# VCU-Bridge: Hierarchical Visual Connotation Understanding via Semantic Bridging

Supplementary Material

## A. More Related Works

While instruction data for MLLMs has evolved toward automated synthesis, existing pipelines rarely produce explicit, verifiable chains connecting perception to connotation. Most approaches prioritize data scale or quality over intermediate reasoning steps that bridge visual observations with abstract interpretations.

**Template-based Generation.** Pipelines such as LLaVA-Instruct-150K [23], InstructBLIP [12], and ShareGPT4V [10] leverage powerful language models to generate large-scale instruction-following data from image captions or dense descriptions. While these methods successfully scale data coverage, they primarily produce *flat* question-answer pairs that lack hierarchical structure. The generated data provides weak supervision for intermediate reasoning steps, leaving the inferential connections between visual evidence and abstract interpretations underspecified. Consequently, models trained on such data learn to map images directly to high-level conclusions without explicitly modeling the semantic bridge that justifies these interpretations.

**Search-based Generation.** Recent work has begun integrating tree search mechanisms to explore diverse reasoning paths. Socratic-MCTS [3] applies MCTS-style procedures at inference time to elicit subquestions and intermediate verification steps, improving interpretability without additional training. ReST-MCTS* [47] employs MCTS with process-level rewards to curate high-quality reasoning trajectories in language modeling. While these methods demonstrate the potential of search-driven data generation, they lack *structured hierarchical progression* from concrete perceptual evidence to abstract connotative meaning. Moreover, they typically do not enforce *inter-level validation* to ensure that each reasoning step logically supports the next, which is critical for learning coherent semantic bridges. In contrast, our MCTS-driven pipeline explicitly constructs complete hierarchical reasoning chains with bottom-up search and inter-level validation, ensuring that generated data teaches models to ground abstract meanings in concrete visual evidence through verifiable intermediate steps.

## B. HVCU-Bench Details

### B.1. Dataset Statistics

As shown in Table 5, HVCU-Bench comprises 1,050 hierarchical multiple-choice QA samples distributed across three task families (Implication Understanding, Aesthetic Appreciation, and Affective Reasoning) and fifteen fine-grained aspects. Each sample is organized as a three-level question–answer chain ($L_{perc}$, $L_{bridge}$, $L_{conn}$) defined on a single image, resulting in a total of 3,150 QA pairs. This design supports both level-wise diagnosis and full-chain evaluation of VCU-Bridge.

Table 5. Statistics of HVCU-Bench across three task families and fifteen fine-grained aspects.

| | Implication Understanding (400) | | | | | Aesthetic Appreciation (350) | | | | Affective Reasoning (300) | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Metaphor | Symbolism | Contrast | Exaggeration | Dislocation | Color | Composition | Font | Graphics | Joy | Affection | Wonder | Anger | Fear | Sadness | |
| #Samples | 319 | 21 | 22 | 22 | 16 | 37 | 122 | 97 | 94 | 25 | 83 | 47 | 31 | 81 | 33 | 1050 |

### B.2. Data Quality Assurance and Validation

To further ensure the quality and reliability of HVCU-Bench, we perform a multi-stage human auditing process on top of the model-driven generation framework described in Section 3.2.

**Validation Protocol.** After automatic format and consistency checks, every candidate hierarchical multiple-choice QA chain is manually inspected by human annotators with access to the original image and all three VCU-Bridge levels. Each question–answer pair is validated by at least five annotators, who independently verify that (i) the correct option at each level is unambiguously supported by the visual content, and (ii) the three levels together form a logically coherent reasoning chain from perception to connotation.

**Consensus Resolution.** We monitor inter-annotator agreement and treat any inconsistencies or potential misinterpretations as flags for further review. Disagreements are resolved through discussion within the annotation team until a single consen-

sus label and reasoning chain are reached; items that remain ambiguous or exhibit artifacts from the generation model are discarded rather than forced to consensus.

**Multi-Round Quality Control.** In total, we conduct three rounds of such quality control, including repeated passes of cross-checking and consolidation, before finalizing the benchmark. As a result, the released HVCU-Bench uses human-validated annotations rather than raw outputs from Gemini-2.5-Pro, substantially reducing the risk that systematic biases or errors from the generation model propagate into our evaluation data.

### B.3. Dataset Samples

To provide concrete illustrations of the hierarchical reasoning structure in HVCU-Bench, we present representative samples spanning all three task families and fifteen fine-grained aspects. Each sample demonstrates a three-level question–answer chain from perception to connotation, illustrating how the benchmark grounds abstract interpretations in concrete visual evidence through explicit semantic bridges.

**List of Samples**



## C. Implementation Details

**Data Generation.** We generate approximately 10k hierarchical instruction data using Gemini 2.5-Pro and our MCTS-driven pipeline (Section 3.3) from 1k images. For each image, MCTS explores a 3-level reasoning tree with exploration constant $c = 2.0$, expanding up to 5 candidate nodes per step with quality filtering (threshold 0.65) and diversity control (threshold 0.75). Tree capacity is limited to 8, 12, and 15 nodes per level. We extract the top-10 paths per image, yielding approximately 10k three-level hierarchical chains, totaling around 30k question-answer pairs. All data follows an open-ended QA format rather than multiple-choice, ensuring the model learns genuine hierarchical reasoning instead of selection patterns.

**Model Training.** Our Qwen3-VL-4B-Bridge is instruction-tuned from Qwen3-VL-4B-Instruct on this data using LoRA with rank 32 and $\alpha = 64$. Training is performed for 3 epochs with a learning rate of $2.0 \times 10^{-5}$, batch size of 128, and AdamW optimizer, implemented using the LLaMA-Factory [50]. All experiments are conducted on NVIDIA A100 GPUs with 80GB of memory, requiring approximately 8 GPU-hours.

**Evaluation Protocol.** For HVCU-Bench evaluation, models are prompted to select from four options (A/B/C/D) for each multiple-choice question in a zero-shot setting. All models are evaluated with the temperature set to 0 to ensure deterministic outputs. We use a rule-based parser to extract answer choices from model outputs, treating unparseable or invalid responses as incorrect. Open-source models are evaluated locally with their default inference configurations, while proprietary models (e.g., GPT-4o) are accessed via official APIs with temperature=0.

Table 6. Performance of Qwen3-VL-4B-Bridge on HVCU-Bench. $+\Delta$ denotes gain over "base" setting.

| Implication Understanding | | | | Aesthetic Appreciation | | | | Affective Reasoning | | | | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | |
| **"Base" Setting** | | | | | | | | | | | | |
| 90.50 | 86.75 | 61.25 | 50.00 | 90.29 | 74.29 | 66.57 | 46.57 | 93.33 | 83.67 | 63.00 | 45.67 | 47.41 |
| **"Context" Setting** | | | | | | | | | | | | |
| 90.50 | 87.25 | 71.75 | 58.25 +8.25 | 90.29 | 76.57 | 85.14 | 58.57 +12.00 | 93.33 | 87.00 | 73.33 | 59.67 +14.00 | 58.83 +11.42 |

Table 7. Detailed Results on general benchmarks. The best score in each metric is **in-bold**.

| Model | MMBench | | | | | HallusionBench | | | MMStar | | | | | | | MMMU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CN(CC) | CN(Dev) | EN(Dev) | RU(Dev) | Avg. | qAcc | fAcc | aAcc | CP | FP | IR | LR | ST | MA | Avg. | Acc |
| Qwen3-VL-4B-Instruct | **68.82** | 81.36 | **83.93** | 77.42 | 77.88 | **36.48** | 36.71 | **60.25** | 77.03 | 60.22 | 69.43 | 53.26 | 43.72 | 41.02 | 57.45 | 48.56 |
| Qwen3-VL-4B-Bridge | 68.43 | **82.13** | 83.76 | **78.18** | **78.18** | 36.04 | **36.99** | 59.62 | **78.06** | **60.39** | **71.60** | **62.02** | **50.39** | **65.80** | **64.71** | **51.78** |

# D. More Experimental Details

## D.1. Detailed Results of Qwen3-VL-4B-Bridge

**Evaluation on HVCU-Bench.** Table 6 provides the complete results corresponding to Figure 4 (left), reporting the HVCU-Bench results of Qwen3-VL-4B-Bridge under both the "base" and "context" settings. In the "base" setting, Qwen3-VL-4B-Bridge achieves strong perceptual accuracy ($Acc_{perc} > 90\%$ across all tasks) while substantially improving $L_{bridge}$, $L_{conn}$, and $Acc_{full}$, reaching an overall score of 47.41%. The "context" setting reveals even stronger performance: Qwen3-VL-4B-Bridge achieves 58.83%, outperforming the base model Qwen3-VL-4B-Instruct by +3.73% (55.10% in Table 8). When transitioning from "base" to "context", Qwen3-VL-4B-Bridge shows substantial gains in full-chain accuracy, improving by +8.25%, +12.00%, and +14.00% on Implication Understanding, Aesthetic Appreciation, and Affective Reasoning, respectively. These results demonstrate that VCU-Bridge effectively teaches hierarchical reasoning chains that maintain superior performance across both evaluation settings, confirming that our training approach enhances the model's intrinsic reasoning capabilities rather than merely optimizing for specific evaluation conditions.

**Evaluation on General Benchmarks.** Table 7 provides the complete results corresponding to Figure 4 (right), comparing Qwen3-VL-4B-Instruct and Qwen3-VL-4B-Bridge on four general benchmarks: MMBench, HallusionBench, MMStar, and MMMU. We evaluate using the LMMS-Eval [48] toolkit and report the official metrics defined by each benchmark. For benchmarks with multiple sub-metrics (MMBench and MMStar), we also report the average score across all sub-metrics for easier comparison. Qwen3-VL-4B-Bridge maintains strong performance on MMBench and HallusionBench while achieving substantial gains on MMStar (Avg. +7.26%) and MMMU (+3.22%). Aggregated across all four benchmarks, Qwen3-VL-4B-Bridge achieves an average improvement of +2.53%. These results confirm that the hierarchical training data does not cause overfitting to HVCU-Bench, but instead induces hierarchical reasoning patterns that transfer to diverse external benchmarks.

## D.2. "Context" Setting of HVCU-Bench

Table 8 presents the full HVCU-Bench results under the "context" setting, which can be compared against the "base" setting results in Table 2. In the "context" setting, models are given predictions from preceding levels as additional context, allowing us to explicitly probe whether access to lower-level information facilitates higher-level connotative reasoning.

**Universal Improvement Across Models.** The results reveal a consistent trend: when provided with hierarchical context, almost every model exhibits substantial improvements across $Acc_{bridge}$, $Acc_{conn}$, and $Acc_{full}$. This phenomenon spans the entire spectrum of model capabilities, from weaker architectures like LLaVA-1.6-7B (+9.85%) to stronger models like GPT-4o (+15.94%). This universal improvement provides robust empirical evidence for strong inter-level dependency, confirming that connotative reasoning is causally linked to the availability and quality of foundational information rather than being an isolated capability that could operate independently of hierarchical context.

**Model-Specific Bottlenecks.** The magnitude of these gains reveals different limitations across models. Stronger models such as GPT-4o and Qwen3-VL-8B achieve the largest absolute improvements (+15.94% and +14.70% respectively), suggesting that their primary bottleneck in the "base" setting is missing spontaneous connectivity rather than insufficient reasoning

Table 8. Overall results on HVCU-Bench with "context" setting.

| Model | Model Size | Implication Understanding | | | | Aesthetic Appreciation | | | | Affective Reasoning | | | | *Score* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | |
| **Basic Reference** | | | | | | | | | | | | | | |
| GPT-4o | - | 95.50 | 89.75 | 76.50 | 65.00 | 95.43 | 82.29 | 87.71 | 72.86 | 91.33 | 86.00 | 80.67 | 66.67 | 68.18 |
| **Open-Source MLLMs** | | | | | | | | | | | | | | |
| Qwen3-VL-Instruct | 4B | 86.75 | 85.50 | 70.75 | 54.50 | 90.57 | 72.86 | 74.00 | 53.14 | 90.33 | 86.00 | 74.33 | 57.67 | 55.10 |
| Qwen3-VL-Instruct | 8B | 93.50 | 90.00 | 74.75 | 62.75 | 91.71 | 74.00 | 82.57 | 59.43 | 94.33 | 89.00 | 76.00 | 64.67 | 62.28 |
| LLaVA-1.6 | 7B | 81.75 | 68.00 | 54.50 | 32.75 | 79.14 | 43.71 | 50.00 | 18.29 | 92.00 | 65.00 | 30.00 | 18.67 | 23.24 |
| LLaVA-1.6 | 13B | 84.75 | 80.25 | 63.00 | 44.75 | 84.86 | 57.71 | 52.00 | 29.14 | 94.33 | 78.00 | 37.67 | 27.33 | 33.74 |
| Deepseek-VL2-tiny | MoE 1B/3B | 88.25 | 65.25 | 55.75 | 34.25 | 89.71 | 47.71 | 60.57 | 27.43 | 93.33 | 68.33 | 33.00 | 23.00 | 28.23 |
| Deepseek-VL2 | MoE 4.5B/27B | 93.75 | 84.25 | 67.50 | 53.75 | 95.14 | 59.43 | 54.00 | 33.43 | 96.33 | 83.67 | 62.00 | 52.00 | 46.39 |
| Gemma3 | 4B | 76.50 | 78.25 | 63.50 | 40.75 | 68.86 | 65.14 | 82.57 | 35.43 | 87.00 | 75.00 | 50.00 | 33.00 | 36.39 |
| Gemma3 | 12B | 87.50 | 88.00 | 74.50 | 57.00 | 82.86 | 72.57 | 84.86 | 50.00 | 90.67 | 86.00 | 74.33 | 59.67 | 55.56 |
| InternVL3.5 | 4B | 82.50 | 80.75 | 66.75 | 46.00 | 82.86 | 65.43 | 64.00 | 36.86 | 91.00 | 82.33 | 79.00 | 60.33 | 47.73 |
| InternVL3.5 | 8B | 82.00 | 84.75 | 66.00 | 47.25 | 84.00 | 70.57 | 71.14 | 45.43 | 86.00 | 81.67 | 68.00 | 50.33 | 47.67 |
| Phi-4-Multimodal-Instruct | 6B | 90.25 | 84.75 | 68.00 | 54.50 | 90.29 | 61.71 | 50.86 | 32.86 | 90.00 | 86.33 | 59.67 | 45.00 | 44.12 |
| Phi-3.5-Vision-Instruct | 4B | 84.25 | 84.00 | 71.50 | 51.25 | 88.29 | 63.43 | 64.00 | 40.00 | 91.33 | 81.33 | 64.67 | 49.33 | 46.86 |

capacity. In contrast, weaker models also improve, confirming they follow the same hierarchical logic, but their performance ceiling remains constrained by the absolute quality of their underlying perception ($Acc_{perc}$), a limitation that even perfect reasoning cannot fully offset. Together, these differential patterns reveal that hierarchical visual understanding requires both robust perception and effective cross-level connectivity.

## D.3. Application on Diverse Model Variants

To assess the generality of our hierarchical supervision across different model scales, architectures, and reasoning paradigms, Table 9 presents the impact of hierarchical instruction tuning on three categories of model variants.

**Scalability to Larger Models.** We instruction-tune Qwen3-VL-8B-Instruct on the same data to obtain Qwen3-VL-8B-Bridge. The results reveal a striking pattern: while the 8B model already achieves near-saturated perceptual accuracy ($Acc_{perc} > 93\%$ across all tasks), hierarchical tuning yields substantial gains at higher reasoning levels, with $Acc_{full}$ improving by +4.25%, +8.57%, and +1.00% on the three task families. This asymmetric improvement profile suggests that larger models already possess strong capabilities at individual levels, but lack the explicit pathways connecting them. Our hierarchical data acts as a structural scaffold that bridges the model's latent low-level perceptual knowledge with its high-level reasoning capacity, enabling cross-level integration. Rather than teaching new perceptual skills to an already-saturated foundation, the training data unlocks the model's ability to systematically propagate visual evidence through intermediate semantic reasoning to abstract interpretations, transforming isolated competencies into a coherent reasoning chain.

**Architectural Robustness.** To verify architectural generality, we apply the same data to LLaVA-1.6-7B, which employs a fundamentally different vision-language fusion mechanism. The model achieves substantial improvements in $Acc_{full}$ (+14.50% on Implication Understanding, +6.28% on Aesthetic Appreciation), with an overall score gain of +7.60%. Notably, the Aesthetic Appreciation task exhibits a slight $Acc_{perc}$ regression (-3.43%), likely reflecting LLaVA-1.6's relatively limited baseline perceptual capacity on this specific task. However, this minor drop at the perceptual level does not prevent significant gains at $Acc_{bridge}$ and $Acc_{conn}$, which is particularly revealing: it demonstrates that hierarchical supervision can successfully teach higher-level reasoning even when lower-level perception remains imperfect or unstable. In other words, the effectiveness of our training framework does not depend on a flawless perceptual foundation, but rather on establishing coherent cross-level reasoning pathways that can function robustly across diverse architectural designs and varying baseline capabilities, confirming the architectural universality of our hierarchical supervision approach.

**Reasoning-Specialized Models.** To examine whether models with built-in multi-step reasoning capabilities still benefit from hierarchical supervision, we apply the same data to Qwen3-VL-4B-Thinking, a reasoning-specialized variant. Despite surpassing the standard Qwen3-VL-4B-Instruct by +10.50% in $Acc_{full}$ on Implication Understanding, the Thinking model exhibits a notable weakness on Aesthetic Appreciation, where $Acc_{conn}$ drops by 27.71% compared to its non-thinking counterpart. This suggests that the extended reasoning process may interfere with aesthetic judgments that rely more on holistic perception than sequential logic. After hierarchical tuning, Qwen3-VL-4B-Thinking-Bridge achieves an overall score gain of +6.14%, with the most pronounced improvement on Aesthetic Appreciation ($Acc_{full}$ +14.00%), precisely where the baseline was weakest. This demonstrates that hierarchical supervision complements built-in reasoning capabilities, and is particularly effective at addressing task-specific weaknesses through structured intermediate reasoning pathways.

Table 9. Performance on diverse model variants. $+\Delta$ denotes gain over the model before instruction tuning.

| Model | Implication Understanding | | | | Aesthetic Appreciation | | | | Affective Reasoning | | | | *Score* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | |
| **Larger Model** | | | | | | | | | | | | | |
| Qwen3-VL-8B-Instruct | 93.50 | 89.50 | 59.50 | 50.75 | 91.71 | 73.43 | 63.43 | 44.00 | 94.33 | 84.67 | 60.00 | 48.00 | 47.58 |
| Qwen3-VL-8B-Bridge | 93.75 | 91.25 | 62.25 | 55.00 +4.25 | 91.71 | 78.29 | 70.00 | 52.57 +8.57 | 94.33 | 87.33 | 60.33 | 49.00 +1.00 | 52.19 +4.61 |
| **Heterogeneous Model** | | | | | | | | | | | | | |
| LLaVA-1.6-7B | 81.75 | 58.00 | 40.25 | 18.75 | 79.14 | 36.86 | 33.14 | 9.43 | 92.00 | 58.00 | 19.33 | 12.00 | 13.39 |
| LLaVA-1.6-7B-Bridge | 82.00 | 75.25 | 52.50 | 33.25 +14.50 | 75.71 | 43.43 | 47.14 | 15.71 +6.28 | 92.33 | 61.33 | 22.00 | 14.00 +2.00 | 20.99 +7.60 |
| **Reasoning-Specialized Model** | | | | | | | | | | | | | |
| Qwen3-VL-4B-Thinking | 95.75 | 85.50 | 64.75 | 53.75 | 93.14 | 66.29 | 32.29 | 24.00 | 96.33 | 84.33 | 63.00 | 51.00 | 42.92 |
| Qwen3-VL-4B-Thinking-Bridge | 96.50 | 88.50 | 66.25 | 57.50 +3.75 | 95.43 | 71.43 | 48.86 | 38.00 +14.00 | 97.00 | 85.00 | 64.33 | 51.67 +0.67 | 49.06 +6.14 |

Table 10. Comparison of data generation strategies on HVCU-Bench. $+\Delta$ denotes gain over Direct.

| Model | Implication Understanding | | | | Aesthetic Appreciation | | | | Affective Reasoning | | | | *Score* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | $Acc_{perc}$ | $Acc_{bridge}$ | $Acc_{conn}$ | $Acc_{full}$ | |
| Qwen3-VL-4B-Direct | 89.50 | 87.00 | 60.25 | 48.75 | 88.57 | 74.57 | 64.57 | 44.00 | 92.33 | 83.67 | 58.67 | 43.33 | 45.36 |
| Qwen3-VL-4B-MCTS | 90.00 | 88.00 | 60.50 | 49.00 +0.25 | 89.71 | 75.14 | 65.43 | 45.43 +1.43 | 92.33 | 85.33 | 60.00 | 44.33 +1.00 | 46.25 +0.89 |
| Qwen3-VL-4B-Bridge | 90.50 | 86.75 | 61.25 | 50.00 +1.25 | 90.29 | 74.29 | 66.57 | 46.57 +2.57 | 93.33 | 83.67 | 63.00 | 45.67 +2.34 | 47.41 +2.05 |

### D.4. Comparison with Other Data Generation Approaches

To validate the effectiveness of our MCTS-driven hierarchical generation, we compare it against two baselines: a direct generation approach and an MCTS-only variant without inter-level validation. All methods generate data from the same 1k Implication Understanding images, while Aesthetic Appreciation and Affective Reasoning use entirely different image distributions, enabling cross-distribution generalization assessment. Table 10 presents the results.

**Direct Generation Baseline.** We construct Qwen3-VL-4B-Direct by instruction-tuning the base model on data generated via a straightforward three-level prompting strategy. This baseline prompt instructs the MLLM to create questions at three difficulty levels (basic perception, connection & relationship, high-level reasoning) around the same topic, but does not enforce bottom-up construction or validate that each level's reasoning logically supports the next. While this approach produces superficially similar hierarchical structures, the resulting questions exhibit weak inter-level dependency, with each level potentially addressing independent aspects of the image rather than forming a coherent reasoning chain.

**MCTS-Only Baseline.** To isolate the contribution of MCTS tree search from inter-level validation, we construct Qwen3-VL-4B-MCTS by instruction-tuning on data generated using MCTS exploration but without the validation mechanism that enforces logical coherence between adjacent levels. This baseline retains the benefits of diverse path exploration and top-K selection, but does not verify that each level's reasoning logically supports the next.

**Performance Comparison.** As shown in Table 10, Qwen3-VL-4B-Bridge achieves a +2.05% overall score improvement over Direct. The intermediate MCTS-only baseline allows us to disentangle the two sources of this gain: MCTS tree search alone contributes +0.89% over Direct by exploring diverse reasoning paths, while inter-level validation provides an additional +1.16%, confirming that enforcing logical coherence between levels is the more critical factor. The advantages of our full pipeline are particularly pronounced on tasks from different distributions: Aesthetic Appreciation and Affective Reasoning exhibit substantially larger gains in $Acc_{full}$ (+2.57% and +2.34% over Direct). This cross-distribution generalization directly reflects the structural quality of our training data, confirming that our approach produces training data with both strong inter-level coherence and broad applicability across diverse reasoning tasks.

## E. Limitations

**Limitations of Three-Level Hierarchy.** Although HVCU-Bench and VCU-Bridge provide a first step toward systematically modeling visual connotative hierarchical understanding, several limitations remain. First, our formulation currently instantiates VCU-Bridge as a three-level discrete hierarchy over three task families (Implication Understanding, Aesthetic Appreciation, and Affective Reasoning). While this design captures a broad range of connotative phenomena, it is still a simplified approximation of human visual cognition, which may involve more continuous, overlapping, or multi-path reasoning processes. In particular, some real-world connotation and aesthetic judgments may not decompose cleanly into a single canonical chain from $L_{perc}$ to $L_{conn}$, and our benchmark cannot fully represent such richer structures.

**Scope of Task Coverage.** While HVCU-Bench currently encompasses three major task families (Implication Understanding, Aesthetic Appreciation, and Affective Reasoning), the spectrum of visual connotation extends beyond these categories. For instance, interpreting complex cultural allusions or inferring subtle social dynamics involving multiple characters often requires specialized knowledge that goes beyond the current scope. Additionally, our benchmark focuses on static images, whereas narrative connotations in sequential images or video remain an unexplored frontier. Future work could expand the VCU-Bridge framework to encompass these broader domains, investigating whether the same hierarchical bridging mechanisms apply to these more dynamic forms of visual reasoning.

**Cultural and Subjective Variability.** Visual connotation is inherently subjective and culturally situated. To maximize generality, we employ a rigorous multi-stage human validation process with annotators from diverse cultural backgrounds to ensure consensus, and deliberately select images and design questions emphasizing universal connotative patterns. However, despite these efforts, the interpretations in HVCU-Bench inevitably reflect certain cultural perspectives of the annotators and source data. Some symbolic meanings or aesthetic judgments may not generalize across all cultural contexts. Consequently, while our evaluation treats connotation as having a single "ground truth" for standardized benchmarking, this may not fully capture the diverse nature of human interpretation across cultures, which we acknowledge as a direction for future research.

## F. Ethical Considerations

**Copyright and Licensing.** It is essential to strictly follow all copyright and licensing regulations. All images in HVCU-Bench are sourced from publicly available datasets with appropriate research licenses. Data from sources that do not permit copying or redistribution will be explicitly excluded.

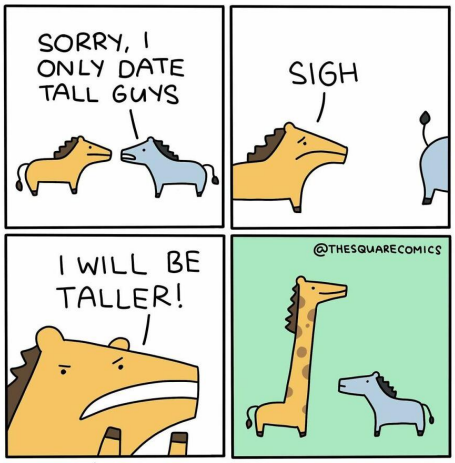
**Data Privacy.** Adherence to privacy laws and ethical standards in data handling is crucial. Annotators are explicitly instructed to avoid selecting images or creating questions that contain personally identifiable information. All selected images undergo privacy review before inclusion in the dataset.

## G. Prompts

We present the complete set of prompts used in both the HVCU-Bench benchmark construction pipeline and the hierarchical instruction data generation pipeline. The prompts are organized into two main categories: (i) *HVCU-Bench Generation and Validation Prompts*, which guide the sequential generation and interleaved validation of hierarchical question-answer chains to ensure logical coherence across $L_{perc}$, $L_{bridge}$, and $L_{conn}$, and (ii) *MCTS-Driven Data Generation Pipeline Prompts*, which support the tree search process for exploring diverse reasoning paths, including node generation at each hierarchical level and quality evaluation of candidate question-answer pairs.

**List of Prompts**

▷ ***Implication Understanding (Metaphor)***



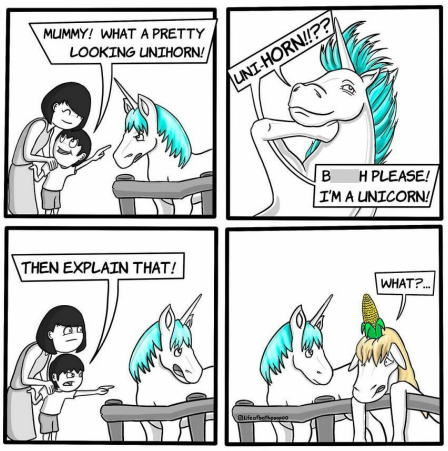

**Level 1 – Perception**
*Question:* What is the primary method of communication used by the characters in the comic strip?
**A**. Characters speak dialogue shown in speech bubbles.
**B**. They use only non-verbal actions like gestures and facial expressions.
**C**. Their communication is shown through thought bubbles above their heads.
**D**. A narrator provides descriptions of their thoughts and actions.
*Ground-truth answer:* **A**.

**Level 2 – Bridge**
*Question:* Based on the sequence of events in the comic, what is the most direct reason the yellow horse decides to become taller?
**A**. To gain a better vantage point for observing its surroundings.
**B**. To explore its own potential for extreme physical transformation.
**C**. To satisfy the blue horse's stated dating preference for tall individuals.
**D**. To develop a more imposing physique for self-defense.
*Ground-truth answer:* **C**.

**Level 3 – Connotation**
*Question:* What could the abrupt transformation of the shorter character into a giraffe in the final panel of the comic strip symbolize in terms of social commentary?
**A**. It symbolizes the importance of personal growth and improvement.
**B**. It represents a critique of the quest for physical perfection and the extremes to which people will go to achieve it.
**C**. The transformation illustrates the beauty of embracing one's unique nature rather than conforming.
**D**. The character symbolizes the absurdity of changing oneself to meet others' arbitrary standards.
*Ground-truth answer:* **D**.

Figure 6. A sample from the *Metaphor* aspect of *Implication Understanding*.

▷ ***Implication Understanding (Contrast)***




**Level 1 – Perception**
*Question:* What object is clearly visible on the head of one of the horse-like creatures in the final panel of the comic strip?
*Options:*
**A**. A flower.
**B**. A small hat.
**C**. A unicorn horn.
**D**. An ear of corn.
*Ground-truth answer:* **D**.

**Level 2 – Bridge**
*Question:* What is the primary misunderstanding that occurs between the child and the "unicorn" character in the initial panels of the comic strip?
*Options:*
**A**. The child is referring to a different animal with an actual corn "horn", but the "unicorn" character believes the child is misidentifying or insulting its own mythical horn.
**B**. The child simply mispronounced the word "unicorn," leading the creature to correct them indignantly.
**C**. The "unicorn" character mistook the child's comment as a criticism of its horn's appearance, thinking it resembled corn.
**D**. The "unicorn" character was primarily offended by the child's informal language, rather than the specific word "unihorn."
*Ground-truth answer:* **A**.

**Level 3 – Connotation**
*Question:* What is the underlying message conveyed through the humorous twist in the final panel of the comic strip, where an ear of corn is revealed instead of a traditional unicorn horn?
*Options:*
**A**. The humor emphasizes our inability to recognize the ordinary when we are obsessed with the extraordinary.
**B**. This humorously indicates that our expectations from mythical tales often overlook the charm and surprise found in nature's simplicities.
**C**. It mocks the convention of seeking deep, hidden meanings in every aspect of art by providing an unexpectedly literal twist.
**D**. It illustrates the importance of deeper investigation and not taking things at face value.
*Ground-truth answer:* **A**.

Figure 7. A sample from the *Contrast* aspect of *Implication Understanding*.

▷ ***Implication Understanding (Exaggeration)***

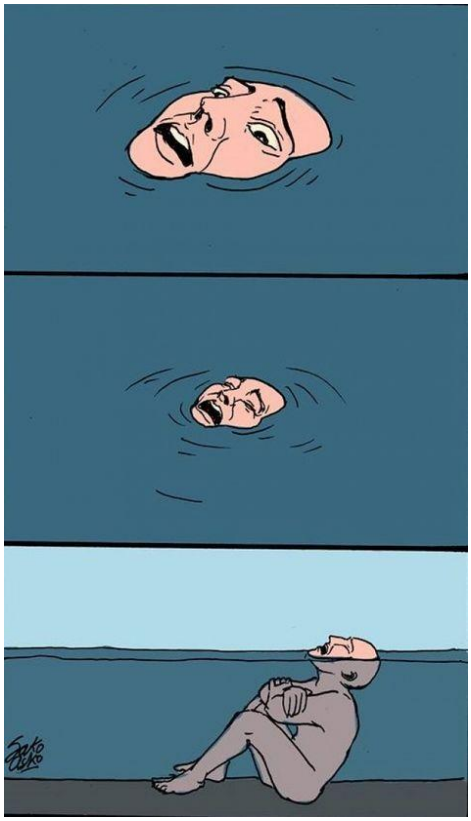

**Level 1 – Perception**
*Question:* What specific item is the bald character asked about in the first panel, and is visibly holding?
*Options:*
**A**. An extra set of headphones.
**B**. A mobile phone.
**C**. A book.
**D**. A pair of sunglasses.
*Ground-truth answer:* **A**.

**Level 2 – Bridge**
*Question:* In the comic, what specific action by the newly arrived character immediately explains why the bald character says "Speak of the Devil"?
*Options:*
**A**. He is holding a smartphone in a public space.
**B**. He is talking very loudly on his phone, disturbing the quiet environment.
**C**. He is about to ask the bald character for a set of headphones.
**D**. He appears to be ignoring the presence of other passengers.
*Ground-truth answer:* **B**.

**Level 3 – Connotation**
*Question:* What does the exaggerated action in the last panel symbolize?
*Options:*
**A**. An overblown representation of how technology can suffocate personal freedoms.
**B**. The frustrations of public transit riders with loud music.
**C**. The extreme discomfort caused by modern technology's invasion into personal space.
**D**. A visual hyperbole of the silent plea for etiquette and respect in shared environments.
*Ground-truth answer:* **D**.

Figure 8. A sample from the *Exaggeration* aspect of *Implication Understanding*.

▷ ***Implication Understanding (Dislocation)***



**Level 1 – Perception**
*Question:* How many distinct panels or frames are present in the image?
*Options:*
**A**. One.
**B**. Four.
**C**. Two.
**D**. Three.
*Ground-truth answer:* **D**.

**Level 2 – Bridge**
*Question:* How does the third panel visually recontextualize the seemingly dire situation?
*Options:*
**A**. Shows consistent water level and horizon line.
**B**. Shows protagonist sitting with crossed arms.
**C**. Reveals shallow depth of the water, proving the initial panic was an overreaction.
**D**. Shifts to a wider shot including surroundings.
*Ground-truth answer:* **C**.

**Level 3 – Connotation**
*Question:* What hidden meaning might the comic be conveying through the exaggerated expressions followed by the reveal?
*Options:*
**A**. It satirizes people's tendency to overreact to non-dangerous situations.
**B**. It shows preparation for worst-case scenarios.
**C**. It illustrates emotional oscillation between terror and relief.
**D**. It mocks sensationalizing everyday events.
*Ground-truth answer:* **A**.

Figure 9. A sample from the *Dislocation* aspect of *Implication Understanding*.

▷ ***Implication Understanding (Symbolism)***


**Level 1 – Perception**
*Question:* What animals are depicted interacting in the comic panels?
*Options:*
**A**. Two dogs.
**B**. A dog and a cat.
**C**. A dog and a squirrel.
**D**. Two cats.
*Ground-truth answer:* **A**.

**Level 2 – Bridge**
*Question:* What is the source of the humor in the final panel?
*Options:*
**A**. The brown dog is actually a human in disguise and speaks perfect English.
**B**. The brown dog claims it cannot "speak" the specific sound "ARF") used by the grey dog, treating a simple bark as a foreign language.
**C**. The grey dog is mute and cannot respond to the brown dog's questions.
**D**. The human owner misunderstands the barking as a conversation about food.
*Ground-truth answer:* **B**.

**Level 3 – Connotation**
*Question:* What social phenomenon is implied by the "language barrier" between the two dogs?
*Options:*
**A**. The biological inability of different species to communicate.
**B**. The tendency for people to talk past each other when angry.
**C**. Even within similar groups, arbitrary linguistic or cultural differences can create barriers to communication.
**D**. Pets often mimic the behavior and language of their owners.
*Ground-truth answer:* **C**.

Figure 10. A sample from the *Symbolism* aspect of *Implication Understanding*.

▷ ***Affective Reasoning (Fear)***

**Level 1 – Perception**
*Question:* What object is the person in the image holding?

*Options:*
**A**. A smartphone.
**B**. A hand mirror.
**C**. A magnifying glass.
**D**. A framed photograph.
*Ground-truth answer:* **B**.

**Level 2 – Bridge**
*Question:* What is the primary emotional dynamic created by the interplay between the woman's facial expression and the visible injuries in the mirror?

*Options:*
**A**. The woman's look of surprise indicates she is seeing the injury for the first time, which makes the viewer feel like a witness to a tragic discovery.
**B**. The woman's expression of horror validates and amplifies the viewer's reaction to the injury, creating a feedback loop of shared shock and disgust.
**C**. The gruesome injuries are the sole source of the viewer's disgust, and the woman's expression simply serves to confirm the reality of the situation.
**D**. The woman's horrified expression creates empathy, which conflicts with the feeling of aversion caused by the injury, resulting in a confusing mix of pity and disgust.
*Ground-truth answer:* **B**.

**Level 3 – Connotation**
*Question:* You think the emotional transfer of this picture is perceived as either direct or indirect? Please explain your perspective.

*Options:*
**A**. The emotional conveyance is direct, but it primarily communicates feelings of surprise and sadness. The composition suggests the character is seeing the scar for the first time, leading to a moment of shocking self-discovery. The viewer directly experiences this surprise through the close-up perspective. The primary emotional impact is not fear, but a sudden, shared realization of a tragic disfigurement.
**B**. The emotional transfer is indirect as it functions on a symbolic level. The magnifying glass represents societal judgment and scrutiny, and the scar symbolizes past trauma. The viewer does not react to the physical image itself but to the abstract concepts of judgment and suffering. This requires an intellectual interpretation of the symbolism before any emotional response is triggered, making the conveyance indirect and contemplative.
**C**. The emotional conveyance of this image is direct. It showcases a close-up of the character's face through a hand mirror, particularly highlighting the scar on the face. This emphasis on detail immediately draws the viewer's attention and sympathy towards the character's experiences, while simultaneously evoking fear and aversion towards the horrifying facial details. The character's expression and the use of the hand mirror enhance the visual impact, allowing viewers to quickly grasp the emotions the image intends to convey. This direct visual technique effectively captures the viewer's attention and prompts them to reflect on the character's inner emotions.
**D**. The emotional transfer is indirect because it relies on the viewer to construct a narrative. The magnifying glass suggests a story of investigation or scientific study, and the scar is a clue. The emotion is not felt immediately but is deduced after the viewer contemplates the character's potential backstory and the events that led to the injury. This cognitive process of storytelling makes the emotional experience indirect and intellectual.
*Ground-truth answer:* **C**.

Figure 11. A sample from the *Fear* aspect of *Affective Reasoning*.

▷ ***Affective Reasoning (Joy)***

**Level 1 – Perception**
*Question:* What prominent, colorful object is being held up by the people amidst the splashing water?
**A**. A large, open umbrella.
**B**. A large rainbow flag.
**C**. A multi-colored beach towel.
**D**. A colorful plastic tarp.
*Ground-truth answer:* **B**.

**Level 2 – Bridge**
*Question:* Why is the atmosphere in the image best characterized as celebratory, rather than simply playful or disorganized?
**A**. The participants' uninhibited splashing and visible smiles are the primary drivers of the mood, indicating a scene of spontaneous, individual fun.
**B**. The combination of a prominent flag and the occupation of a public fountain suggests a demonstration, where the energy is more disruptive than joyful.
**C**. The large rainbow flag provides a shared symbol that gives the energetic activity in the water a sense of collective purpose and joyous expression.
**D**. The energy stems from the refreshing relief the fountain provides on what appears to be a hot day, making the activity a practical response to the environment.
*Ground-truth answer:* **C**.

**Level 3 – Connotation**
*Question:* Which specific elements in the image (such as color, setting, people, etc.) triggered your several main emotional responses? please provide a detailed explanation.
**A**. My emotional response is driven by the overcast sky and the muted color palette of the people's clothing. The grey, cloudy sky casts a somber light over the scene, suggesting a melancholic or pensive mood. The lack of vibrant colors in the attire, apart from the flag, contributes to a sense of uniformity and blandness. These elements together create a feeling of calmness bordering on sadness, reflecting a subdued and uneventful day.
**B**. The vibrant rainbow flag and the splashing water in the image evoke a strong emotional response. The rainbow-colored flag adds vividness and liveliness to the scene, catching the eye and conveying a celebratory atmosphere through its prominent position in the frame. The splashing water enhances the sense of energy and joy, as people play freely and carefreely in the water. Together, these elements create a positive, inclusive, and celebratory mood, bringing feelings of delight and surprise.
**C**. The image evokes a sense of anger and frustration due to the apparent disorder. The sight of people wading in a public fountain, combined with the prominent display of a flag, suggests a protest or public disruption. This disregard for public property and order triggers a negative response, as it feels chaotic and disrespectful. The splashing water, in this context, is not playful but rather a sign of recklessness, leading to feelings of irritation.
**D**. The primary emotional triggers are the dense crowd and the architectural style of the background buildings. The large number of people packed together evokes a sense of claustrophobia and anxiety, while the imposing, formal architecture creates a feeling of being small and insignificant. This combination leads to a feeling of unease and social discomfort, as the scene appears overwhelming and impersonal.
*Ground-truth answer:* **B**.

Figure 12. A sample from the *Joy* aspect of *Affective Reasoning*.

▷ ***Affective Reasoning (Wonder)***

**Level 1 – Perception**
*Question:* What kind of animal is sitting on the rock in the center of the image, looking up towards the sky?
**A**. A wolf.
**B**. A fox.
**C**. A dog.
**D**. A cat.
*Ground-truth answer:* **C**.

**Level 2 – Bridge**
*Question:* How does the use of light in the image primarily establish the relationship between the earthly foreground and the fantastical moon?
**A**. The diffused glow illuminating the mist and mountains establishes a realistic nocturnal setting, creating a dominant atmosphere of quiet serenity and peace.
**B**. "The light selectively leaves the abandoned cabin in deep shadow, making its decay the focal point and evoking a primary sense of melancholy and loss.
**C**. By casting a direct, focused beam from the moon to the dog, the light creates a narrative link that transforms physical distance into a moment of connection and wonder.
**D**. The sharp contrast between the dark landscape and the bright moon serves to isolate the two realms, emphasizing a feeling of insurmountable distance and loneliness.
*Ground-truth answer:* **C**.

**Level 3 – Connotation**
*Question:* Which specific elements (for example, colors, scenes, people, etc.) in the image do you find to evoke the several main emotional reactions from you? please provide a detailed explanation.
**A**. Several elements in this image evoke strong emotional responses. First, the colors: the dark tones in the scene contrast sharply with the soft moonlight, creating a mysterious and serene atmosphere that catches the eye and sparks a sense of wonder. Next, the setting: the abandoned wooden cabin, rugged mountain rocks, and the distant moon form a dreamlike scene, evoking feelings of both loneliness and longing. Finally, the characters: the dog sitting on the rock and the silhouette of a person on the moon add a narrative quality to the image, as if telling a distant and ancient story.
**B**. The image primarily evokes a sense of fear and anxiety. The dark, ominous tones combined with the dilapidated, abandoned cabin suggest a scene of horror or danger. The rugged, sharp rocks and the isolated setting create a feeling of vulnerability and entrapment. The dog appears tense, as if sensing a threat, and the silhouette on the moon is a menacing, watchful figure. This combination of elements fosters a deep sense of unease and foreboding, making the viewer feel tense and scared.
**C**. This image communicates overwhelming sadness and melancholy. The abandoned cabin is a clear symbol of loss and forgotten times, evoking a deep sense of grief. The barren, rocky landscape emphasizes the feeling of utter loneliness and despair. The single dog, sitting alone, appears to be waiting for a companion who will never return, adding a layer of poignant sorrow. The moonlight casts a cold, mournful glow, amplifying the feeling of desolation and making the entire scene feel tragic.
**D**. The image evokes a sense of vibrant energy and excitement. The sharp contrast between the light and dark areas creates a dynamic, high-energy feeling, like a flash of lightning. The rugged mountains suggest adventure and challenge, sparking a desire for action. The cabin, while old, looks like a basecamp for an exciting exploration. The dog is poised as if about to spring into action, and the figure on the moon appears to be a triumphant explorer. The entire scene feels like the beginning of a thrilling journey, filling the viewer with anticipation and exhilaration.
*Ground-truth answer:* **A**.

Figure 13. A sample from the *Wonder* aspect of *Affective Reasoning*.

▷ ***Affective Reasoning (Anger)***

**Level 1 – Perception**
*Question:* What primary emotion is the woman's facial expression conveying?
**A**. Sadness.
**B**. Anger.
**C**. Surprise.
**D**. Joy.
*Ground-truth answer:* **B**.

**Level 2 – Bridge**
*Question:* How does the background interact with the woman's expression?
**A**. The intense red of the background amplifies the raw anger in her expression, creating a more powerful and overwhelming feeling of rage.
**B**. The red background is the primary source of aggression, causing the woman's expression to be interpreted as a reaction of fear or feeling cornered.
**C**. The flat, theatrical quality of the red background contrasts with the realistic expression, creating an emotional ambiguity between genuine rage and a staged performance.
**D**. The chaotic details in the red background, like graffiti and decay, suggest a specific narrative cause for her anger, making it feel targeted at her surroundings.
*Ground-truth answer:* **A**.

**Level 3 – Connotation**
*Question:* What elements create the emotional responses you feel?
**A**. The image primarily conveys a profound sense of sadness and despair. The character's wide-open mouth is not a scream of anger but a cry of anguish and loss. The stark red background isolates her, symbolizing her inner pain and emotional turmoil, trapping her in her grief. Her wide eyes are filled with hopelessness, not rage. The entire composition communicates a deep, personal tragedy, making the viewer feel empathy and a shared sense of sorrow for her suffering.
**B**. The scene appears to be one of theatrical performance, which inspires a sense of awe and surprise. The red is reminiscent of a stage curtain, suggesting a dramatic reveal, and the character's expression is one of pure astonishment, as if witnessing something incredible for the first time. This creates a feeling of wonder and anticipation in the viewer, rather than anger or fear. The scene feels exciting and spectacular, drawing the viewer in with a sense of dramatic tension and curiosity about what she is seeing.
**C**. The primary emotional trigger is the lack of a detailed environment, focusing solely on the character against a plain background. This minimalism creates a sense of loneliness and isolation. The character's expression seems to be one of confusion and vulnerability, as if she is lost or abandoned. The red color does not feel aggressive but rather empty and vast, amplifying her solitude. The main emotions evoked are therefore pity and a gentle melancholy for the character's apparent plight.
**D**. The intense red background and the expression of the character in the image evoke the strongest emotional reaction in me. Red is often associated with passion and power, and here it may symbolize anger or intense emotion. Combined with the character's expression, particularly her wide-open mouth and wide eyes, it conveys rage, fury, or extreme dissatisfaction. This sense of anger is also transmitted to the viewer, becoming the primary emotion evoked. The combination of this expression with the red background enhances the overall tension and unease, creating a sense of oppression that makes the viewer feel fear. Additionally, the intense visual content and color scheme can cause slight discomfort, thereby eliciting a sense of aversion.
*Ground-truth answer:* **D**.

Figure 14. A sample from the *Anger* aspect of *Affective Reasoning*.

▷ ***Affective Reasoning (Affection)***

**Level 1 – Perception**
*Question:* What three main subjects are gathered at the table?
**A**. A man and a young girl.
**B**. A man, a young girl, and a dog.
**C**. A man, a young girl, and a cat.
**D**. A woman, a young boy, and a dog.
*Ground-truth answer:* **B**.

**Level 2 – Bridge**
*Question:* How do their actions and expressions establish the emotional tone?
**A**. The father's expression of gentle nostalgia as he looks down suggests a bittersweet moment, while the daughter's wide-eyed look shows her concern for his feelings.
**B**. The father's gentle gaze towards the girl, combined with her look of happy surprise and the presence of a birthday cake, creates a shared moment of familial joy.
**C**. The scene's happiness stems primarily from the dog, which is being presented as a gift, causing the daughter's excitement and the father's look of satisfaction.
**D**. The daughter's expression of eager anticipation is directed at the cake, while the father's observant look suggests he is waiting for her reaction, creating a sense of suspense.
*Ground-truth answer:* **B**.

**Level 3 – Connotation**
*Question:* Describe the emotional content (valence, arousal, dominance).
**A**. The picture depicts a father and daughter in a moment of disagreement. The father, with a stern expression, holds the dog back, which appears agitated. The daughter looks away from the cake, her face showing disappointment and sadness. The cartoon style contrasts with the underlying tension. The main emotions are sadness, frustration, and tension. This scene has a low valence, a high level of arousal due to the conflict, and a low level of dominance as the characters feel constrained by the negative situation.
**B**. The picture depicts a scene where the father and the family's little dog are celebrating the daughter's birthday together. The father, wearing glasses, looks at his daughter with a kind expression, holding the little dog in his hands. The dog, with its tongue hanging out, gazes at the daughter and the cake, while the daughter looks back at her father and the dog, creating a joyful and heartwarming family atmosphere that brings happiness and delight. On the table, there are two pieces of cake, which are visually appealing, delicious, and surprisingly delightful. The entire scene has been rendered in a cartoonish style, maintaining a certain level of neutrality. The levels of joy and dominance are both high, evoking feelings of happiness and surprise. The arousal level is moderate, reflecting a warm and positive emotional tone.
**C**. The image portrays a family preparing for a pet competition. The father is holding their small dog, showcasing it to his daughter who acts as a judge. The dog seems eager, while the daughter's expression is one of serious concentration. The cakes on the table are rewards. The primary emotions are anticipation, concentration, and a sense of pressure. This results in a moderate valence, high arousal due to the competitive tension, and high dominance reflecting the characters' focus and control.
**D**. This is a scene of a family saying goodbye. The father is about to leave and is holding the family dog for his daughter to pet one last time. The daughter looks up at him with a sad expression. The cake on the table is a farewell treat. The emotions conveyed are sadness, love, and longing. The cartoon rendering softens the sad theme but doesn't erase it. The valence is low, arousal is moderate due to the poignant emotions, and dominance is low, signifying a sense of loss of control over events.
*Ground-truth answer:* **B**.

Figure 15. A sample from the *Affection* aspect of *Affective Reasoning*.

▷ ***Affective Reasoning (Sadness)***


**Level 1 – Perception**
*Question:* What type of animal is the main subject of this close-up photograph?
**A**. A monkey.
**B**. A cat.
**C**. A dog.
**D**. A bear cub.
*Ground-truth answer:* **C**.

**Level 2 – Bridge**
*Question:* How do the different parts of the pug's facial expression interact to create a sense of emotional ambiguity?
**A**. The primary source of ambiguity lies solely within the pug's eyes, which appear both large and curious due to light reflection while also having a sad, downward-turned shape.
**B**. The expression is a result of the pug's physical structure; the bared teeth are a common trait of the breed's underbite and not an emotional sign, while the wide eyes indicate a state of high alert or fear.
**C**. The deep wrinkles on the pug's forehead, suggesting worry, conflict with the dramatic, high-contrast lighting, which gives the image an aggressive and menacing quality.
**D**. The pug's wide, seemingly sad eyes contrast with its bared teeth, which could be interpreted as either aggression or a playful smile, creating an uncertain emotional signal.
*Ground-truth answer:* **D**.

**Level 3 – Connotation**
*Question:* When you look at this image, do you feel any conflicting emotions or uncertainty?
**A**. Upon viewing the image, there is no sense of emotional conflict. The pug's expression is one of clear and unambiguous aggression and anger. The bared teeth are a sign of a snarl, and the tension in its facial muscles indicates hostility. The stark, high-contrast black and white photography amplifies this feeling of menace. The viewer is meant to feel intimidated and fearful, as the dog is asserting its dominance in a threatening manner, leaving no room for emotional ambiguity.
**B**. When viewing this image, one might experience an emotional conflict. The expression of the pug in the picture appears somewhat melancholic yet with a hint of playfulness. Its teeth are slightly exposed, as if it's smiling, but there's a touch of seriousness in its eyes. This mixed expression makes it difficult to determine the dog's emotional state, creating a sense of uncertainty that manifests in the viewer primarily holding a neutral emotion. At the same time, the black and white tones add a sense of oppression to the scene, potentially evoking the viewer's curiosity and sympathy towards the dog's emotional state, leading to feelings of sadness and fear.
**C**. This image presents a straightforward and non-conflicting emotional state of pure curiosity. The pug's head is tilted, and its eyes are focused, suggesting it is intently observing something just out of frame. The slightly open mouth and exposed teeth are characteristic of a dog's relaxed, inquisitive posture. The black and white tones create a dramatic effect that highlights the intensity of the dog's focus. The viewer primarily feels a sense of engagement and wonder, not emotional conflict or sadness.
**D**. There is no emotional uncertainty in this image; it is a clear depiction of physical discomfort and pain. The pug's squinted eyes and bared teeth are not related to any complex emotion but are physical reactions to an unpleasant sensation, such as an injury or illness. The black and white aesthetic lends a somber, clinical quality to the scene, emphasizing the animal's suffering. The viewer's response is one of sympathy and concern, driven by the unambiguous signs of physical distress, not emotional conflict.
*Ground-truth answer:* **B**.

Figure 16. A sample from the *Sadness* aspect of *Affective Reasoning*.

▷ ***Aesthetic Appreciation (Graphic)***


**Level 1 – Perception**
*Question:* What type of animal is the main subject of this advertisement?

**A**. A rabbit
**B**. A hamster
**C**. A puppy
**D**. A kitten

*Ground-truth answer:* **D**.

**Level 2 – Bridge**
*Question:* By comparing the kitten image to the surrounding text and graphics, what is the primary visual tension or conflict within the ad's composition?

**A**. The blur on the kitten is a deliberate technique to create depth of field, pushing it into the background to make the "ADOPT a PET" text the main focus.
**B**. There is a conflict in balance; the visually heavy kitten on the left competes for attention with the text block on the right, dividing the viewer's focus.
**C**. There is a conflict in style; the realistic photograph of the kitten clashes with the flat, illustrative style of the heart graphics and typography.
**D**. There is a conflict in sharpness; the kitten, the emotional focus, is blurry and indistinct, while the text and logos are crisp and clear.

*Ground-truth answer:* **D**.

**Level 3 – Connotation**
*Question:* Does this design have any issues that affect its effectiveness? If so, what are the main impacts?

**A**. Yes, the lack of sharpness in the main subject significantly weakens the design's effectiveness. Even though the text is legible, this visual flaw undermines the emotional connection with the audience and damages the organization's perceived professionalism.
**B**. Yes, the inconsistent and overly playful typography creates a sense of disorganization that is difficult to read quickly. This confusing hierarchy detracts from the message's urgency and makes the brand appear less credible.
**C**. No, the design does not have significant issues; in fact, the soft focus on the subject creates an artistic, dreamy effect that captures attention. This stylistic choice encourages the viewer to focus on the main "Adopt a Pet" message, enhancing its communicative power.
**D**. The design has a minor issue with image clarity, but its impact is minimal. The strong, legible headline and clear contact information effectively compensate for this, ensuring the core message of adoption is still successfully communicated.

*Ground-truth answer:* **A**.

Figure 17. A sample from the *Graphic* aspect of *Aesthetic Appreciation*.

▷ ***Aesthetic Appreciation (Color)***


**Level 1 – Perception**
*Question:* What is the color of the large circular area on the left side of the image where the text is located?

**A**. Light Green
**B**. Light Yellow
**C**. Light Blue
**D**. White

*Ground-truth answer:* **A**.

**Level 2 – Bridge**
*Question:* Why might a potential adopter have difficulty reading the contact information for the "Claytoni Shelter" in this advertisement?

**A**. The font size for the contact information is too small compared to the headline, making it the primary reason it is difficult to read.
**B**. The white text of the address and phone number has very low color contrast against the light green background, making it nearly illegible.
**C**. The circular text box is placed over a visually complex part of the background image, which interferes with the text.
**D**. The large word "WAITING" is so visually dominant that it draws the eye away from the contact details, making them hard to find.

*Ground-truth answer:* **B**.

**Level 3 – Connotation**
*Question:* Does this design have any issues that affect its effectiveness? If so, what are the main impacts?

**A**. No, the design is highly effective and communicates a gentle, caring brand identity. The soft pastel color palette creates a calming, approachable atmosphere that enhances the emotional appeal of the subject matter and strengthens the brand's positioning.
**B**. While the text is somewhat faint, the overall minimalist aesthetic is clean, modern, and visually pleasing. The subtle color treatment is a reasonable trade-off that supports a non-aggressive tone, resulting in a minor, acceptable impact on readability.
**C**. Yes, the design's color choices create significant legibility problems that obscure vital information. Even if the palette intends to be soft and gentle, this choice severely undermines communication effectiveness and damages the organization's credibility by appearing unprofessional.
**D**. Yes, the design's typography feels dated and generic, which fails to capture attention effectively. This weak font choice makes the brand seem uninspired and reduces the overall persuasive power of the message, even with a strong central image.

*Ground-truth answer:* **C**.

Figure 18. A sample from the *Color* aspect of *Aesthetic Appreciation*.

▷ ***Aesthetic Appreciation (Font)***


**Level 1 – Perception**
*Question:* What color is the handwritten text that reads “limited offer”?

**A**. Beige
**B**. Black
**C**. Red
**D**. White

*Ground-truth answer:* **C**.

**Level 2 – Bridge**
*Question:* How do the visual characteristics of the “SPRING SALE” text and the “limited offer” text create different impressions for the viewer?

**A**. The “SPRING SALE” text is bold to show importance, while the “limited offer” text is red to serve as a warning to the customer.
**B**. The “SPRING SALE” text establishes a modern, minimalist theme, while the “limited offer” text adds a personal, artistic touch.
**C**. Both text elements are designed primarily to grab attention, using different fonts to distinguish the main headline from the secondary condition.
**D**. The “SPRING SALE” text uses an elegant, formal font suggesting high quality, while the “limited offer” text uses a casual, handwritten font suggesting informal urgency.

*Ground-truth answer:* **D**.

**Level 3 – Connotation**
*Question:* Does this design present any significant issues affecting its overall effectiveness? If so, what are the primary high-level impacts on communication and brand perception?

**A**. The design has a minor issue with the handwritten font feeling slightly out of place, but its impact is minimal. The clarity of the main “Spring Sale” headline and the high-quality photography are strong enough to maintain the brand’s premium positioning and ensure the message is communicated effectively.
**B**. Yes, the design suffers from an incongruous text element that clashes with the established sophisticated aesthetic. Even though the urgent message is legible, this stylistic mismatch undermines the brand’s perceived professionalism and creates a visual dissonance that can cheapen the overall impression.
**C**. No, the design is highly effective because the mix of formal and informal typography creates a dynamic contrast that captures attention. This stylistic choice makes the promotion feel more accessible and urgent, ultimately boosting engagement without harming the core brand identity.
**D**. Yes, the placement of the text elements creates a cluttered and unbalanced composition that competes with the main image. This poor spatial organization weakens the visual hierarchy, making it harder for viewers to process the information efficiently and grasp the key message at a glance.

*Ground-truth answer:* **B**.

Figure 19. A sample from the *Font* aspect of *Aesthetic Appreciation*.

▷ ***Aesthetic Appreciation (Composition)***


**Level 1 – Perception**
*Question:* What is the primary object that serves as the background for most of the text in the image?

**A**. A cutting board
**B**. A piece of parchment paper
**C**. A baker's peel
**D**. A serving platter

*Ground-truth answer:* **A**.

**Level 2 – Bridge**
*Question:* How is the quote "A party without cake is just a meeting" spatially organized in relation to the cutting board illustration?

**A**. The quote begins inside the top of the cutting board and flows downwards, with only the author's name appearing outside the board.
**B**. The entire quote is contained within the boundaries of the cutting board, with the most important words enlarged for emphasis.
**C**. The main subjects, "party" and "cake", are positioned outside the board to draw attention, while the rest of the phrase is inside.
**D**. The first two words, "A party", are positioned outside the cutting board, while the remainder of the quote is located inside it.

*Ground-truth answer:* **D**.

**Level 3 – Connotation**
*Question:* Does this design exhibit any issues that compromise its overall effectiveness? If so, what are the primary consequences for the viewer's experience and the design's perceived quality?

**A**. No, the design does not have significant issues; in fact, its unconventional typography is effective at capturing attention. This dynamic arrangement adds visual interest and energy, reinforcing the celebratory theme of the quote.
**B**. While the separation of the first two words is slightly unconventional, it creates a unique visual entry point into the quote. The overall charming, handcrafted aesthetic is strong enough to compensate for this minor quirk, ensuring the design remains effective and endearing.
**C**. Yes, the muted, monochromatic color palette makes the design feel dated and unexciting. This lack of vibrant color fails to convey the joyful nature of a "party", thereby weakening the message's emotional impact and reducing its overall memorability.
**D**. Yes, the design suffers from disjointed text placement that fragments the central message. Even if the playful font is appealing, this structural flaw disrupts reading flow and undermines the design's sense of polish and professionalism.

*Ground-truth answer:* **D**.

Figure 20. A sample from the *Composition* aspect of *Aesthetic Appreciation*.

## Prompt for $L_{conn}$ Generation

# ROLE
You are an expert AI Benchmark Analyst. Your task is to deconstruct a given question and its context to find the correct answer, select the best distractors, and state the core reasoning.

# TASK
Analyze the provided JSON data, which contains ground-truth information, a `question`, and a list of potential `options`. Perform the following steps:
1. Use the image to understand the ground truth. If an image is provided, the `explanation` text is supplementary.
2. Read the `question` and evaluate all `options` against the image.
3. Identify the single best `option` that is most strongly supported by the ground truth.
4. From the remaining incorrect options, select the **three most plausible and confusing distractors**.
5. Write a concise, one-sentence `reasoning` that explains *why* the correct option is correct, based *strictly* on the image.

# INPUT CONTEXT
The input is a single JSON object containing the raw data for a question. It contains the following structure:
- `explanation`: A text string providing context for the scenario.
- `question`: The question to be analyzed.
- `options`: An array of strings, where one is the correct answer and the others are potential distractors.

# OUTPUT FORMAT
Return a single JSON object with the following structure:

```
{
  "level": 3,
  "level_name": "Connotation",
  "question": "The original question text",
  "options": [
    {
      "option_text": "Text of the correct option",
      "is_correct": true
    },
    {
      "option_text": "Text of the first distractor",
      "is_correct": false
    },
    {
      "option_text": "Text of the second distractor",
      "is_correct": false
    },
    {
      "option_text": "Text of the third distractor",
      "is_correct": false
    }
  ],
  "reasoning": "The one-sentence explanation for why the correct answer is correct,
               based on the provided ground truth."
}
```

# IMPORTANT RULES
- The `question` in the output must be identical to the input `question`.
- The output `options` array must contain exactly **four** options: the single correct answer and the three most plausible distractors selected from the original list.
- Ensure exactly one option is marked as `is_correct: true`.
- The `reasoning` is the most critical part. It must be clear and directly derivable from the image.
- Return ONLY the JSON object. Do not explain or add extra text.

Input:

```
{json_data}
```

Figure 21. Prompt for $L_{conn}$ generation on HVCU-Bench.

## Prompt for $L_{bridge}$ Generation

# ROLE
You are an expert AI Benchmark Crafter. You create questions that build on each other to test understanding.

# TASK
You are given the analysis of a Level 3 question. Your task is to create a Level 2 (Comprehensive Understanding) question that logically precedes it.
The Level 2 question should address the "how" or "why" of the situation, bridging Level 1 facts and the Level 3 conclusion. It must be answerable using the image and help users understand the `reasoning` in the `level_3_analysis`.

# INPUT CONTEXT
The input is a JSON object containing the following keys:
- `explanation`: Supplementary text context. The primary ground truth is the image.
- `level_3_analysis`: The L3 QA and reasoning. Your L2 question must be a logical prerequisite for this.

# OUTPUT FORMAT
Return a single JSON object for the Level 2 question:

```
{
  "level": 2,
  "level_name": "Semantic Bridge (Comprehensive Understanding)",
  "question": "Your newly generated Level 2 question",
  "options": [
    {
      "option_text": "The correct answer text",
      "is_correct": true
    },
    {
      "option_text": "A plausible but incorrect distractor",
      "is_correct": false
    },
    {
      "option_text": "Another plausible but incorrect distractor",
      "is_correct": false
    },
    {
      "option_text": "A third plausible but incorrect distractor",
      "is_correct": false
    }
  ]
}
```

# IMPORTANT RULES
- The generated question and all options must be derived *only* from the image.
- The question should focus on understanding the relationships, causes, or implications described in the provided context.
- **DISTRACTOR DIFFICULTY REQUIREMENTS FOR LEVEL 2 (CRITICAL):**
  - Distractors must require **deliberate analysis** to rule out, not just superficial pattern matching.
  - All distractors must be from the same **semantic/visual domain** as the correct answer and share overlapping features.
  - Each distractor should be **partially correct**—it may align with part of the scene, concept, or logic—but contains a **crucial flaw**, such as:
    - Oversimplification of a process or mechanism
    - Misidentification of cause and effect
    - Confusion between similar entities, directions, or states
    - Overgeneralization from a local detail
  - Include at least one distractor that reflects a **common but incorrect heuristic**, such as:
    - Selecting the most visually salient item regardless of function
    - Assuming temporal sequence implies causality
    - Mistaking correlation for explanation
  - Avoid options that are impossible, irrelevant, or rely on external knowledge.
  - Test the ability to: differentiate superficial vs. deep correctness, resolve ambiguous cues, and recognize subtle misalignments.
- Ensure there are exactly four options and only one is marked `is_correct: true`.
- Return ONLY the JSON object. Do not explain or add extra text.

{retry_guidance}

Input:

```
{
  "explanation": {explanation_text},
  "level_3_analysis": {level_3_data}
}
```

Figure 22. Prompt for $L_{bridge}$ generation on HVCU-Bench.

## Prompt for $L_{perc}$ Generation

# ROLE
You are an expert AI Benchmark Crafter. You create questions that build on each other to test understanding.

# TASK
You are given the analysis of a Level 3 question and the complete data for a generated Level 2 question. Your task is to create a **very simple** Level 1 (Perception) question that serves as the logical first step in this hierarchy.
The Level 1 question should ask about the **most prominent and easily verifiable element** in the image. The goal is to create a straightforward, entry-level question that almost any observer could answer easily. It should be significantly easier than the Level 2 question. It is the "what" that precedes the "how/why" of Level 2.

# INPUT CONTEXT
The input is a JSON object containing the following keys:
- `explanation`: A text string providing context for the scenario. If an image is part of the input, this explanation is supplementary information. The primary ground truth is defined by the image.
- `level_3_analysis`: A JSON object containing the Level 3 question, its correct answer, and the reasoning behind it. This provides the high-level context.
- `level_2_qa`: A JSON object containing the generated Level 2 question, its options, and the correct answer. Your Level 1 question should be a logical prerequisite for this question.

# OUTPUT FORMAT
Return a single JSON object for the Level 1 question:

```
{
  "level": 1,
  "level_name": "Perception",
  "question": "Your newly generated, simple Level 1 question about
              a prominent and central fact",
  "options": [
    {
      "option_text": "The correct factual answer",
      "is_correct": true
    },
    {
      "option_text": "A plausible but incorrect factual distractor",
      "is_correct": false
    },
    {
      "option_text": "Another plausible but incorrect factual distractor",
      "is_correct": false
    },
    {
      "option_text": "A third plausible but incorrect factual distractor",
      "is_correct": false
    }
  ]
}
```

# IMPORTANT RULES
- The generated question and all options must be derived *only* from the image.
- The question must be about a specific, observable, and **obvious** factual detail.
- The options must be plausible, with one clear correct answer based on the image.
- Ensure there are exactly four options, and only one is marked `is_correct: true`.
- Return ONLY the JSON object. Do not explain or add extra text.

{retry_guidance}

Input:

```
{
  "explanation": {explanation_text},
  "level_3_analysis": {level_3_data},
  "level_2_qa": {level_2_data}
}
```

Figure 23. Prompt for $L_{perc}$ generation on HVCU-Bench.

## Prompt for $L_{bridge} \rightarrow L_{conn}$ Validation

# ROLE
You are a meticulous AI assistant specializing in logical and hierarchical analysis.

# TASK
Evaluate if knowing the answer to "Level 2 QA" provides a foundational building block that helps in reasoning about or answering "Level 3 QA".

**CRITICAL REQUIREMENTS:**
1. **Logical dependency**: Level 2 information must be DIRECTLY useful or necessary for Level 3
2. **Difficulty progression**: Level 2 MUST be more objective, concrete, and simpler than Level 3
3. **Hierarchical coherence**: Level 2 must provide intermediate knowledge that Level 3 builds upon
4. **Complexity standard**: Level 2 should involve basic analysis/interpretation while Level 3 requires deeper reasoning, hidden meanings, or abstract concepts

**VALIDATION STANDARDS:**
- Level 2 questions should involve straightforward analysis or interpretation
- Level 3 questions should require complex reasoning, metaphorical understanding, or deeper insights
- There must be a clear logical connection where Level 2 knowledge helps answer Level 3
- If Level 2 is not noticeably simpler and more concrete than Level 3, validation should FAIL

# REMEMBER
Your primary source of truth is the image.

# QA Pairs
**Level 2 QA** (The foundational knowledge - should be more objective/simple)
- Question: {question_l2}
- Correct Answer: {answer_l2}

**Level 3 QA** (The complex question that should build upon Level 2)
- Question: {question_l3}
- Correct Answer: {answer_l3}

# OUTPUT FORMAT
Respond with ONLY a JSON object with the following structure:

```
{
  "is_helpful": <boolean, true if Level 2 helps with Level 3 AND is significantly simpler, otherwise false>,
  "confidence": <float, your confidence in the "is_helpful" assessment from 0.0 to 1.0>,
  "reasoning": "<string, a brief explanation focusing on logical dependency and difficulty progression.
               If false, explain why Level 2 is not sufficiently simpler or helpful>"
}
```

Figure 24. Prompt for hierarchical validation on HVCU-Bench ($L_{bridge} \rightarrow L_{conn}$).

## Prompt for $L_{perc} \rightarrow L_{bridge}$ Validation

# ROLE
You are a meticulous AI assistant specializing in logical and hierarchical analysis.

# TASK
Evaluate if knowing the answer to "Level 1 QA" provides a foundational building block that helps in reasoning about or answering "Level 2 QA".

**CRITICAL REQUIREMENTS:**
1. **Logical dependency**: Level 1 information must be DIRECTLY useful or necessary for Level 2
2. **Difficulty progression**: Level 1 MUST be significantly more objective, concrete, and simpler than Level 2
3. **Hierarchical coherence**: Level 1 must provide basic, factual knowledge that Level 2 builds upon
4. **Objectivity standard**: Level 1 should focus on observable facts (colors, objects, numbers, basic actions) while Level 2 involves interpretation or analysis

**VALIDATION STANDARDS:**
- Level 1 questions should be answerable by direct observation
- Level 2 questions should require reasoning, interpretation, or analysis
- There must be a clear logical connection where Level 1 knowledge helps answer Level 2
- If Level 1 is not noticeably simpler and more objective than Level 2, validation should FAIL

# REMEMBER
Your primary source of truth is the image.

# QA Pairs
**Level 1 QA** (The foundational knowledge - should be most objective/simple)
- Question: {question_l1}
- Correct Answer: {answer_l1}

**Level 2 QA** (The intermediate complexity question that should build upon Level 1)
- Question: {question_l2}
- Correct Answer: {answer_l2}

# OUTPUT FORMAT
Respond with ONLY a JSON object with the following structure:

```
{
  "is_helpful": <boolean, true if Level 1 helps with Level 2 AND is significantly simpler, otherwise false>,
  "confidence": <float, your confidence in the "is_helpful" assessment from 0.0 to 1.0>,
  "reasoning": "<string, a brief explanation focusing on logical dependency and difficulty progression.
                If false, explain why Level 1 is not sufficiently simpler or helpful>"
}
```

Figure 25. Prompt for hierarchical validation on HVCU-Bench ($L_{perc} \rightarrow L_{bridge}$).

## Prompt for Level-1 Node Generation on the MCTS Tree

Your task is to generate a basic perception question based on the given context.

**Context:**
- Target Level: {target_level} - {level_description}
- {retry_guidance}

**Instructions:**
1. Create a question that focuses on direct visual elements, objects, colors, positions, or basic attributes
2. The question should be foundational and provide a building block for more complex reasoning
3. Ensure the question is different from existing questions in the hierarchy
4. The difficulty should be: {difficulty_guidance}
5. The answer MUST be concise ($\leq$ 30 words).

You must respond with a JSON object in exactly this format:

```
{
  "question": "Your generated question here",
  "answer": "The correct answer",
  "reasoning": "Brief explanation of why this question fits level 1"
}
```

Generate a level 1 basic perception question now.

Figure 26. Prompt for level-1 generation on data generation pipeline.

## Prompt for Level-2 Node Generation on the MCTS Tree

Your task is to generate a connection-level question based on the parent context.

**Parent Context:**
- Parent Question: {parent_question}
- Parent Answer: {parent_answer}
- Target Level: {target_level} - {level_description}
- {retry_guidance}

**Instructions:**
1. Build upon the parent question/answer to create a more complex question
2. Focus on relationships, connections, implications, or broader understanding
3. The question should logically follow from the parent but require additional reasoning
4. The difficulty should be: {difficulty_guidance}
5. The answer MUST be concise ($\leq$ 40 words).
6. Ensure clear logical progression from the parent question

You must respond with a JSON object in exactly this format:

```
{
  "question": "Your generated question here",
  "answer": "The correct answer",
  "reasoning": "Brief explanation of how this connects to the parent and why it fits level 2"
}
```

Generate a level 2 connection question now.

Figure 27. Prompt for level-2 generation on data generation pipeline.

## Prompt for Level-3 Node Generation on the MCTS Tree

Your task is to generate a high-level reasoning question based on the parent context.

**Parent Context:**
- Parent Question: {parent_question}
- Parent Answer: {parent_answer}
- Target Level: {target_level} - {level_description}
- {retry_guidance}

**Instructions:**
1. Build upon the parent question/answer to create a highly complex question
2. Focus on abstract reasoning, inference, analysis, implications, or deep understanding
3. The question should require sophisticated thinking beyond basic observation or connection
4. The difficulty should be: {difficulty_guidance}
5. Ensure the question represents the pinnacle of reasoning complexity for this hierarchy
6. The answer MUST be concise ($\leq$ 50 words).

You must respond with a JSON object in exactly this format:

```
{
  "question": "Your generated question here",
  "answer": "The correct answer",
  "reasoning": "Brief explanation of how this builds on the parent and why it requires high-level reasoning"
}
```

Generate a level 3 high-level reasoning question now.

Figure 28. Prompt for level-3 generation on data generation pipeline.

## Prompt for Evaluation of New Nodes on the MCTS Tree

You are an expert evaluator for hierarchical visual question-answer datasets. Evaluate the quality of the target Q&A pair.

**Context (for reference only):**
- Previous Level {parent_level} Question: {parent_question}
- Previous Level {parent_level} Answer: {parent_answer}

**Target Q&A to Evaluate:**
- Level {child_level} Question: {child_question}
- Level {child_level} Answer: {child_answer}
- Expected Level: {child_level} ({level_description})

**Core Evaluation Criteria (3 key dimensions):**
1. **Logical Coherence**: Is the question internally consistent and does the answer logically follow from the question?
2. **Difficulty Appropriateness**: Does the question match the cognitive demands of Level {child_level}? Does it demonstrate the expected depth of thinking?
3. **Image Alignment**: Do both the question and answer accurately reflect and align with the provided image content?

**Evaluation Rules:**
- Evaluate each dimension on a 0–1.0 scale with precision
- Consider overall quality holistically
- Focus reasoning on identifying current limitations and areas needing improvement
- High scores (0.8+) for excellent Q&A pairs, medium scores (0.5–0.7) for adequate ones, low scores (¡0.5) for problematic ones

You must respond with JSON in exactly this format:

```
{
  "quality_score": 0.0-1.0,
  "reasoning": "Identify specific issues and areas that need improvement (focus on limitations, not strengths)"
}
```

Provide your evaluation now.

Figure 29. Prompt for evaluator on data generation pipeline.