CAN MLLMS DECODE THE CREATIVE LEAP? INTRODUCING C4 FOR CROSS-CONCEPT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Creative capabilities of MLLMs matter in design, communication, education, and human–AI collaboration, yet remain difficult to evaluate because explicit targets and reward signals are scarce compared with accuracy-oriented tasks. Cross-concept understanding is a core cognitive capacity underlying receptive creativity. It enables a perceiver to recover intended meaning from non-obvious but meaningful conceptual relations. We operationalize item construction as cross-concept encoding and model inference as cross-concept decoding. We introduce C4, a cognition-inspired evaluation framework for Chengyu (Chinese idiom)-based Cross-Concept Creativity. Its encoding component maps target slots to imageable substitute concepts along bridge paths in a manually annotated and third-party-reviewed cross-concept network, enabling batch generation with explicit structure, difficulty indexed by bridge count and depth, and exact answers. Using this framework, we instantiate the C4 Evaluation Set (C4-Eval), comprising 184 synthetic items and 37 human-created cross-concept chengyu figures collected from online sources. We manually construct and review cross-concept relations, bridge paths, and reasoning processes for the collected figures. Each C4-Eval item is instantiated in five task settings, yielding 884 primary answer-recovery cases. Across ten evaluated MLLMs, the strongest closed models reach 50.7% and 48.0% primary accuracy, while open-source models remain substantially lower. Candidate constraints improve accuracy sharply, but bridge hints and explanation requests provide only modest gains. These results expose a substantial gap in how current MLLMs decode creatively encoded meaning through cross-concept relations. Release details are provided in the appendix and code and dataset can be found at `https://anonymous.4open.science/r/C4-06BF`.

1 INTRODUCTION

Creative capabilities are increasingly important as multimodal large language models (MLLMs) participate in design, education, communication, and human–AI collaboration (Yin et al., 2024; Zhou & Lee, 2024; Doshi & Hauser, 2024). Evaluating these capabilities, however, remains difficult (Huang et al., 2025; Chakrabarty et al., 2024). Many accuracy-oriented tasks provide explicit targets and direct reward signals, while creative tasks involve novelty, appropriateness, and potentially multiple valid interpretations (Runco & Jaeger, 2012; Huang et al., 2025). These properties make modeling objectives harder to specify and scalable evaluation harder to construct, leaving creative understanding less systematically studied in current multimodal benchmarks (Huang et al., 2025; Chakrabarty et al., 2024). Cognitive psychology provides a principled foundation for evaluating creative understanding. Creativity combines originality with effectiveness (Runco & Jaeger, 2012). Converging cognitive accounts place cross-concept connection at the center of creative cognition. Creative cognition emphasizes generative recombination (Finke et al., 1992), associative theory links creativity to remote conceptual connections (Mednick, 1962), structure mapping formalizes analogical alignment (Gentner, 1983), and conceptual integration explains how meaning emerges from partially shared input spaces (Fauconnier & Turner, 1998). Together, these accounts identify the ability to connect concepts that are not obviously adjacent and traverse the resulting interpretable relation as a core cognitive capacity underlying receptive creativity.

We refer to this capacity as **cross-concept understanding** and to each recoverable link as a cross-concept relation. A red fruit and a physicist share no surface similarity, yet an apple links them

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

through a well-known anecdote. Recovering the apple from such displaced clues illustrates the cognitive operation that supports creative interpretation.

This capability admits a natural two-direction formulation, illustrated in Figure 1. We define **cross-concept encoding** as the substitution process. A communicator starts from a source concept, projects it through meaningful associative paths, and expresses it as displaced but imageable cues. We define **cross-concept decoding** as the recognition direction. An observer sees the displaced cues, hypothesizes the latent relations, and follows them in reverse to recover the source.

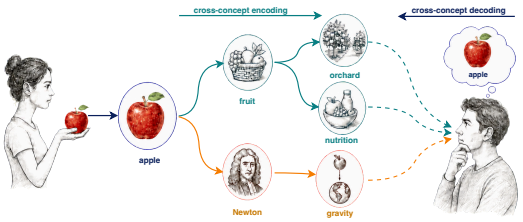


Figure 1: Cross-concept encoding and decoding. A communicator encodes an apple into indirect concepts through distinct associative paths. An observer integrates the resulting clues and decodes the latent source.

Grounding this formulation requires a target space whose answers are stable and whose internal structure supports rich associations. Chinese four-character idioms, or **chengyu**, satisfy both requirements. A chengyu is a conventional expression with a fixed written form, which yields exact answers. Its characters, fragments, pronunciations, referenced objects, and cultural allusions all provide anchor points for associative substitution. This combination preserves culturally grounded creative associations while providing exact targets and inspectable paths. We use chengyu as the carrier and propose C4, a cognition-inspired evaluation framework for Chengyu-based Creative Cross-Concept understanding that makes creativity measurable at scale. The C4 framework organizes benchmark construction and evaluation around a manually annotated chengyu-oriented cross-concept network. Two annotators independently selected concept slots from each target chengyu, associated outward from those slots, and built cross-concept bridges toward imageable landing concepts. A second round combined internal cross-checking between the two annotators with review by a third annotator, and produced the final network. On top of the network, the encoding component fixes a target chengyu, selects slots, follows bridge paths to substitute concepts, and renders the substitutes as visual clues, so item batches carry explicit latent structure, graded difficulty, and exact answers.

Using the C4 framework and the reviewed network, we instantiate a concrete evaluation set, the C4 Evaluation Set (C4-Eval). C4-Eval contains synthetic items at four difficulty levels, L1 through L4, indexed by bridge count and bridge depth, together with human-created figures collected from the web and manually annotated with the same cross-concept structure. The C4 framework supports multiple task forms. Four answer-bearing settings, T1 through T4, constitute the primary evaluation, and a known-answer explanation setting, T5, is excluded from the score and reserved for analysis. We evaluate ten closed and open multimodal models under a strict protocol. The main findings are threefold. First, the task remains far from saturated. The strongest model reaches 50.7% primary accuracy. Second, candidate constraints add 17.3–56.0 points over open recovery, indicating that models often hold enough signal to recognize the answer in a small set but fail to decode it freely. Third, open models under native-context, no-output-cap policy lag behind the strongest closed models, even when documented reasoning wrappers are parsed through an explicit final-answer extraction policy. Together, these results reveal a substantial gap in the ability of current MLLMs to decode creatively encoded meaning.

In summary, this paper makes three contributions:

- We operationalize cross-concept understanding as encoding and decoding and develop the C4 framework around a manually annotated and reviewed chengyu-oriented network with explicit bridge paths and difficulty levels.
- We instantiate the framework as C4-Eval, which contains 184 synthetic items and 37 web-collected figures across 84 targets and 884 primary cases with manually constructed cross-concept networks.
- We evaluate ten MLLMs under a strict protocol, organize the analysis around explicit research questions on decoding bottlenecks and slot and bridge-depth difficulty, and systematically analyze explanation fidelity and shared failures.

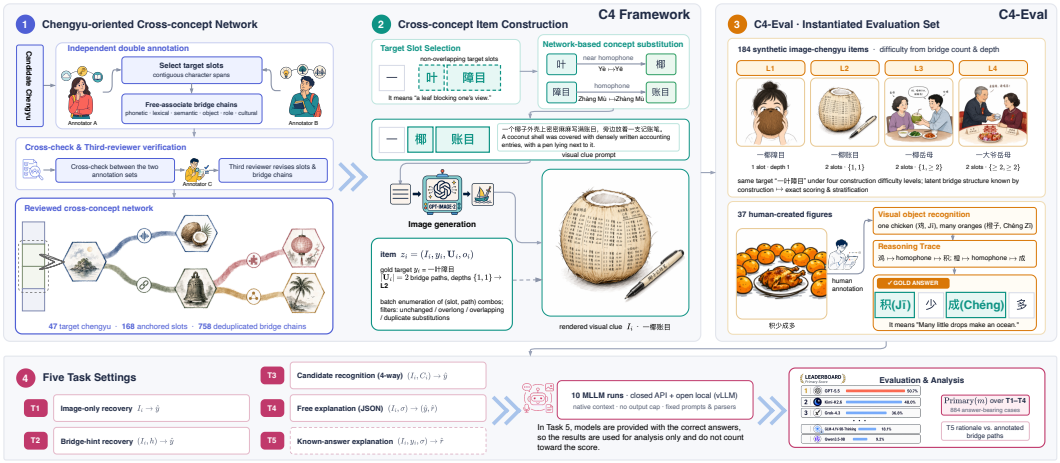


Figure 2: Overview of the C4 framework and C4-Eval. A reviewed chengyu-oriented cross-concept network supports controlled synthetic item construction, while human-created figures receive object, reasoning-trace, and gold-answer annotations. Both sources enter five task settings and a unified MLLM evaluation, where T1–T4 determine the primary score and T5 is reserved for rationale analysis.

2 Related Work

Multimodal Reasoning Benchmarks. Multimodal benchmarks span image-conditioned question answering, controlled and real-image compositional reasoning, external knowledge, grounded language, and image-text compositionality (Antol et al., 2015; Goyal et al., 2017; Johnson et al., 2017; Hudson & Manning, 2019; Marino et al., 2019; Suhr et al., 2019; Thrush et al., 2022). VQA and VQAv2 emphasize image-grounded answering, while CLEVR and GQA probe controlled and real-image compositional structure. MME, MM-Vet, MMMU, and CMMMU broaden coverage of perception, cognition, and domain reasoning across visual formats (Fu et al., 2025; Yu et al., 2024; Yue et al., 2024; Zhang et al., 2024a). REBUS targets indirect visual semantics in image-based wordplay through image recognition, string manipulation, hypothesis testing, and multi-step reasoning (Gritsevskiy et al., 2024). These benchmarks ground answers in visible content, compositional relations, symbolic clues, or domain knowledge. C4 instead targets conventional creative expressions linked to images indirectly through annotated bridge paths. Complementary task views separate open recovery, candidate recognition, and explanation, while fixed targets and explicit paths support exact scoring and analysis by source and difficulty.

Creative Cognition and Conceptual Bridging. Cognitive accounts ground the cross-concept formulation of creative understanding. Creativity is commonly characterized by both novelty and appropriateness (Runco & Jaeger, 2012). Creative cognition studies how retrieval, recombination, imagery, and exploration produce new but interpretable structures (Finke et al., 1992; Beaty et al., 2016). Associative theory links creativity to remote conceptual connections (Mednick, 1962), structure mapping formalizes analogical alignment (Gentner, 1983), and conceptual integration explains how meaning emerges from partial input spaces (Fauconnier & Turner, 1998). Across these accounts, creative interpretation depends on recognizing and traversing relations between concepts that are not obviously adjacent. The C4 framework operationalizes these mechanisms through a cross-concept network, where imageable substitutes and conventional target slots are connected by explicit bridge paths. Bridge count and depth translate associative distance into structured difficulty levels, while the retained paths make each hypothesized relation inspectable.

Creative-Task Benchmarks. Creativity benchmarks cover divergent generation, lateral problem solving, and multimodal creation. Psychometric tests compare model and human creativity, while BRAINTEASER and LatEval assess lateral thinking (Bellemare-Pepin et al., 2026; Jiang et al., 2023; Huang et al., 2024). Generative benchmarks evaluate literary outputs, visual humor, and image-conditioned creative responses. Related studies examine causal interventions and human–AI

effects on novelty and diversity (Chakrabarty et al., 2024; Zhang et al., 2024b; Fang et al., 2025; Huang et al., 2025; Zhou & Lee, 2024; Doshi & Hauser, 2024). These approaches typically rely on diversity proxies, task success, or open-ended judgments. C4-Eval advances receptive creative evaluation by requiring models to recover a fixed conventional target from a creative image and connect visible clues to target slots through annotated bridge paths. Exact recovery supports reproducible scoring, while the cross-concept network distinguishes visual-recognition failures from decoding failures.

3 Benchmark Design

3.1 Cross-Concept Network Construction

The foundation of the C4 framework is a manually annotated chengyu-oriented cross-concept network. We build this network once and reuse it across all synthetic encoding, rather than generating associations separately for each item. Constructing the network ahead of item generation brings two benefits. Every substitute concept in every item traces back to a human-written and reviewed associative path, and the same slot and path inventory supports systematic enumeration of item batches. Annotation proceeds in two rounds. First, two annotators worked independently on each target chengyu. They marked one or more contiguous character spans as candidate concept slots, then associated outward from every slot and wrote one- or multi-step bridge chains toward alternative concepts. A chain may traverse phonetic, lexical, semantic, object, role, part–whole, or culturally grounded associations. Each node in one chain provides a candidate landing concept for replacing the source slot and rendering a visual clue. The two annotators produced separate slot schemes and chain sets for the same targets. The second round combines internal verification with external review. The two annotators first cross-checked each other's slot boundaries and bridges, flagging spans that break morpheme structure and bridges whose steps lack a recoverable relation. A third reviewer then examined both annotation sets, adjudicated the flagged records, and revised slot boundaries and bridges where needed. We parse the reviewed records into exact chengyu-anchored spans, flatten alternative slot schemes, merge duplicate chains by target, effective slot, and node sequence, and retain source provenance for each surviving path. The resulting network contains 47 target chengyu, 168 anchored slots, and 758 deduplicated bridges. Figure 2 places this annotation stage before item-level path selection. The network is a reusable encoding resource.

3.2 Cross-Concept Encoding

Given the reviewed network, the encoding component turns a target chengyu into an image item with known latent structure. The process fixes a target y, selects one or more non-overlapping target slots, and follows one bridge path from each selected slot to an imageable substitute concept. It then composes the substituted phrase by writing each substitute back into its slot position, and prompts an image generator with a scene description that renders the substitutes as visual clues. The image shows the substitutes. The target itself never appears. Enumerating valid combinations of target slots and bridge paths produces item batches whose cross-concept structure is known by construction, because every item records which slots were replaced, which paths were followed, and how deep each path runs. Scene prompts are reviewed to prevent leakage of the gold chengyu, the substituted phrase as a title, or explicit task meta-language. The target stays fixed while slot choices, relation types, and bridge depths vary, which enables exact scoring and stratification by bridge count and depth across batches. The frame-

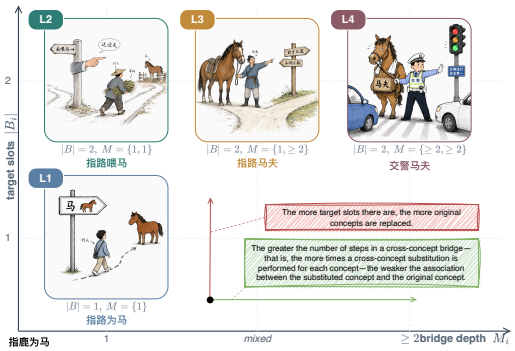


Figure 3: Encoding principles for the four synthetic levels. Each realized example is placed at its own encoding coordinates, with bridge depth on the x-axis and target-slot count on the y-axis. L3 falls between the depth-1 and depth-$\geq$2 columns since its two paths differ. All four points instantiate the same target, *zhǐ lù wéi mǎ* ("calling a deer a horse"), under the four structures.

work expands by adding reviewed target chengyu and bridge paths to the network, and every new path immediately becomes available to the encoder.

Synthetic difficulty follows directly from the encoding parameters. Level 1 uses one slot and one bridge step. Level 2 uses two non-overlapping slots, each with one bridge step. Level 3 uses two non-overlapping slots where one path is one-step and the other runs two or more steps. Level 4 uses two non-overlapping slots where both paths run two or more steps. Figure 3 places a realized example at its encoding coordinates, jointly indexed by bridge depth and target-slot count. The rendered content illustrates each level, while Equation 3 defines it from the annotated path set $\mathcal{B}_i$ and depth multiset M_i. The level scheme provides a structured proxy for cross-concept integration demand. A one-step path should usually be easier to reverse than a multi-step path, and two slots should usually require more integration than one.

3.3 C4-Eval Composition

We instantiate the C4 framework as C4-Eval with synthetic and web-collected sources, summarized in Table 1. The synthetic subset is produced by the encoding component and follows four construction-defined difficulty levels. The web-collected subset consists of human-created cross-concept chengyu figures gathered from online sources. For each figure, we manually reconstruct and review its cross-concept relations, bridge paths, and reasoning process, placing both sources under the same annotation schema. The L1–L4 labels describe synthetic construction parameters. The reviewed bridge paths and reasoning processes characterize the web-collected figures, whose difficulty also reflects visual density, artistic convention, cultural context, and creator-specific encoding choices.

Table 1: Composition and evaluation scale of C4-Eval.

Category	Breakdown	Count
Synthetic items	L1–L4 (47/47/46/44)	184
Web-collected figures	Online sources	37
All items	84 target chengyu	221
Evaluation cases	T1–T5	1,105
Primary cases	T1–T4	884

3.4 Formal Task Definition

Let $\mathcal{Y}$ be the chengyu inventory and let each evaluation item be

$$z_i = (I_i, y_i, \mathcal{B}_i, o_i), \qquad y_i = (c_{i1}, c_{i2}, c_{i3}, c_{i4}) \in \mathcal{Y}. \tag{1}$$

Here I_i is the image, y_i is the gold four-character chengyu, $\mathcal{B}_i$ is the set of annotated bridge paths, and o_i records the source type, web-collected or synthetic. Written in the decoding direction, a bridge path carries a visible substitute concept back to one target slot:

$$b_{ij} : v_{ij0} \to v_{ij1} \to \cdots \to v_{ijm_{ij}} \mapsto c_{is(j)}. \tag{2}$$

The depth m_{ij} counts bridge steps, and $s(j)$ identifies the target slot reached by path j. Encoding traverses these paths from slot to substitute, and a solver must traverse them in reverse. For synthetic items, the difficulty level is defined by bridge count and the multiset of bridge depths $M_i = \{m_{ij} : b_{ij} \in \mathcal{B}_i\}$:

$$D(z_i) = \begin{cases} L1,\ |\mathcal{B}_i| = 1,\ M_i = \{1\}; & L2,\ |\mathcal{B}_i| = 2,\ M_i = \{1, 1\}; \\ L3,\ |\mathcal{B}_i| = 2,\ M_i = \{1, \geq 2\}; & L4,\ |\mathcal{B}_i| = 2,\ M_i = \{\geq 2, \geq 2\}. \end{cases} \tag{3}$$

Equation 3 assigns L1–L4 labels only to synthetic items. The web-collected figures retain their manually constructed bridge paths and reasoning processes without synthetic difficulty labels.

Cross-concept encoding must also leave the item solvable. We impose both conceptual non-identity and bridge recoverability. Let $V(I_i)$ denote the visible concepts a solver can extract from the image. The visible concepts need not literally instantiate the target chengyu, but the reviewed annotations must license a path to it:

$$V(I_i) \not\Rightarrow_{\text{literal}} y_i, \qquad V(I_i) \overset{\mathcal{B}_i}{\Rightarrow} y_i. \tag{4}$$

The first condition prevents direct target illustration from collapsing cross-concept decoding into literal recognition. The second prevents the answer from becoming arbitrary. The evaluation target remains y_i, and the bridge count and depths in $\mathcal{B}_i$ define the synthetic difficulty level. The annotations describe the intended latent relations and are not model inputs except when a task setting supplies a general bridge hint.

3.5 TASK SUITE

Each item is instantiated into five reader-facing task settings. We use T1–T5 in the paper to avoid exposing internal experimental identifiers. All five settings probe the decoding direction under different amounts of support. T1 is image-only recovery, where the model sees the image and must output the chengyu. T2 adds a general bridge hint stating that the image may express a chengyu through homophony, character decomposition, substitution, role association, object association, or related cross-concept relations. The hint names representative relation types and never reveals item-specific bridge paths. T3 gives a four-way candidate set containing the gold chengyu and three distractors drawn from the C4-Eval inventory, and the model must output the chengyu string itself rather than an option letter. T4 asks for a free answer plus structured explanation in JSON. T5 gives the gold chengyu and asks only for a structured explanation. Formally, let h be the bridge-hint text, C_i the candidate set, σ the requested JSON schema, and $\hat{r}$ a free-text or structured rationale. The task views are

$$\begin{aligned} &T_1 : I_i \mapsto \hat{y}, && T_2 : (I_i, h) \mapsto \hat{y}, && T_3 : (I_i, C_i) \mapsto \hat{y}, \\ &T_4 : (I_i, \sigma) \mapsto (\hat{y}, \hat{r}), && T_5 : (I_i, y_i, \sigma) \mapsto \hat{r}. \end{aligned} \tag{5}$$

The five settings separate distinct sources of difficulty. T1 measures unconstrained decoding from the image. T2 tests whether an explicit statement of the bridge phenomenon helps. T3 measures recognition when the answer search space is collapsed to a candidate set. T4 tests whether asking for an explanation changes answer recovery and whether the model can satisfy a structured output format. T5 separates explanation quality from answer search.

4 EXPERIMENTS

We organize the empirical study around four research questions on cross-concept decoding. **RQ1** asks how well current MLLMs decode cross-concept chengyu items overall. **RQ2** asks where the decoding bottleneck lies, in visual cue recognition, in relation inference, or in open answer-space search. **RQ3** asks which encoding parameter contributes more difficulty, the number of target slots or the depth of bridge paths. **RQ4** asks whether models reconstruct the encoded bridge paths faithfully once the answer is known, and what shared failure patterns emerge on items that defeat every model.

4.1 EXPERIMENT SETTINGS

We evaluate ten MLLMs on all 221 items in C4-Eval under the five task settings. The primary comparison uses the 884 answer-bearing T1–T4 cases, while T5 supplies 221 known-answer explanations for targeted analysis. The evaluated models split into two inference types. The API type covers GPT-5.5 (OpenAI, 2025), Kimi-K2.6 (Kimi Team, 2026), Grok-4.3, MIMO-v2.5 (Xiao et al., 2026), and Mistral-Large-3. The Local type covers GLM-4.1V-9B-Thinking (Hong et al., 2025), InternVL3.5-8B (Wang et al., 2025), Gemma3n-E4B-it (Gemma Team et al., 2025), Nemotron-Nano-12B-v2-VL (NVIDIA et al., 2025), and Qwen3.5-9B (Bai et al., 2025). Local models were served with vLLM using NVIDIA A40, L40S, A100, H100 NVL, H200, and RTX PRO 6000 Blackwell GPUs. We report results by task setting, source, and synthetic construction level. T5 is analyzed separately because its prompt contains the target answer.

The current metric is strict exact answer match after lightweight normalization. Let $\nu(\cdot)$ be the normalization function and let $\mathcal{A}_i$ contain the accepted gold strings for item i. Each $\mathcal{A}_i$ contains exactly the canonical chengyu string and alias sets are supported by the scorer but unused. For model m, task t, and predicted answer $\hat{y}_{i,t}^{(m)}$, exact recovery is

$$e_m(i,t) = \mathbf{1}\left[\nu\left(\hat{y}_{i,t}^{(m)}\right) \in \{\nu(a) : a \in \mathcal{A}_i\}\right]. \tag{6}$$

The primary leaderboard score is

$$\mathrm{Primary}(m) = \frac{1}{4N} \sum_{i=1}^{N} \sum_{t=1}^{4} e_m(i,t), \qquad N = 221. \tag{7}$$

T5 is excluded from Equation 7 because its prompt supplies the gold chengyu. For T5, we compare the generated rationale with the annotated visible cues, bridge paths, and target-fragment alignment rather than treating answer recovery as evidence of understanding.

Table 2: Primary accuracy and task-form breakdown (left), and paired outcomes over 221 items (right). "T1 only" and "T4 only" count disagreements between image-only and free-explanation recovery. "Cand. only" counts T3 successes for which all three open settings, T1, T2, and T4, fail.

Model	T1	T2	T3	T4	Primary	T1 only	T4 only	Cand. only
GPT-5.5	35.7	40.3	87.3	39.4	**50.7**	10	18	89
Kimi-K2.6	32.6	39.4	85.5	34.4	48.0	16	20	90
Grok-4.3	23.5	24.4	75.1	24.0	36.8	6	7	99
MIMO-v2.5	16.7	19.0	67.0	21.3	31.0	6	16	90
GLM-4.1V-9B	5.0	7.2	53.4	6.8	18.1	1	5	104
InternVL3.5-8B	5.4	5.0	52.5	8.1	17.8	3	9	96
Gemma3n-E4B-it	2.3	2.3	58.4	2.7	16.4	1	2	122
Mistral-Large-3	2.7	3.6	42.1	5.4	13.5	0	6	83
Nemotron-Nano-12B	1.4	0.9	46.2	1.4	12.4	0	0	99
Qwen3.5-9B	6.3	4.1	22.2	4.1	9.2	7	2	37

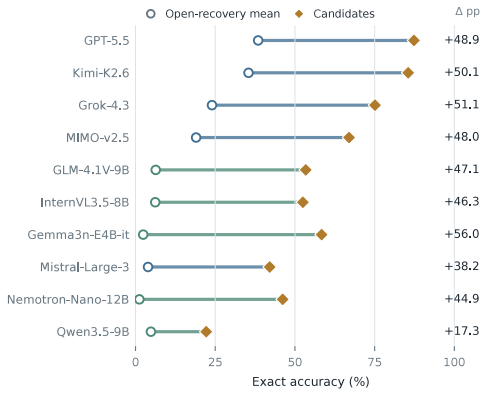
4.2 RQ1: Overall Decoding Performance

RQ1 asks how well current MLLMs decode cross-concept chengyu items. Table 2 shows the completed official leaderboard together with the task-form breakdown. Closed models and open models separate into two clear tiers, and the top score of 50.7% sits far above the best open score of 18.1%. The results show that cross-concept decoding remains far from solved. Even the top model misses roughly half of the primary cases. The two strongest closed models track each other closely in aggregate accuracy, yet they diverge by several points once individual task views are compared, so overall parity can hide task-specific differences. The best open model sits closer to the weakest closed model than to the top of the closed tier. Scale and training recipe therefore dominate the current ranking, yet no recipe brings a model near the ceiling.

4.3 RQ2: Locating the Decoding Bottleneck

RQ2 asks where cross-concept decoding fails. Three stages could be responsible: missing visual cues, failing to infer the latent relations from recognized cues, or failing to search the open answer space once usable relations are available. The task suite separates these stages because T3 removes the search burden while keeping perception and relation inference unchanged. Table 2 also shows the task-form profiles. The candidate constraint produces the most striking effect, lifting accuracy far above the open settings for every evaluated model, including the weakest ones. Figure 4a summarizes this gap. Across the evaluated models, candidate recognition exceeds the mean of T1, T2, and T4 by 17.3 to 56.0 percentage points. The image therefore supplies enough evidence for partial alignment once the target enters a small candidate set, even when that evidence cannot drive unconstrained retrieval. Visual perception is not the limiting stage for most items, since invisible cues could not let a collapsed answer space raise accuracy by tens of points.

The bridge hint isolates the relation-inference stage rather than the search stage. T2 names possible relation types without identifying the relevant cue, substituted concept, or target slot, and it produces much smaller gains than candidates. T4 behaves like the other open settings rather than approaching candidate-level accuracy. Item-level pairing sharpens the picture. The right block of Table 2 pairs T1 and T4 for every item and separately counts candidate-only recovery. Requiring an explanation changes which items are solved rather than producing a uniform shift. Most models recover more items only in T4 than only in T1, though the direction reverses for the weakest model. Explanation prompting can therefore help on some items and hurt on others, even with a fixed image and answer space. Candidate-only recovery quantifies the search bottleneck directly. Roughly half to the majority of the closed models' T3 successes occur on items missed by all three open settings, and the share rises further for most open models, reaching at least three quarters of all T3 successes. The answer to RQ2 follows from these contrasts. Visual cue recognition is largely intact, relation inference is partially available and surfaces once the search space collapses, and open answer-space search is the dominant bottleneck across all evaluated models.



(a) Candidate constraint lift.

Collected	Synthetic bridge level →			
Human	L1	L2	L3	L4
GPT-5.5	42.6	76.6	54.3	39.7	37.5
Kimi-K2.6	40.5	68.6	51.1	39.1	38.1
Grok-4.3	26.4	60.1	43.1	25.5	25.6
MIMO-v2.5	14.9	54.3	29.3	25.0	27.8
GLM-4.1V-9B	9.5	31.4	14.9	13.0	19.9
InternVL3.5-8B	10.1	34.0	13.3	14.7	14.8
Gemma3n-E4B-it	12.8	24.5	13.3	15.2	15.3
Mistral-Large-3	8.8	25.5	10.1	10.3	11.4
Nemotron-Nano-12B	5.4	17.0	12.8	10.3	15.3
Qwen3.5-9B	6.8	22.9	6.4	4.3	4.5

(b) Difficulty profile.

Figure 4: Candidate constraint lift and difficulty profile. (a) Candidate-recognition accuracy minus the mean of the three open recovery settings. The lift is large for most models, showing that answer-space search is a central bottleneck. (b) Accuracy across source and synthetic levels.

4.4 RQ3: Slots versus Bridge Depth

RQ3 asks which encoding parameter contributes more difficulty. The level design permits a direct comparison because adjacent levels change exactly one parameter. Moving from L1 to L2 adds a second slot while keeping every path one-step. Moving from L2 to L3 deepens one path while keeping two slots. Moving from L3 to L4 deepens the second path. Figure 4b shows performance by source and synthetic level. Accuracy falls sharply from L1 to L2, drops more moderately from L2 to L3, and shows no further decline from L3 to L4, reversing slightly for several models.

The answer to RQ3 is that slot count dominates. Averaged across all ten models, the L1-to-L2 slot transition produces roughly three times the mean accuracy drop of the L2-to-L3 depth transition, and additional depth beyond the first multi-step path yields no further measurable difficulty. The level buckets contain different targets and rendered images, so these adjacent comparisons describe associations with the annotated encoding rather than controlled causal effects. The web-collected subset stays difficult for all models, scoring below the synthetic average for every model. Human creators combine dense composition, cultural context, and free-form encoding choices, so this subset checks that the synthetic difficulty scheme does not exhaust the phenomenon.

4.5 RQ4: Explanation Fidelity and Shared Failures

RQ4 asks whether models reconstruct the encoded bridge paths faithfully and what items defeat every model. We address the first part with a paired T4–T5 case study and the second with a corpus-level bad-case analysis.

Known-answer explanation. T5 isolates explanation behavior after answer retrieval has been removed. We analyze each explanation in two stages. First, we compare the reported perceptual cues with the annotated visual elements. Second, we compare the proposed cue-to-target links with the gold bridge paths and their aligned target fragments. T4 on the same image provides an independent cross-check because both its answer and rationale are generated without access to the target. Figure 5 aligns the shared visual evidence with the annotated L3 paths and the paired model diagnoses. For GPT-5.5, the paired outputs agree at the perceptual layer but diverge at the bridge layer. In visual recognition, T4 identifies the labeled strawberry, the labeled monk, and the soldier helmets, and T5 reports the same three elements, ruling out object recognition as the bottleneck here. In bridge linking, T4 correctly predicts the target and explicitly reconstructs both annotated paths, including the culturally mediated association from monk to wooden fish and then to wood. T5 instead maps the strawberry directly to "grass and wood," omits the monk path, and explains the scene through the conventional meaning that vegetation is being treated as soldiers. The explanation stays fluent but does not recover the encoding that generated the image. The correct T4 prediction proves that the image contains sufficient evidence for both target recovery and the intended paths, so

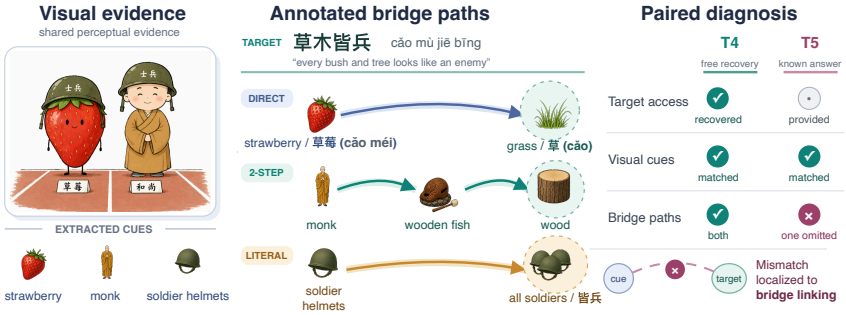


Figure 5: Paired T4–T5 diagnosis for *cǎo mù jiē bīng* ("every bush and tree looks like an enemy"). Both settings recover the same visual cues, while only T4 reconstructs both annotated bridge paths.

the T5 mismatch arises at the decoding step that links a recognized entity through an intermediate cultural concept to its slot. A correct answer field or valid JSON alone does not establish explanatory faithfulness.

Bad cases shared across models. We examine items that resist every model. Across the ten evaluated models, 83 of the 221 items are never solved by any model in any of the three open settings, and 6 items remain unsolved even when candidates are provided. The open-unsolved items concentrate in the higher synthetic levels, matching the RQ3 finding that two-slot encoding drives shared failure. One fully unsolved item makes the shared failure mode concrete. The L3 item encodes *yī yè zhàng mù* ("a leaf blocks one's view") as the phrase "one coconut mother-in-law," following a one-step near-homophone path from leaf to coconut and a two-step path from block-the-eye through a homophone to mother-in-law. The rendered image shows a mother-in-law at a dinner table with a coconut and straw at hand. Model outputs reveal decoding anchored on the surface. GPT-5.5's structured rationale correctly lists the mother-in-law label, the coconut, and the straw, then assembles them literally into *yuè mǔ cì zì* ("General Yue's mother tattoos characters on his back"), treating the visible concepts as target fragments instead of reversing their bridges. Under candidates, every strong model selects the distractor *yè gōng hào lóng* ("Lord Ye loves loongs"), which shares the leaf character with the gold target, rather than the target itself. All ten models solve none of the four answer-bearing tasks on this item, yet the strongest models reconstruct both encoded paths in T5 once the answer is given. The failure is therefore a decoding failure rather than a perceptual or knowledge gap, since visible substitutes are consumed literally instead of being traversed backward through their encoded relations.

5 Conclusion

We introduce C4, a cognition-inspired framework that makes receptive creative understanding measurable through cross-concept encoding and decoding. We instantiate the framework as C4-Eval, combining synthetic items with human-created figures collected from online sources under a shared annotation schema. C4 builds on a manually annotated chengyu-oriented cross-concept network and integrates structured item construction, task instantiation, answer extraction, exact scoring, and stratified analysis. Across ten MLLMs, even the strongest model recovers only about half of the primary cases, leaving the leaderboard far from saturated. Candidate constraints produce large gains, while open models remain far behind the strongest closed systems. Further analysis identifies open answer-space search as the dominant bottleneck, shows that slot count influences difficulty more strongly than bridge depth, and finds that models often consume visible substitutes literally rather than decode their cross-concept relations. These findings demonstrate substantial headroom in cross-concept understanding and establish C4 as a structured, extensible foundation for evaluating how MLLMs decode creatively encoded meaning across concepts.

AI USE STATEMENT

Generative AI enters this work in the following roles. *Creating synthetic datasets.* Human annotators constructed and reviewed the cross-concept network, including target slots, bridge paths, and landing concepts. For each synthetic item, an LLM generated a scene description from the selected network nodes, and every scene description was manually reviewed and revised before rendering. The 184 synthetic images were then rendered from these descriptions by GPT Image 2 (Section 3.2 and Appendix D). Target chengyu, bridge paths, substituted phrases, difficulty labels, and gold answers are determined by the human-annotated network rather than by the generators. *Supporting data annotation.* The structured explanation annotations of visible cues, bridge units, and target-fragment alignment were written manually for the 37 web-collected figures. For the 184 synthetic items, an LLM drafted these annotations from the reviewed scene descriptions and the annotated bridge chains, and the drafts were manually checked and corrected. *Object of study.* The ten evaluated MLLMs are the subject of the experiments. Their outputs constitute the experimental results and are scored by deterministic exact match (Equation 6), and no LLM is used as a judge for any reported score. *Editing.* We used LLMs to polish the wording of the manuscript and an LLM-based coding assistant (Claude Code) to adapt its LaTeX layout to the ICLR template without changing the scientific content. We did not use generative AI for translation. We have reviewed all AI-assisted work. One of the authors checked every LLM-generated scene description and synthetic explanation annotation against the annotated bridge paths and reviewed all AI-edited text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 specifies the construction of the cross-concept network, the encoding procedure, and the synthetic difficulty levels, and Equations 1–5 formally define items, bridge paths, difficulty levels, the recoverability conditions, and the five task views. Section 4.1 and Equations 6–7 define answer normalization, exact-match scoring, and the primary score. Appendix C documents the two-round annotation and review protocol and the relation families used for bridge annotation. Appendix D describes slot and path enumeration, filtering, the image generator and resolution, and the deterministic item selection. Appendix E gives the task settings, distractor selection, normalization, and final-answer extraction rules. Appendix F lists the interface, context length, hardware, decoding setting, and software versions for every evaluated model, and Appendix G reports correct-answer counts for every model and task so that all reported percentages can be recomputed. As described in Appendix B, the evaluation set, reviewed annotations, synthetic images, task instances, original and scored model outputs, and construction and evaluation code will be released upon publication. An anonymized version of the construction and evaluation code, annotations, and scene metadata is provided in the supplementary material and at `https://anonymous.4open.science/r/C4-06BF`. The release of the web-collected figures will respect the rights of their original creators, and item-level predictions for all 221 items allow every reported score to be recomputed. Each reported result corresponds to one complete evaluation pass. Results for API models depend on provider-managed serving and may change as endpoints are updated, and GPT-5.5 does not expose a temperature parameter.

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. doi: 10.1109/ICCV.2015.279.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631.

Roger E. Beaty, Mathias Benedek, Paul J. Silvia, and Daniel L. Schacter. Creative cognition and brain network dynamics. *Trends in Cognitive Sciences*, 20(2):87–95, 2016. doi: 10.1016/j.tics.2015.10.004.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1), 2026. doi: 10.1038/s41598-025-25157-3.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024. doi: 10.1145/3613904.3642731.

Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290.

Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, Kai Chen, and Dahua Lin. Creation-MMBench: Assessing context-aware creative intelligence in MLLM. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 447–456, 2025. doi: 10.1109/ICCV51701.2025.00049.

Gilles Fauconnier and Mark Turner. Conceptual integration networks. *Cognitive Science*, 22(2): 133–187, 1998. doi: 10.1207/s15516709cog2202_1.

Ronald A. Finke, Thomas B. Ward, and Steven M. Smith. *Creative Cognition: Theory, Research, and Applications*. The MIT Press, 1992. ISBN 9780262272650. doi: 10.7551/mitpress/7722.001.0001.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2025. doi: 10.48550/arXiv.2306.13394.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170, 1983. doi: 10.1016/S0364-0213(83)80009-3.

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. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Andrew Gritsevskiy, Arjun Panickssery, Aaron Kirtland, Derik Kauffman, Hans Gundlach, Irina Gritsevskaya, Joe Cavanagh, Jonathan Chiang, Lydia La Roux, and Michelle Hung. REBUS: A robust evaluation benchmark of understanding symbols. *arXiv preprint arXiv:2401.05604*, 2024. doi: 10.48550/arXiv.2401.05604.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. doi: 10.48550/arXiv.2507.01006.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. LatEval: An interactive LLMs evaluation benchmark with incomplete information from lateral thinking puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 10186–10197, 2024. doi: 10.63317/44hwsvdgxk8q.

Zhongzhan Huang, Shanshan Zhong, Pan Zhou, Shanghua Gao, Marinka Zitnik, and Liang Lin. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3830–3846, 2025. doi: 10.1109/TPAMI.2025.3539433.

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. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14317–14332, 2023. doi: 10.18653/v1/2023.emnlp-main.885.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.

Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026. doi: 10.48550/arXiv.2602.02276.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3190–3199, 2019. doi: 10.1109/CVPR.2019.00331.

Sarnoff Mednick. The associative basis of the creative process. *Psychological Review*, 69(3):220–232, 1962. doi: 10.1037/h0048850.

NVIDIA, Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, et al. NVIDIA Nemotron Nano V2 VL. *arXiv preprint arXiv:2511.03929*, 2025. doi: 10.48550/arXiv.2511.03929.

OpenAI. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*, 2025. doi: 10.48550/arXiv.2601.03267.

Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, 2019. doi: 10.18653/v1/P19-1644.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5228–5238, 2022. doi: 10.1109/CVPR52688.2022.00517.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. doi: 10.48550/arXiv.2508.18265.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, et al. MiMo-V2-Flash technical report. *arXiv preprint arXiv:2601.02780*, 2026. doi: 10.48550/arXiv.2601.02780.

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. doi: 10.1093/nsr/nwae403.

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. In *Proceedings of the International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.48550/arXiv.2311.16502. Oral presentation; arXiv:2311.16502.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, and Jie Fu. CMMMU: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*, 2024a. doi: 10.48550/arXiv.2401.11944.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, Robert Mankoff, and Robert Nowak. Humor in AI: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125264–125286, 2024b. doi: 10.52202/079017-3978.

Eric Zhou and Dokyun Lee. Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, 3(3):pgae052, 2024. doi: 10.1093/pnasnexus/pgae052.

A Theoretical Scope and Operational Assumptions

The theoretical contribution of C4 is a cognition-grounded formulation of receptive creative understanding for MLLMs. It connects three levels of description. Cognitive theories motivate cross-concept connection as a core mechanism of creative interpretation. The encoding and decoding formulation identifies the two directions of the process. The cross-concept network then makes the formulation operational through target slots, bridge paths, landing concepts, and recoverable answers.

The evaluation rests on four explicit assumptions. First, an item has a conventional target whose written form is fixed. Second, the visible concepts differ from a literal depiction of that target. Third, every intended substitution has a reviewed bridge path back to a target slot. Fourth, the annotated path records the intended construction but does not exclude other plausible interpretations of the same image. Exact recovery measures whether a model reaches the intended conventional target. The L1–L4 levels describe construction parameters. They do not assert an intrinsic psychological scale independent of targets and images.

The formal objects in the main paper define an item, a bridge path, the four synthetic levels, the recoverability condition, the five task views, and the primary score. The experiments test the resulting predictions at the benchmark level. Candidate constraints isolate answer-space search, paired T1 and T4 outcomes measure prompt-form sensitivity, level comparisons separate slot count from bridge depth, and T5 enables direct inspection of visual recognition and bridge reconstruction.

B Data and Resource Release

Upon publication, we will release the reviewed cross-concept annotations, the 184 synthetic images, task instances, model outputs used in the paper, and the code required to construct and evaluate the benchmark. Table S1 summarizes these resources. The manual annotations of the 37 web-collected figures, including visible cues, bridge units, target-fragment alignment, and reviewed rationales, will be released in full, and the figures themselves will be distributed in a form that respects the rights of their original creators. We will license the code under Apache-2.0 and the data, annotations, prompts, and benchmark metadata under CC BY-NC 4.0. An anonymized version of the code and construction metadata is available for review at `https://anonymous.4open.science/r/C4-06BF`.

Table S1: Resources included in the publication release.

Resource	Contents	Reproducibility role
Cross-concept annotations	Reviewed target slots, bridge chains, landing concepts, relation types, and annotation provenance	Supports inspection and reuse of the encoding resource
C4-Eval items	Images or image descriptions, canonical answers, source labels, difficulty levels, and aligned bridge annotations	Reconstructs the evaluated item collection
Task instances	T1–T5 prompts, candidate sets, expected answers, and explanation annotations	Reproduces the five evaluation settings
Model predictions	Original responses and scored predictions for every reported model	Recomputes task scores and stratified analyses
Construction and evaluation code	Network processing, item construction, task instantiation, scoring, and analysis procedures	Reproduces benchmark construction and reported results

B.1 Data Card

C4-Eval contains 221 items covering 84 target chengyu. The synthetic subset contains 47 L1, 47 L2, 46 L3, and 44 L4 items. The collected subset contains 37 human-created figures. Each item has five task views. This produces 1,105 evaluation cases, of which 884 answer-bearing T1–T4 cases determine the primary score.

Each item includes a source type, image, canonical answer, task prompt, and accepted-answer list. Synthetic items additionally include the selected target slots, complete bridge chains, landing concepts, substituted phrase, generation prompt, and L1–L4 label. Collected items include visible cues, bridge units, target-fragment alignment, and a reviewed rationale. The benchmark contains no personal attributes or private user records.

C Cross-Concept Network Annotation

Two annotators independently processed each of the 47 synthetic target chengyu. Each annotator selected one or more contiguous character spans that could function as concept slots. They then associated outward from each slot and wrote one-step or multi-step chains that ended at an imageable concept. The two annotation sets were retained separately during the first round.

In the second round, the annotators cross-checked slot boundaries and bridge steps. They flagged spans that broke the intended lexical unit and steps that lacked a recoverable relation. A third reviewer examined both annotation sets, adjudicated flagged records, and revised slots and chains. The reviewed annotations were normalized to exact character spans. Alternative slot schemes were retained, duplicate chains were merged by target, effective slot, and node sequence, and source provenance was preserved. The final resource contains 47 targets, 168 anchored slots, and 758 deduplicated bridge chains.

Table S2: Relation families used during bridge annotation. A chain may combine several families.

Relation	Annotation criterion
Phonetic	Homophony or near homophony links two spoken forms
Lexical	A character, word, or conventional expression licenses the next node
Semantic	The next node is connected by meaning, category, or functional similarity
Object	An object conventionally evokes another object or its use
Role	An agent, profession, or social role evokes a related concept
Part–whole	A salient component links to an object, event, or category
Cultural	Shared stories, customs, named entities, or conventional knowledge license the step

Every retained chain must satisfy slot alignment, conceptual non-identity, stepwise recoverability, and imageability of its landing concept. Slot alignment requires the chain to terminate at the annotated target span. Conceptual non-identity prevents a landing concept from simply repeating the target. Stepwise recoverability requires each adjacent pair to have an interpretable relation. Imageability requires the final node to support a concrete visual scene.

D Synthetic Construction and Quality Control

The encoder enumerates non-overlapping target slots and reviewed bridge chains. L1 selects one one-step chain. L2 selects two one-step chains. L3 selects one one-step chain and one chain of depth two or greater. L4 selects two chains of depth two or greater. The build retains at most one item for each target and level. It removes unchanged substitutions, overlapping slots, malformed phrases, and duplicate items. This process yields 184 items from 188 possible target-level cells.

Scene descriptions were generated by an LLM from the selected network nodes, and all of them were manually reviewed. Review removed direct appearances of the gold chengyu, titles that repeated the substituted phrase, and explicit task instructions. The final 184 images were generated with GPT Image 2 at a resolution of 1024 by 1024 pixels. The release includes each final image together with its target, level, substituted phrase, scene description, prompt, and gold bridge paths. Item selection, candidate construction, and task instantiation are deterministic once the reviewed annotations are fixed. Gold explanation annotations for synthetic items were drafted by an LLM from the reviewed scene description and the annotated bridge chains and then manually checked, whereas those for collected figures were written manually.

E TASK INSTANTIATION AND SCORING

Table S3 summarizes the five task settings. All task prompts are written in Chinese to match the language of the target expressions. T3 distractors are selected from the 84-answer inventory. Candidates that share more Chinese characters with the gold target are ranked first, and ties are resolved deterministically.

Table S3: Required outputs and supplied information for the five task settings.

Task	Required output	Information supplied
T1	Chengyu	Image only
T2	Chengyu	Image and a general cross-concept hint
T3	Chengyu	Image and four candidate chengyu
T4	Chengyu and rationale	Image and the structured explanation schema
T5	Rationale	Image, gold chengyu, and the structured explanation schema

T1–T3 score one predicted chengyu, while T4 scores the answer supplied with the structured rationale. Normalization removes whitespace and common ASCII and Chinese punctuation. The accepted-answer set contains only the canonical four-character target. T5 is excluded from primary accuracy because its prompt supplies the target.

GLM-4.1V-9B-Thinking emits a visible reasoning segment followed by an explicit final-answer payload, which supplies its scored prediction. Qwen3.5-9B is scored only when it produces a final answer. Responses that terminate before that answer remain incorrect. The publication release includes the original responses and final scored predictions.

T5 explanations are analyzed in two stages. The first stage compares reported perceptual cues with the annotated visual elements. The second compares cue-to-target links with the annotated bridge chains and target fragments. T4 on the same item provides an answer-hidden cross-check.

F EXPERIMENTAL CONFIGURATION

Each official row is one complete pass over all 1,105 task cases. Temperature is set to zero when the endpoint exposes that parameter. No output-token cap is applied. API models use their provider-managed serving context and documented default interaction mode. Local inference runs on Linux with Python 3.11.11, CUDA 13.0.2, and vLLM 0.23.0. Each local process uses one listed NVIDIA accelerator with 40 to 141 GB of GPU memory.

Table S4: Final inference conditions. API hardware is managed by the corresponding provider.

Model	Interface	Context	Hardware	Final decoding and scoring setting
GPT-5.5	Provider API	Provider managed	Provider managed	Temperature omitted, no output cap, exact scoring
Kimi-K2.6	Provider API	Provider managed	Provider managed	Temperature 0, no output cap, exact scoring
Grok-4.3	Provider API	Provider managed	Provider managed	Temperature 0, no output cap, exact scoring
MIMO-v2.5	Provider API	Provider managed	Provider managed	Temperature 0, no output cap, exact scoring
Mistral-Large-3	Provider API	Provider managed	Provider managed	Temperature 0, no output cap, exact scoring
GLM-4.1V-9B-Thinking	vLLM	65,536	H100 NVL	Temperature 0, no output cap, final-answer extraction
InternVL3.5-8B	vLLM	8,192	A100	Temperature 0, no output cap, native chat behavior
Gemma3n-E4B-it	vLLM	32,768	A100	Temperature 0, no output cap, instruction-tuned image-first template
Nemotron-Nano-12B-v2-VL	vLLM	8,192	L40S	Temperature 0, no output cap, native chat behavior
Qwen3.5-9B	vLLM	262,144	A40, L40S, A100, H100 NVL, H200, RTX PRO 6000 Blackwell	Temperature 0, no output cap, strict final-answer scoring

The task prompts, candidate sets, normalization, and scoring rules are fixed across models. Each reported result is computed from one complete evaluation pass.

G Extended Results

Table S5 reports the task-level numerators used to compute the main leaderboard. Every task has 221 cases. Primary accuracy divides the sum of T1–T4 correct counts by 884.

Table S5: Correct-answer counts. T5 reports exact recovery of its supplied answer field and is not included in Primary.

Model	T1	T2	T3	T4	Primary	T5
GPT-5.5	79	89	193	87	448	221
Kimi-K2.6	72	87	189	76	424	221
Grok-4.3	52	54	166	53	325	221
MIMO-v2.5	37	42	148	47	274	216
GLM-4.1V-9B-Thinking	11	16	118	15	160	221
InternVL3.5-8B	12	11	116	18	157	221
Gemma3n-E4B-it	5	5	129	6	145	221
Mistral-Large-3	6	8	93	12	119	221
Nemotron-Nano-12B-v2-VL	3	2	102	3	110	218
Qwen3.5-9B	14	9	49	9	81	152

The source and level breakdowns use the same four answer-bearing tasks. The 37 collected items contribute 148 primary cases. L1 and L2 each contribute 188 cases, L3 contributes 184, and L4 contributes 176. The publication release includes item-level predictions, allowing every percentage and paired count in the paper to be recomputed.

G.1 Additional Construction Cases

The synthetic target *zhǐ lù wéi mǎ* (“call a deer a horse”) illustrates how one target supports all four construction levels. L1 replaces *zhǐ lù* (“point at a deer”) with its homophone meaning “give directions.” L2 adds the one-step homophonic path from “be a horse” to “feed a horse.” L3 extends the second path from feeding a horse to a horse keeper. L4 also extends the first path from giving directions to a traffic officer. The four images therefore vary slot count and bridge depth while keeping the target fixed.

The collected case *máo sè dùn kāi* (“become suddenly enlightened”) uses two culturally grounded paths. A blocked toilet evokes the target fragment *máo sè* through its spoken description. Newton opening the blockage evokes *dùn kāi* through Newton’s transliterated Chinese name and the opening action. The annotation separates recognition of the toilet, Newton, and the opening action from reconstruction of the two target-fragment links.

H Chengyu Reference

Tables S6–S11 list every target in C4-Eval. Pinyin is italicized. The English text gives the conventional meaning used for reviewer orientation.

Table S6: C4-Eval chengyu reference, part 1 of 6.

Chengyu	Pinyin	Conventional meaning
一叶障目	*yī yè zhàng mù*	A single leaf blocks one's view, describing a narrow or obscured perspective
一网打尽	*yī wǎng dǎ jǐn*	Catch or eliminate everything in one sweep
上纲上线	*shàng gāng shàng xiàn*	Elevate a minor issue into a matter of principle or ideology
东施效颦	*dōng shī xiào pín*	Imitate another person blindly and produce the opposite effect
井底之蛙	*jǐng dǐ zhī wā*	A frog at the bottom of a well, describing a person with a narrow outlook
亡羊补牢	*wáng yáng bǔ láo*	Mend the pen after losing sheep, meaning that correction is still worthwhile
交浅言深	*jiāo qiǎn yán shēn*	Speak with undue intimacy despite a shallow acquaintance
以德服人	*yǐ dé fú rén*	Win people over through virtue
倾家荡产	*qīng jiā dàng chǎn*	Lose all family property and resources
凿壁偷光	*záo bì tōu guāng*	Study diligently under hardship, literally by borrowing light through a wall
分庭抗礼	*fēn tíng kàng lǐ*	Meet or contend with another party on equal terms
分道扬镳	*fēn dào yáng biāo*	Go separate ways
别开生面	*bié kāi shēng miàn*	Create a fresh style or situation
勾心斗角	*gōu xīn dòu jiǎo*	Scheme and struggle against one another

Table S7: C4-Eval chengyu reference, part 2 of 6.

Chengyu	Pinyin	Conventional meaning
半路出家	*bàn lù chū jiā*	Enter a profession or field relatively late
单刀赴会	*dān dāo fù huì*	Attend a dangerous meeting alone and with courage
南辕北辙	*nán yuán běi zhé*	Act in a direction contrary to one's stated goal
卧薪尝胆	*wò xīn cháng dǎn*	Endure hardship while preparing to achieve a long-term goal
口蜜腹剑	*kǒu mì fù jiàn*	Speak sweetly while harboring malicious intent
叶公好龙	*yè gōng hào lóng*	Profess a love that disappears when confronted with the real thing
叹为观止	*tàn wèi guān zhǐ*	Acclaim something as reaching the height of excellence
各有千秋	*gè yǒu qiān qiū*	Each has distinctive merits
后患无穷	*hòu huàn wú qióng*	Leave endless trouble for the future
命中注定	*mìng zhōng zhù dìng*	Be predestined
唇亡齿寒	*chún wáng chǐ hán*	Two parties are so interdependent that harm to one endangers the other
喜形于色	*xǐ xíng yú sè*	Show joy visibly on one's face
多管闲事	*duō guǎn xián shì*	Meddle in matters that do not concern oneself
大可不必	*dà kě bù bì*	Be entirely unnecessary

Table S8: C4-Eval chengyu reference, part 3 of 6.

Chengyu	Pinyin	Conventional meaning
天罗地网	*tiān luó dì wǎng*	An inescapable net spread everywhere
头头是道	*tóu tóu shì dào*	Speak coherently and with apparent reason
奴颜婢膝	*nú yán bì xī*	Behave in a servile and obsequious manner
孜孜不倦	*zī zī bù juàn*	Work diligently and tirelessly
守株待兔	*shǒu zhū dài tù*	Wait idly for an unlikely opportunity
对牛弹琴	*duì niú tán qín*	Address an unreceptive or inappropriate audience
对酒当歌	*duì jiǔ dāng gē*	Sing over wine and reflect on the brevity of life
小鸟依人	*xiǎo niǎo yī rén*	Appear delicate, affectionate, and dependent
平易近人	*píng yì jìn rén*	Be amiable and approachable
得意忘形	*dé yì wàng xíng*	Become carried away by success
德才兼备	*dé cái jiān bèi*	Possess both virtue and ability
心服口服	*xīn fú kǒu fú*	Be sincerely and completely convinced
怒不可遏	*nù bù kě è*	Be unable to restrain one's anger
惊弓之鸟	*jīng gōng zhī niǎo*	Remain highly fearful after a previous shock

Table S9: C4-Eval chengyu reference, part 4 of 6.

Chengyu	Pinyin	Conventional meaning
愚昧无知	*yú mèi wú zhī*	Be ignorant and unenlightened
拖泥带水	*tuō ní dài shuǐ*	Handle something in a sloppy, indirect, or protracted way
拱手相让	*gǒng shǒu xiāng ràng*	Yield something willingly to another person
指鹿为马	*zhǐ lù wèi mǎ*	Deliberately call a deer a horse and invert the truth
推心置腹	*tuī xīn zhì fù*	Treat another person with complete sincerity
掩耳盗铃	*yǎn ěr dào líng*	Deceive oneself while ignoring an obvious reality
握手言欢	*wò shǒu yán huān*	Shake hands and reconcile
断章取义	*duàn zhāng qǔ yì*	Quote or interpret something out of context
有勇无谋	*yǒu yǒng wú móu*	Be brave but lack strategy
朝三暮四	*cháo sān mù sì*	Be capricious or repeatedly change one's position
杯弓蛇影	*bēi gōng shé yǐng*	Develop unfounded fear by mistaking a reflection for a snake
欺世盗名	*qī shì dào míng*	Deceive the public to obtain an undeserved reputation
正人君子	*zhèng rén jūn zǐ*	An upright and honorable person
死有余辜	*sǐ yǒu yú gū*	Commit crimes so serious that even death cannot atone for them

Table S10: C4-Eval chengyu reference, part 5 of 6.

Chengyu	Pinyin	Conventional meaning
毛遂自荐	*máo suì zì jiàn*	Volunteer oneself for a task
泥古不化	*nì gǔ bù huà*	Adhere rigidly to old conventions
滥竽充数	*làn yú chōng shù*	Pass oneself off as qualified by hiding among experts
狐假虎威	*hú jiǎ hǔ wēi*	Borrow another's power to intimidate others
狗仗人势	*gǒu zhàng rén shì*	Bully others by relying on powerful backing
画饼充饥	*huà bǐng chōng jī*	Use empty promises or fantasies to satisfy a real need
积少成多	*jī shǎo chéng duō*	Accumulate small amounts into a large total
程门立雪	*chéng mén lì xuě*	Show deep respect for a teacher and sincere dedication to learning
空前绝后	*kōng qián jué hòu*	Be unprecedented and unlikely to be repeated
素不相识	*sù bù xiāng shí*	Have never been acquainted
纹丝不动	*wén sī bù dòng*	Remain completely motionless
结草衔环	*jié cǎo xián huán*	Repay kindness with lasting gratitude
罄竹难书	*qìng zhú nán shū*	Have more crimes or faults than can be recorded
老态龙钟	*lǎo tài lóng zhōng*	Appear aged and physically infirm

Table S11: C4-Eval chengyu reference, part 6 of 6.

Chengyu	Pinyin	Conventional meaning
背道而驰	*bèi dào ér chí*	Act in direct opposition to the proper direction or principle
自相矛盾	*zì xiāng máo dùn*	Contradict oneself
舍本逐末	*shě běn zhú mò*	Neglect essentials while pursuing minor details
茅塞顿开	*máo sè dùn kāi*	Become suddenly enlightened
草木皆兵	*cǎo mù jiē bīng*	See every bush and tree as an enemy because of extreme fear
趋炎附势	*qū yán fù shì*	Curry favor with powerful people
趾高气扬	*zhǐ gāo qì yáng*	Behave arrogantly and overbearingly
路人皆知	*lù rén jiē zhī*	Be known to everyone
门可罗雀	*mén kě luó què*	Receive so few visitors that birds could gather at the gate
门庭若市	*mén tíng ruò shì*	Be crowded with visitors like a marketplace
闻鸡起舞	*wén jī qǐ wǔ*	Rise at cockcrow to practice with diligence
飞扬跋扈	*fēi yáng bá hù*	Behave in a domineering and unrestrained manner
鱼目混珠	*yú mù hùn zhū*	Pass off an inferior object as something valuable
黑白混淆	*hēi bái hùn xiáo*	Confuse right and wrong