Drawing Conclusions: Representation or Reasoning in New Yorker Caption Matching

Matthew T. Bouchard
Stanford University
mattb415@stanford.edu

Abstract

Understanding why a New Yorker cartoon caption fits an image is a demanding test of multimodal intelligence: a model must not only see the cartoon but also get the joke. In the contest's five-way matching task, the input is a cartoon, a human-written scene description, and five candidate captions, and the output is the caption that belongs with the cartoon. The best published model reaches only about 62% accuracy on this task, while expert humans reach about 94%.

We ask whether that gap can be closed by stronger representations alone, or whether it reflects a need for candidate-level reasoning. Across CLIP baselines, fine-tuning, larger embedding models, and gated cross-attention that enriches image and caption embeddings with the scene description, no representation-focused method reliably breaks the 61–62% band, and a knowledge-rich encoder reaches that ceiling with no training at all. Frontier vision–language models approach human performance, but they change scale, encoder, scoring, and reasoning at once, so we treat them as an upper-bound probe rather than a controlled ablation; our controlled reasoning-style interventions help weak encoders but quickly saturate. Caption matching, in short, is not just about embedding cartoons and captions closer together, but about reasoning over the scene, the incongruity, and why one caption fits best—getting machines in on the joke will take more than better representations.

1. Introduction

Understanding why a cartoon is funny is a demanding test of multimodal intelligence. It asks a model not merely to recognize what is drawn, but to connect the scene to an unstated, often absurd premise that makes a caption land. The New Yorker Caption Contest matching task, introduced by Hessel et al. [7], gives this challenge a concrete form: given a cartoon, a description of its scene, and five candidate captions (one correct caption and four distractors drawn from other contests), select the caption that best fits (Fig. 1). It is a deceptively simple setup that turns on the hardest part of humor: the hidden connection between the image and the joke.

Annotation (`from_description`): *identical-looking men entering a room; a woman on the phone stares at them.* `uncanny`: *all the men look the same; this is not possible.*

Candidates: (A) **"On second thought, bring the name tags."** (B) "I guess this rules out a church wedding." (C) "In our opinion, you won't make it past the holidays." (D) "This way, it's a business trip." (E) "Remind me to close the curtains tonight."

Figure 1. A matching instance (Contest #584). Given the cartoon, a human scene annotation, and five candidate captions, one true winner (A, bold) and four distractors drawn from other contests, the model must select the winner. Chance is 20%.

What makes the task scientifically useful is a striking and still not fully explained gap. The strongest method in the original study reaches only about 62% accuracy, while expert humans reach about 94% [21]. Is the bottleneck representation (how faithfully an encoder embeds cartoons and captions into a comparable space) or reasoning (the ability to work out why a particular caption resolves the scene's incongruity)? The two diagnoses point to very dif-

ferent research directions, yet prior work does not fully separate them. This paper tests that distinction and finds that representation-focused methods saturate well below human performance.

Input and output. The input to our algorithm is a cartoon image I, a textual scene annotation D, and five candidate captions $C = \{c_1, \ldots, c_5\}$; the output is a predicted label $\hat{y} \in \{0, 1, 2, 3, 4\}$ identifying the best-matching caption. We attack the representation-versus-reasoning question from three directions: we build and ablate a CLIP-based cross-attention model that enriches the image and each candidate caption with the scene description before scoring; we sweep several representation-side design choices, including encoder scale, resolution, scoring rule, encoder family, and fine-tuning; and we probe reasoning-capable frontier vision–language models and training-free reasoning-style caption expansion, with leakage controls to keep the comparison honest.

2. Related Work

The benchmark. Hessel et al. [7] introduced the New Yorker Caption Contest as three tasks (matching, quality ranking, and explanation) and found a large gap between models and humans; we build directly on their matching task and reproduce their strongest configuration. Zhang et al. [21] scaled the contest to over 250 million human ratings and showed that even GPT-4 and Claude underperform top human contestants at *generating* captions, with alignment methods such as RLHF and DPO failing on this creative task; their crowd data establishes the human-expert ceiling (94.3%) we use as our upper anchor.

Vision underutilization. A recent survey, "Words Over Pixels" [8], documents a broad pattern in multimodal LLMs: they often lean on textual cues and underuse visual information. Our results are consistent with that observation but sharpen it for this benchmark: visual representation alone does not appear to be the binding constraint; candidate-level reasoning accounts for much of the remaining gap.

Representation backbones. Our models use CLIP [13] as a unified image–text encoder. We also evaluate alternatives along the representation axis: OpenCLIP encoders trained on LAION-2B [3], whose own scaling study found that the pretraining *distribution*, not merely scale, drives downstream behavior, matching our finding that these larger-corpus encoders underperform the original CLIP here; SigLIP 2 [17], a newer multilingual encoder family; and Gemini Embedding 2 [6], a natively-multimodal embedding model that reaches our representation ceiling with no task-specific training.

Cross-modal fusion. Our gated cross-attention enrichment builds on the transformer attention mechanism [18] and the established line of vision–language fusion models (ViLBERT [11], LXMERT [16], and BLIP-2 [9]) and the instruction-tuned models such as Flamingo [1] and LLaVA [10] from which today's frontier systems descend. Our own components are standard: a ViT image encoder [5] and, for the text-only baseline, DistilBERT [14].

Humor as structured reasoning. Incongruity-resolution accounts of humor hold that "getting" a joke means detecting an incongruity and then resolving it [15, 2, 4]; this is the conceptual basis for our hypothesis and for framing the task as reasoning. Most computational work nevertheless treats humor as black-box prediction. A notable recent exception is the Incongruity-Resolution Supervision framework of Vural et al. [19], which, like us, casts humor understanding as structured incongruity-resolution reasoning and supervises intermediate reasoning traces. Our contribution is complementary and diagnostic: rather than supervising reasoning, we compare representation-focused variants against reasoning-capable systems to estimate where the remaining matching gap lies. More broadly, the nuanced picture emerging from this literature, and reinforced by our results, is that current multimodal models can reliably perceive and describe a cartoon, yet still struggle with the cultural and inferential leap that makes a caption funny—which suggests the harder part lies more in reasoning than in perception.

3. Methods

3.1. Problem formulation and metrics

We treat caption matching as 5-way classification. The input is a cartoon image I, a textual scene annotation D, and a set of five candidate captions $C = \{c_1, \ldots, c_5\}$, exactly one of which is the correct caption for that cartoon; the other four are real winning captions drawn from *other* contests. The output is a label $y \in \{0, 1, 2, 3, 4\}$ identifying the matching caption. We formulate matching as a scoring problem: a model assigns each candidate a scalar compatibility score with the cartoon, optionally conditioned on the annotation,

$$s_i = f_\theta(I,\, D,\, c_i) \in \mathbb{R}, \qquad i = 1, \ldots, 5, \tag{1}$$

Every candidate is scored *independently against the same* (I, D) context, and the model predicts $\hat{y} = \arg\max_i s_i$. In our CLIP-style models, s_i is a temperature-scaled cosine similarity between image and caption embeddings; in the text-only and frontier baselines, it is produced by the corresponding scoring procedure. We train with a label-smoothed five-way softmax cross-entropy over the candidate scores ($\varepsilon = 0.1$),

$$p_i = \frac{\exp(s_i)}{\sum_{j=1}^{5} \exp(s_j)}, \qquad \mathcal{L} = -\sum_{i=1}^{5} \tilde{y}_i \log p_i, \tag{2}$$

with $\tilde{y}_y = 1 - \varepsilon + \varepsilon/5$ and $\tilde{y}_{i \neq y} = \varepsilon/5$. Chance is 20%. The original benchmark distinguishes a *From-Pixels* setting, where the model sees the rendered cartoon, from a *From-Description* setting, where human annotations stand in for visual understanding. Our cross-attention models use both the cartoon image and the provided scene annotation, treating D as auxiliary semantic context rather than as a replacement for the image.

Our primary metric is matching accuracy, the fraction of items for which the model picks the true winner. Because the test split is small (N=528), we attach a 95% Wilson confidence interval [20] to each accuracy (a form that stays reliable for proportions on small samples), and we compare two systems on the *same* items with the exact McNemar test [12], which asks whether they truly differ on the items where their answers disagree. This guards against reading seed-level noise as a real effect, since our systems cluster tightly.

3.2. Representation: CLIP backbone and gated cross-attention enrichment

Design principles. All of our trained models share a single CLIP ViT-L/14 backbone [13] for both modalities. A unified, contrastively-pretrained encoder is what makes cross-modal attention meaningful: separately pretrained vision and text models would require learning additional alignment from very little data, making cross-modal attention unstable. For the same reason we keep CLIP's native dimension and its cosine scoring rather than learning a new head. Projecting to a lower dimension or relearning the score risks destroying the calibrated geometry that already gives $\sim$61% for free, and with only $\sim$700 unique cartoons (§4) a heavier trainable fusion would overfit. We therefore add only a lightweight gated cross-attention module that the model can learn to ignore if it does not help.

Hypothesis. Plain CLIP scores image and caption independently and never consults the scene. The benchmark's `from_description` annotation, however, encodes the incongruity that makes a cartoon funny—a scene description plus an explicit `uncanny:` clause naming what is absurd. Grounded in incongruity-resolution accounts of humor [15, 2, 4], we hypothesized that cross-attending to this annotation before scoring would let the matcher exploit an incongruity signal that cosine discards, and so break the ceiling.

Architecture. The CLIP image encoder produces image hidden states, and the CLIP text encoder produces token states for the annotation D and for each candidate c_i (D and the captions are tokenized to 77 tokens with the CLIP tokenizer; the `uncanny:` clause survives this cap). A cross-attention block $\mathrm{CA}(X, Y)$ lets a query stream X attend to a context Y via multi-head attention with residual connections and layer norm, stacked over ℓ=2 layers. Crucially, in our design the annotation D is the shared key/value for both branches: image tokens attend to D, and each candidate caption independently attends to the same D; the image and captions never attend directly to each other (Fig. 2). The image branch enriches the image embedding h_I with scene context,

$$\tilde{h}_I = \mathrm{CA}(h_I,\, D), \qquad \hat{h}_I = \alpha_{\mathrm{img}}\, \tilde{h}_I + (1 - \alpha_{\mathrm{img}})\, h_I, \quad (3)$$

where $\alpha_{\mathrm{img}} = \sigma(g_{\mathrm{img}})$ is a learned scalar gate; $\hat{h}_I$ is passed through CLIP's `visual_projection` and ℓ_2-normalized to give z_I. The caption branch (added in the dual variant) enriches each candidate against the *same* annotation, in CLIP's native text dimension with no projection (caption and annotation already share the text space),

$$\tilde{c}_i = \mathrm{CA}(c_i,\, D), \qquad \hat{c}_i = \alpha_{\mathrm{cap}}\, \tilde{c}_i + (1 - \alpha_{\mathrm{cap}})\, c_i, \quad (4)$$

followed by CLIP's `text_projection` and normalization to give z_{c_i}. Scoring is the temperature-scaled cosine similarity in CLIP's aligned space, $s_i = t\, z_I^\top z_{c_i}$, with t a learned temperature (CLIP's logit scale).

Design progression. We reached the final model in three steps, each fixing what the previous one exposed. *Image enrichment* (v3.1) enriches only the image; *dual enrichment* (v3.2, Fig. 2) adds the symmetric caption branch described above; and *early projection* (v3.3, our best) changes *where* the image enrichment happens. v3.1 and v3.2 cross-attend in CLIP's raw 1024-d vision space and apply `visual_projection` *afterward*, so the gate shapes a vector that the projection then redistorts, and an untrained 768→1024 bridge is needed to lift the 768-d description into that space. Early projection instead projects *first* (patches into the 768-d joint embedding space) and cross-attends there, the exact space where the cosine score lives, so the gate shapes the final scoring vector directly and no projection follows it. Because enrichment now happens at 768-d, the description is used natively on both branches and the untrained bridge is removed; the caption branch, the two gates, and cosine scoring are otherwise unchanged from v3.2. The two gates' converged values are revealing, and we report them in §5; the v3.1 and v3.3 architectures are shown in Appendix B.

3.3. Probing the reasoning axis

To ask whether the ceiling reflects representation or reasoning, we add three training-free probes that separate representation-focused changes from reasoning-capable decision procedures. (i) *Caption expansion*: a language model rewrites each candidate into an explicit statement of why it could be funny in context, and we score a late fusion $\alpha\, s_i^{\mathrm{exp}} + (1 - \alpha)\, s_i^{\mathrm{cap}}$ of the expanded and original CLIP scores on the frozen backbone. (ii) *Knowledge-rich encoder*: we swap CLIP for a natively-multimodal embedding

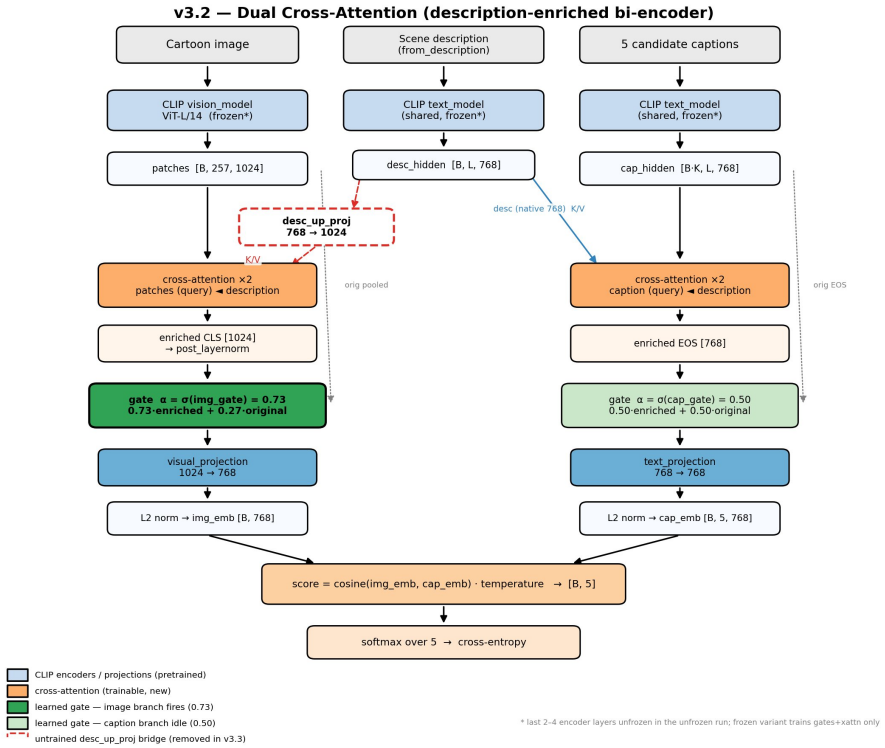


Figure 2. **Our dual cross-attention model** (*dual enrichment*). On a shared CLIP ViT-L/14 backbone, the image patches and each candidate caption are separately enriched by cross-attending to the human scene description, each behind its own learned gate, before temperature-scaled cosine scoring and a five-way softmax. The image gate opens to $\alpha_{\text{img}} \approx 0.73$ while the caption gate stays inert at $\alpha_{\text{cap}} \approx 0.50$: the model reports that half of the dual architecture is unused. The mechanism and the early-projection refinement are described in §3.2; baselines in Fig. 3.

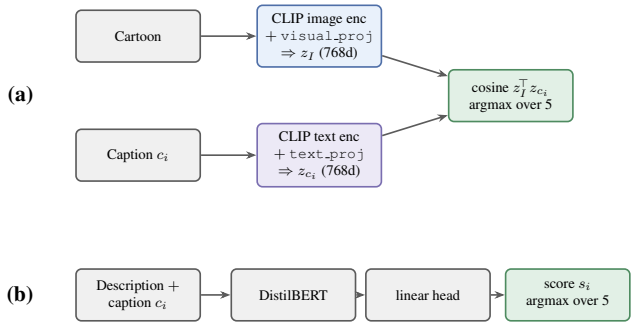


Figure 3. **Baseline architectures.** (a) CLIP cosine: dual-encoder similarity with no enrichment, evaluated zero-shot and with fine-tuning. (b) A text-only DistilBERT scores each *description + caption* pair with a linear head (no image). See §5 for results.

model (Gemini Embedding 2) and score zero-shot, with no training, to test whether a stronger *representation* alone moves the ceiling. (iii) *Frontier upper bound*: we present the cartoon and all five captions to frontier vision–language models (GPT-5, Claude Opus, and Gemini 3.5 Flash) in a single zero-shot choice, with no scene description, to estimate how far reasoning-capable systems can reach while recognizing that they also change scale, encoder, and scoring procedure. We describe this last setup and its controls next.

3.4. Frontier evaluation and leakage controls

A provider-identical harness gives each frontier model the cartoon image and the five candidate captions *only* (no scene description or human annotation) and asks for one sentence of reasoning followed by a single-letter answer (see Appendix C). We run one deterministic pass over the 528-item test split, with Gemini at its highest thinking setting and GPT-5 at low reasoning effort (the latter restricted due to API latency constraints).

Because the contest is heavily licensed and may sit in pretraining, the same harness runs two leakage controls. The *blank* control replaces the cartoon with a solid gray square while keeping the five captions, so a model that scores above chance must be working from the captions or from memory, not the picture. The *shuffle* control is stronger: each item keeps its own captions but is shown the *previous* item's real cartoon, so the correct answer is the caption for a cartoon the model never sees—a genuinely vision-grounded model is driven to chance, while a pure caption-text or look-up shortcut would survive.

3.5. Training regimes and implementation

Fine-tuning regimes. To separate the contribution of fine-tuning from that of architecture, we state the training regimes explicitly and study all of them: (a) *zero-shot* cosine, with no task-specific training; (b) *partial fine-tuning*, the default for our enrichment models, which unfreezes the last 4 vision and last 2 text transformer blocks alongside the new modules; (c) a *frozen-backbone* control that trains only the enrichment head; and (d) a *full fine-tune* used to faithfully reproduce the benchmark.

A note on capacity. These models carry far more trainable parameters than the data can constrain (only ∼700 unique cartoons), so all of them overfit. We document this and its consequences in §5.

What we implemented. We build on the HuggingFace `transformers` CLIP implementation and pretrained weights, and on the HuggingFace distribution of the benchmark; the CLIP encoders and tokenizer are library components. Everything task-specific is ours: the gated cross-attention enrichment modules for both branches, the early-projection routing, the scoring and label-smoothed objective, the training and evaluation loops, the reasoning-fusion probe, and the provider-agnostic frontier-evaluation harness with its leakage controls.

4. Task and Data

We use the *matching* configuration of the New Yorker Caption Contest benchmark [7], distributed as `jmhessel/newyorker_caption_contest`: each cartoon is paired with five candidate captions and the task is to pick the contest winner (Fig. 1). Each cartoon also comes with a human annotation, used by the *From-Description* protocol: a literal scene description, an `uncanny:` clause naming what is absurd, and a list of salient entities.

We use fold 0 of the benchmark's five-fold split: 9,792 training, 531 validation, and 528 test *instances*. One property of the data drives everything that follows: these are (cartoon, candidate-set) tuples, not distinct images. Because the contest produces only one cartoon per week, the matching split is built from only ∼700 unique cartoons. This is a hard structural ceiling—no amount of training adds new visual jokes—and it is the root cause of the overfitting we document in §5. For preprocessing we follow each encoder's native pipeline: images are resized to the encoder's expected resolution (224, or 336 for the high-resolution reproduction) and normalized with CLIP statistics. Only the Hessel reproduction uses the light, label-preserving augmentation from that recipe; our cross-attention experiments use deterministic CLIP preprocessing, since the binding constraint is the number of distinct cartoons, not pixel-level variation.

5. Experiments and Results

5.1. Experimental setup

Our primary metric is five-way matching accuracy, the fraction of examples for which the highest-scoring caption is the ground-truth caption. We report 95% Wilson confidence intervals for single-model accuracies and paired McNemar tests for matched comparisons on the same test items.

We evaluate on the matching configuration of the benchmark (§4). The frontier and reproduction experiments use CLIP ViT-L/14 at 336×336 with pad-to-square preprocessing (bicubic resize, no center-crop) so the whole cartoon stays in view; our cross-attention experiments use ViT-L/14 at 224×224. This preprocessing choice is consequential: the default center-crop sees only the middle of the cartoon and scores 48.7% zero-shot, whereas pad-to-square recovers ∼62%—a ∼12-point gain from crop handling alone.

Optimization and hyperparameters. All trainable models use AdamW, 10% linear warmup followed by cosine decay, gradient clipping at 1.0, bf16 mixed precision, and mini-batches of 16 on a single NVIDIA A40. Our enrichment models use differential learning rates, 3×10^{-5} for newly initialized modules and 3×10^{-6} for unfrozen CLIP layers; the Hessel reproduction full-fine-tunes CLIP at 5×10^{-6}. We tuned only regularization (weight decay 0.05, dropout 0.15, label smoothing 0.1, and freezing all but the last 2–4 encoder layers) because the model otherwise reached 100% training accuracy within a few epochs. We keep the best-validation checkpoint by early stopping.

Cross-validation. A single 528-item split is noisy at this dataset size, so for our representation-ceiling models we also run the benchmark's native 5-fold cross-validation. Our strongest zero-shot encoder, Gemini Embedding 2, scores 62.2% ± 1.85% across folds (pooled 62.2%, 95% Wilson CI [60.3, 64.0], consistent with its 61.2% single-fold result in Table 2); the remaining systems are reported on the standard single test fold. Chance is 20% and a strong human solver reaches ≈ 94% [21].

5.2. The representation ceiling

Custom architectures do not significantly beat CLIP-style cosine scoring. Table 1 compares our trained models (baseline designs in Fig. 3). Zero-shot CLIP cosine (image + caption) already reaches 48.7%, and a description+caption text model with no image reaches the same 48.7%. Under these baselines, the text-only signal is competitive with zero-shot image+caption CLIP. Partial fine-tuning and cosine scoring (*FT-cosine*) helps substantially, raising CLIP cosine to 58.5%; but our three cross-attention variants (image, dual, and early-projection) remain in the same band at 58.3/59.3/60.4%. All trained variants are statistically tied under paired McNemar. Early pro-

Table 1. Trained models on the matching task (CLIP ViT-L/14, n=528 test). All trained variants are statistically tied under paired McNemar—within this saturated, small-data regime, no architecture we tried beats plain cosine.

Model	Test acc. (%)
Random chance	20.0
Description + caption (DistilBERT)	48.7
CLIP cosine, image + caption (zero-shot)	48.7
CLIP cosine, image + caption (fine-tuned)	58.5
Image enrichment	58.3
Dual enrichment	59.3
Early projection	**60.4**

jection is nonetheless our strongest and the one we carry forward (§3.2): projecting before enriching lets the gate act directly in the 768-d joint scoring space and removes the untrained up-projection bridge, the cleanest of the three designs, though its edge over the dual and image variants stays within that McNemar noise. One caveat matters for how we read this: every trained model saturates the tiny training set (§5.4), so the fair conclusion is that the cross-attention machinery buys nothing *in this small-data regime*—not that architecture could never help given more cartoons. (Two fine-tuned CLIPs recur below: the partial *FT-cosine* at 58.5% here, and a full fine-tune at 336px that reproduces Hessel, the *Hessel-repro* at 61.2%, in Fig. 4.)

No representation-focused variant breaks ∼61–62%. Figure 4 tests the main representation-side choices available to us (all From-Pixels). A faithful reproduction of Hessel et al. [7] (full fine-tune, CLIP ViT-L/14@336) reaches 61.17% with a 95% interval of $[57.0, 65.2]$ that contains their reported 62.3%—we reproduce the benchmark. Dropping resolution to 224 costs almost nothing; swapping in LAION-2B encoders [3] *lowers* accuracy despite far larger pretraining, token-level MaxSim scoring underperforms pooled cosine, and SigLIP 2 [17] reaches only ∼48%. Scale, resolution, scoring granularity, and pretraining corpus all leave the ceiling near 61%, suggesting that pretraining *distribution* matters more here than scale alone.

A knowledge-rich encoder reaches the ceiling for free. Swapping CLIP for Gemini Embedding 2 and scoring zero-shot, no training at all, reaches 61.17% $[57.0, 65.2]$, on par with our Hessel-repro CLIP. This suggests the 61–62% plateau is not an artifact of our fine-tuning procedure: a stronger pretrained embedding model reaches it immediately, but does not move beyond it.

The gates show asymmetric use of the description. The early-projection model's learned gates are interpretable and reproduce across runs: the image gate opens to α_{img}=0.73 (scene context genuinely helps image scoring) while the caption gate remains at its neutral initialization, α_{cap}=0.50 (the model does not learn to rely on caption-side enrichment). The model itself reports that half of the dual architecture is inert.

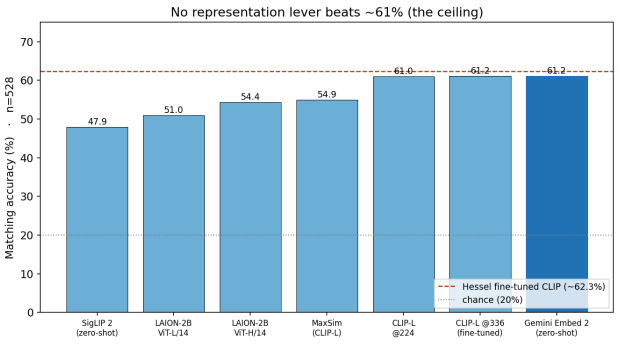


Figure 4. Representation-side design choices fail to break the ∼61–62% ceiling against the 62.3% Hessel reference line. Larger pretraining (LAION-2B), higher resolution, token-level scoring (MaxSim), and a different encoder family (SigLIP 2) do not move the ceiling; the pretraining distribution does.

5.3. Reasoning-capable systems break the ceiling

Frontier models clear the wall. Table 2 reports zero-shot frontier accuracy on the same 528 items. Where no representation-focused method exceeds the 61–62% band, Gemini 3.5 Flash reaches 93.56% $[91.1, 95.4]$, with GPT-5 at 87.88% and Claude Opus 4.7/4.8 at 82.01/80.87%—approaching the ≈ 94% human level. (That human number is an imported anchor from Zhang et al. [21], measured on different cartoons, not on our 528 items.) The gap that no representation-focused method could close, these systems close almost entirely. We read this cautiously: the frontier models differ from our trained models in several ways at once: their own stronger encoders, a joint forced choice that weighs all five captions together rather than scoring each by cosine, and far greater scale. So the jump reflects that *combination*, not reasoning in isolation. The ordering is also not a clean capability ranking (a fast model tops two larger ones), which, as we discuss next, is also consistent with the models having seen differing amounts of this licensed corpus.

How much is grounding, how much exposure? Because the contest is heavily licensed and likely appears in pretraining, we test whether frontier accuracy reflects genuine grounding. Two controls rule out the simplest shortcuts: a blank image drops Gemini to chance (22% on a 50-item probe) and the shuffle control (its own captions over the wrong cartoon, on the full 528 items) drops it to 19.89%, so the 93.56% depends on seeing the *right* cartoon—the model is neither guessing from captions alone nor ignoring the image.

Gemini shows signs of contest exposure—but it does not help it cheat. The blank control is revealing: where GPT-5 and Claude (each on a 100-item blank probe) mostly

Table 2. Zero-shot frontier models against our best representation model and an imported human anchor (image + 5 captions, n=528; the human number is from Zhang et al. on different cartoons). Intervals are 95% Wilson on our 528-item test split; a dash marks figures imported from prior work (the Hessel reference and the human anchor), not measured here. These systems close most of the gap no representation-focused method could—a combination of representation, joint reasoning, and scale (see text).

Model	Acc. (%)	95% CI
Hessel et al. ref [7]	62.3	—
Early projection (ours)	60.4	[56.2, 64.5]
Gemini Embedding 2 (0-shot)	61.2	[57.0, 65.2]
Claude Opus 4.8	80.9	[77.3, 84.0]
Claude Opus 4.7	82.0	[78.5, 85.0]
GPT-5	87.9	[84.8, 90.4]
Gemini 3.5 Flash	**93.6**	[91.1, 95.4]
Human expert [21]	94.3	—

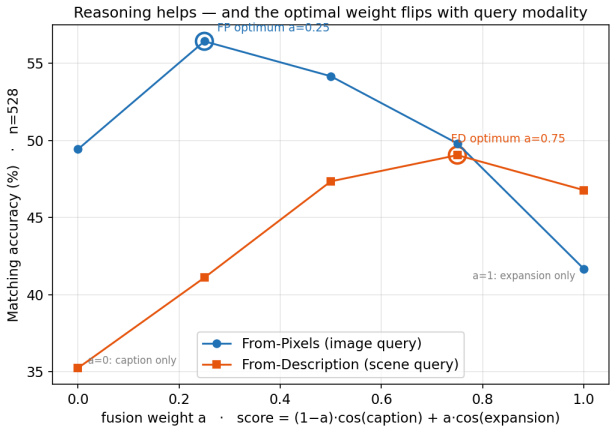


Figure 5. Late-fusion weight α between the LLM caption-expansion score and CLIP's. The optimum flips with the query modality (α=0.25 From-Pixels vs. α=0.75 From-Description), reflecting how much textual reasoning each protocol needs.

admit they cannot see an image and then guess, Gemini reaches into pretraining—of its 50 blank responses, roughly a fifth invoke "winning caption" or finalist look-up language and a few cite specific *Contest #NNN* numbers, direct evidence that the dataset sits in its training data. *Yet the recall is inert as a shortcut: blank accuracy stays at chance and the shuffle control collapses Gemini to* 19.89% *even with a real (wrong) cartoon in view.* Recall lets it recite captions; only *seeing the correct cartoon* lets it answer, so the 93.56% reflects genuine visual grounding, not retrieval.

We still cannot fully exclude that recognising a particular cartoon–caption pair helps on *some* items, since that would pass both controls; the clean test is a temporal holdout on cartoons published after each model's training cutoff, which we leave to future work. We therefore read the frontier numbers as an *upper bound* on what reasoning-capable systems reach, not as a pure measure of reasoning.

Training-free reasoning helps, partially. Figure 5 adds an LLM caption-expansion step and late-fuses its score with CLIP's on the frozen backbone. It helps significantly: From-Pixels rises from 49.43% to 56.44% (+7.0pp, McNemar p=10^{-4}) and From-Description from 35.23% to 49.05% (+13.8pp, p<10^{-4}). The optimal fusion weight *flips* with the query modality (From-Pixels peaks at α=0.25, From-Description at α=0.75): when the model already sees the image it needs little textual reasoning, when it sees only text it needs much more. Stacking the same expansion on the stronger Gemini encoder, however, adds only +3.4pp (61.17% → 64.58%) and does *not* reach significance (p=0.06): on an encoder that already carries world knowledge, the reasoning step has largely saturated. (We run many paired comparisons in this paper, so we read an isolated result near p=0.05 as suggestive, not conclusive.)

Prompt-level reasoning is not enough. A structured chain-of-thought scaffold (describe → enumerate the incongruity → judge each candidate) on GPT-5 does *not* help: 86.55% vs. 87.88% direct, a non-significant −1.33pp (paired McNemar p=0.35). The reasoning that closes the gap is the model's own internal, decision-level reasoning, not an external prompt structure—consistent with the embedding-level result that injected reasoning saturates.

5.4. Overfitting is structural

All enrichment models reach 100% training accuracy within two epochs while validation plateaus near 60%—a ∼40pp gap—and the label-smoothed cross-entropy hits its theoretical floor (≈ 0.39) just as fast, leaving no residual gradient. The cause is the data, not the model: the matching split is built from only ∼700 unique cartoons. A frozen-backbone control makes this concrete—freezing CLIP cuts trainable parameters from 92M to 26M yet leaves test accuracy essentially unchanged (57.6 vs. 58.3%), so the overfitting lives in the lightweight head, not the backbone. Standard mitigations (label smoothing, dropout, weight decay, early stopping) stabilize training but do not raise the plateau, because the ceiling is set by the number of distinct cartoons, not by regularization. The same saturation underlies a related null result: varying what the early-projection model attends to (nothing, the scene, the `uncanny:` clause, or the full human rationale) moves test accuracy only within the ±1.3pp seed band, so within this regime added scene access does not help (consistent with the inert caption gate reported above).

5.5. Error analysis

The failures are consistent with a reasoning gap rather than a perception gap. Consider Contest #575 (see Ap-

Table 3. Per-model predictions on a hard "hidden-connection" item (Contest #575; gold = B, ★). Each representation model, including our early-projection model, lands on a different distractor, while both frontier reasoners recover the winner.

Candidate caption	Picked by
A. Evolution can be so tacky.	CLIP (0-shot)
B. Let's pick up the pace. They're billing by the hour. ★	GPT-5, Gemini
C. Well, five acres of popcorn back there says you were.	—
D. I'm sure it works. I was involved in the original research.	Early-proj.
E. Just water for me, thanks.	CLIP (FT)

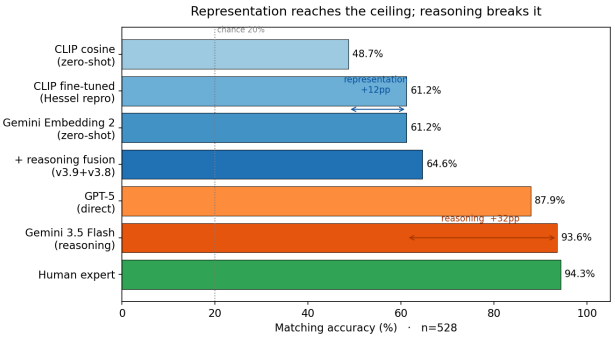


Figure 6. The representation–reasoning decomposition (test set, n=528). Accuracy climbs from chance (20%) toward the human level (94.3%) in two stages. Refining *representation*, from zero-shot CLIP cosine (48.7%) to the Hessel reproduction or a natively-multimodal encoder (both ∼61%), gains roughly 12 points and then saturates. Frontier vision–language models (87.9–93.6%) close most of the rest, combining stronger representation, joint reasoning over the candidates, and scale; a controlled training-free reasoning step (64.6%) helps only modestly.

pendix A, Figure 7), in which armored conquerors on horseback lead a long column of suited businessmen across a plain; the winning caption—"Let's pick up the pace. They're billing by the hour."—bridges the ancient-conquest scene to modern corporate billing, an unstated connection between two incongruous domains. Table 3 shows each model's pick. Every representation model (zero-shot CLIP, fine-tuned CLIP, and our early-projection model) lands on a *different* distractor by matching surface content, while both frontier reasoners recover the winner. The case is representative: across the 183 such hard items in the test split, the early-projection model misses 58.5% (107/183). The pattern holds throughout—representation locates *what* is in the cartoon; only reasoning recovers *why* a caption resolves it.

5.6. Discussion

Our results split the climb from chance to the human level into two parts (Fig. 6). Representation saturates early: zero-shot CLIP at the default center-crop reaches 48.7%, and representation-side improvements (pad-to-square preprocessing, fine-tuning, and a stronger encoder) reach a ceiling near 61% (the Hessel reproduction and a natively-multimodal encoder both land there), a gain of roughly 12 points that no representation-side design choice we tried reliably breaks. A representation tells the model *what* is in the cartoon, and the controls confirm it is genuinely used. The second part, from ∼61% to the ∼94% frontier systems reach, we resist attributing entirely to "reasoning": those systems change encoder, scoring (a joint choice over all five captions, not independent cosine), and scale at once, so the jump is best read as that bundle. What we *can* isolate about reasoning comes from the controlled probes.

What the controlled probes say. Three controlled probes point the same way: adding an explicit reasoning step (caption expansion) helps a weak encoder (+7pp) but barely a strong one; a chain-of-thought prompt does not help a frontier model with headroom; and giving our own model the full human rationale does not help at all. The missing ingredient is the model's own internal reasoning about *why* a caption resolves the scene. Caption matching is thus better seen as a reasoning task with a visual-grounding requirement than a representation task: the grounding is necessary but easy to reach, which is why representation-centric methods plateau far below human performance.

Why a *Flash* model tops the table. Gemini 3.5 Flash's lead should not be over-interpreted as a clean capability ranking. Its performance may reflect both stronger native multimodal grounding and greater exposure to the licensed contest corpus.

6. Conclusion

We asked whether New Yorker caption matching is bottlenecked by representation or reasoning. Representation-focused methods saturate near 61–62%: fine-tuning, larger pretraining, alternative encoders, and our cross-attention enrichment all remain in that band, while a knowledge-rich encoder reaches it with no training. Frontier systems close most of the gap to the ∼94% human anchor, suggesting that joint candidate-level reasoning and scale, not representation alone, drive the remaining improvement.

These conclusions come with limitations. Our strongest models are closed API systems, which constrains reproducibility; our trained models overfit the small set of unique cartoons regardless of regularization; and although our controls rule out the simplest shortcuts, the benchmark's licensing means we cannot fully exclude that some frontier accuracy is recognition rather than reasoning. The cleanest next steps are a temporal holdout on post-cutoff cartoons to settle contamination, and an open-weight vision–language model to make the reasoning claim reproducible and analyzable. Teaching a model to truly get the joke, it seems, is not just about how it sees the cartoon, but about how it reasons over what it sees.

Contributions and Acknowledgements

This was a solo project. The author thanks the CS231N teaching staff for the course, feedback, and project guidance. The implementation builds on HuggingFace `transformers` and the HuggingFace distribution of the New Yorker Caption Contest benchmark [7]; training and evaluation were run on a single NVIDIA A40 GPU.

Generative AI tools, including Anthropic Claude and OpenAI ChatGPT, were used for concept explanation, editing and formatting assistance, code debugging, analysis support, and generating charts, graphs, and architecture diagrams. The underlying ideas, experimental design, methodology, analysis, and conclusions are the author's own, and all AI-assisted material was reviewed and verified by the author.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] S. Attardo and V. Raskin. Script theory revis(it)ed: Joke similarity and joke representation model. *HUMOR: International Journal of Humor Research*, 4(3–4):293–347, 1991.

[3] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023.

[4] S. Coulson. *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge University Press, 2001.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] Google DeepMind. Gemini embedding 2: A native multimodal embedding model from Gemini, 2026.

[7] J. Hessel, A. Marasović, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, and Y. Choi. Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 688–714, 2023.

[8] A. Jain, M. Vatsa, and R. Singh. Words over pixels? Rethinking vision in multimodal large language models. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*, pages 10481–10489, 2025.

[9] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

[10] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[11] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[12] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, 2019.

[15] J. M. Suls. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In J. H. Goldstein and P. E. McGhee, editors, *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*. Academic Press, 1972.

[16] H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[17] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[19] H. M. Vural, D. Kukul, E. E. Ozlu, D. E. Arikan, B. Mankoff, E. Erdem, and A. Erdem. Learning to think like a cartoon captionist: Incongruity-resolution supervision for multimodal humor understanding, 2026.

[20] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[21] J. Zhang, L. Jain, Y. Guo, J. Chen, K. L. Zhou, S. Suresh, A. Wagenmaker, S. Sievert, T. Rogers, K. Jamieson, R. Mankoff, and R. Nowak. Humor in AI: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.10522.

Appendix

A. Additional Cartoon Examples

To give a fuller sense of the cartoons the matching task draws on, we show a few more examples. Figure 7 is the cartoon behind the error analysis in §5, which the text only describes; Figure 8 shows two further cartoons that illustrate the range of scenes, drawing styles, and visual gags a model must handle.

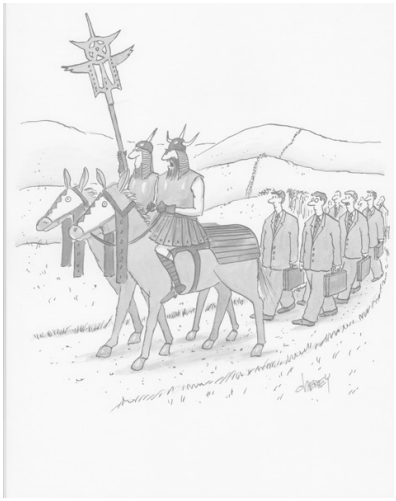
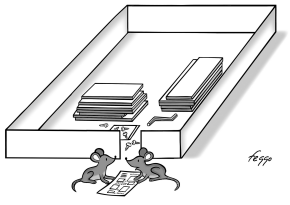


Figure 7. Contest #575, the cartoon behind our error analysis. Armored conquerors on horseback lead a column of suited businessmen across a plain; the winning caption—"Let's pick up the pace. They're billing by the hour."—bridges the ancent-conquest scene to modern corporate billing. Every representation-only model in our study misses this unstated connection; both frontier reasoners recover it.



"Let's just go with the open floor plan."

"Looks like you're already familiar with the side effects."

Figure 8. Two further cartoons with their winning captions (Contests #667, left, and #700, right). The humor turns on incongruities that are easy for a person to see but hard to state—part of what makes the matching task a test of reasoning rather than perception.

B. Architecture details

The body shows our dual model (Fig. 2). For completeness, Fig. 9(a) shows the image-only predecessor (v3.1), which enriches only the image branch and leaves the captions as plain CLIP, and Fig. 9(b) shows the early-projection model (v3.3, our best), which applies `visual_projection` *before* cross-attending (so both branches enrich in the 768-d joint scoring space) and drops the untrained 768→1024 description bridge (§3.2).

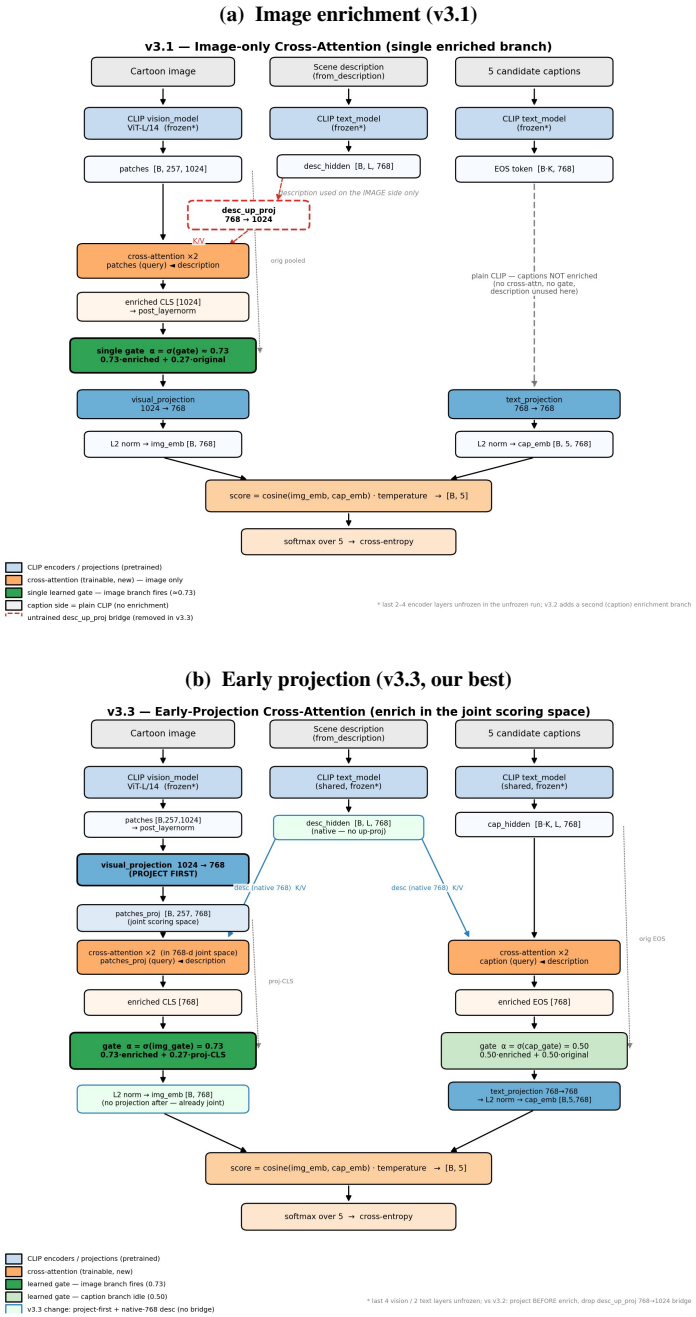


Figure 9. **Architecture variants bracketing our dual model** (Fig. 2). **(a)** *Image enrichment* (v3.1): only the image attends to the scene description; the candidate captions pass through plain CLIP (single learned gate $\alpha_{\text{img}} \approx 0.73$). **(b)** *Early projection* (v3.3, our best): `visual_projection` is applied first, so both branches cross-attend in the 768-d joint scoring space and the untrained 768→1024 bridge is removed; the caption branch, the two gates, and cosine scoring are otherwise as in v3.2.

C. Frontier Evaluation Prompt

Every frontier model received the identical zero-shot prompt below, together with the cartoon image and the five candidate captions; no scene description was provided.

```
This image is a cartoon from The New
Yorker Cartoon Caption Contest.  Below
are five candidate captions labeled
A through E. Exactly one of them is
the caption that was actually paired
with this cartoon; the other four are
distractors from other cartoons.  Choose
the single caption that best matches THIS
cartoon.

A. [caption A]  ...  E. [caption E]

Give exactly ONE sentence of reasoning
inside <think></think>, then your
final choice as a single letter inside
<answer></answer>.
```

The answer is parsed by preferring the contents of `<answer>`, then a trailing "answer: X," then the last lone A–E letter in the reply. The chain-of-thought variant (§5) keeps this format but inserts a three-step instruction before the answer (describe the scene, name the incongruity, then judge each candidate) rather than a single sentence of reasoning.