the interpretability and reasoning quality of VLMS.

Question: Analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible

Think 1: *Got it, let's analyze the image step by step. First, identify the main subject: a person practicing archery. Let's look at the person's attire: wearing a red cap, white shirt, and a black glove on the left hand (since they're holding the bow). The bow is a traditional wooden longbow, with a string and a sight maybe? Wait, the bow has a wooden structure, typical of recurve or longbow. Then the targets: there are three circular targets with orange and blue colors, mounted on a white backdrop, probably a net or a wall. The setting is an outdoor area with dry grass, trees in the background, and a fence or structure with wires. Let's check the person's posture: drawing the bow back, aiming at the targets, which suggests they're in the process of shooting an arrow. The environment looks like an archery range, maybe a practice area with multiple targets for different distances. Also, the person's stance is focused, indicating concentration on the target. Let's list details: clothing (red cap, white t-shirt), equipment (wooden bow, black glove), targets (three circular with concentric circles, orange center), setting (outdoor, grassy area, trees, fence). Now, explain the reasoning: the person is engaged in archery, which is a sport or recreational activity involving shooting arrows with a bow. The setup of multiple targets suggests a practice session, maybe at a club or a dedicated archery range. The presence of the glove is important for grip and protection. The bow's design (wooden, traditional) might indicate traditional archery rather than modern compound bows. The targets are standard archery targets, often used for scoring based on hitting the center. The environment is open, with natural elements like trees, so it's likely a dedicated outdoor range.*

Answer 1: To analyze the image step by step, we break down elements, context, and details: ### 1. **Main Subject & Activity** The central figure is a person practicing **archery** a sport/recreational activity involving shooting arrows with a bow. The individual is in the process of aiming, with their posture (drawing the bow, focused gaze) indicating they are about to release an arrow. ### 2. **Person's Appearance & Attire** - **Headwear**: Wearing a bright red baseball cap, likely for sun protection. - **Clothing**: A white short-sleeve t-shirt, suitable for physical activity. - **Hand Protection**: A black glove on the left hand (holding the bow), which is common in archery for grip, stability, and protection. ### 3. **Archery Equipment** - **Bow**: A traditional wooden longbow (or recurve bow, given its shape). It has a wooden structure with a string, and the person is drawing the string back to aim. - **Arrows** (implied, though not visible): Archers typically use arrows with tips for hitting targets; the bow's position suggests an arrow is loaded or being drawn. ### 4. **Targets & Setting** - **Targets**: Three circular archery targets are mounted on a white backdrop (likely a net or wall). Each target has concentric circles (orange center, blue outer rings), which are standard for scoring (closer to the center = higher score). - **Environment**: The setting is an **outdoor archery range** dry grass, trees in the background, and a fence/wire structure. This is a dedicated space for archery practice, with open space for safety and multiple targets for varied distances. ### 5. **Context & Purpose** The scene suggests a **practice session** (not a competitive event, as there's no audience or formal setup). The multiple targets imply the archer is working on accuracy at different distances. The natural, open environment is typical of community archery clubs or recreational archery ranges. ### 6. **Additional Observations** - The archer's focused stance (body aligned with the target, eyes on the bullseye) highlights the skill and concentration required in archery. - The wooden bow indicates traditional archery (as opposed to modern compound bows, which use mechanical aids). In summary, the image shows a person practicing archery outdoors at a dedicated range, using traditional equipment and standard targets, with a focus on skill development.

Think 2: *So, let's analyze the image step by step. First, look at the main subject: a person wearing a red cap, white sleeveless shirt, and dark wristband. They're holding a traditional bow, likely a longbow, and aiming at an archery target. The background has multiple archery targets with concentric circles (standard for archery), set against a grassy field with trees. The sky is partly cloudy. The person's posture is focused, indicating they're engaged in archery practice or competition. The setting is an outdoor archery range, structured with targets and open space. Now, summarize all observed elements: subject's attire, equipment, target details, environment. First, describe the central figure: A person wearing a bright red baseball cap, a white sleeveless or short-sleeved shirt, and a dark wristband on their right arm. They are holding a longbow with both hands, positioned in a shooting stance, aiming at an archery target. Then, the background: Multiple archery targets are visible, each with concentric colored rings (typical of Olympic or standard archery targets). The targets are set against a backdrop of trees and a grassy field. The sky appears to be partly cloudy, suggesting an outdoor setting. Additional details: The person's posture is focused, indicating concentration on aiming. The longbow is wooden, and the person's stance is typical of traditional archery. The environment is an open, grassy area, likely a designated archery range, with natural vegetation (trees) in the distance. Now, compile this into a coherent description, ensuring all key elements are covered: the individual, their attire and equipment, the archery targets, and the outdoor setting.*

Answer 2: To analyze the image, we first identify key elements: 1. **Central Figure**: A person wearing a bright red baseball cap, a white sleeveless shirt, and a dark wristband on their right arm. They hold a traditional longbow with both hands, adopting a focused shooting stance (arms extended, bowstring drawn, aiming posture). 2. **Archery Equipment/Targets**: Multiple archery targets with concentric colored rings (standard for competitive archery) are visible in the background. These targets are positioned on a grassy field, indicating an outdoor archery range. 3. **Environment**: The setting is an open, grassy area with trees in the distance and a partly cloudy sky, suggesting a rural or park - like location for archery practice or competition. ### Step - by - Step Reasoning: - **Subject's Appearance**: The person's attire (red cap, white shirt, wristband) and the longbow (wooden, traditional design) are noted. Their posture (body turned, arms positioned to aim) shows concentration on target acquisition. **Background Elements**: The archery targets' design (concentric rings) confirms the sports context. The grass and trees establish an outdoor, recreational or competitive space. - **Contextual Relationships**: The person's stance and equipment directly relate to the targets; their focused posture implies practice or participation in archery. In summary, the image depicts an individual engaged in archery on an outdoor range, with distinct visual details of their attire, equipment, the archery targets, and the surrounding environment.

Figure 2: An example of a sample image and reasoning instance from the ImageNet-Think dataset. Each instance consists of a **Question** that prompts the model to analyze the image step by step and provide detailed reasoning. The **Think** section illustrates the reasoning trace, where the model generates step-by-step observation and contextual details about the scene; these intermediate outputs are referred to as *thinking tokens*. Following this, the dataset consists of multiple **Answer** fields that represent the final summarized outputs. This structure captures both the process (*thinking tokens*) and the outcome (*final answers*), enabling explicit evaluation of reasoning quality in addition to correctness of the end prediction. The example shown demonstrates how models explain visual details (attire, equipment, target setting, and actions) before producing coherent, task-relevant answers. Such instances highlight the dataset's ability to disentangle step-by-step reasoning from final responses, making it valuable for studying reasoning quality, interpretability, and multimodal chain-of-thought behaviors.