Drawing Conclusions: Representation or Reasoning in New Yorker Caption Matching

Spoiler: it's reasoning, by about 32 points.

Matthew T Bouchard

CS231N Spring 2026, Stanford University

Introduction

Understanding why a cartoon is funny is a demanding test of multimodal intelligence: a model must link the scene to an unstated, often absurd premise.
The New Yorker Caption Contest matching task (Hessel et al. [1]) makes this concrete: given a cartoon and a scene annotation, pick the winning caption from five. Figure 1 shows the difficulty: nothing in the drawing says death rays; the joke is a leap the pixels never make.
The best published model reaches only ~62%, while expert humans reach ~94%. What is that gap made of?

Problem Formulation: Given cartoon image I, scene annotation D, and five candidate captions C = {c_1,…,c_5}, predict y ∈ {0,…,4}, the winning caption. Distractors are real winners of other contests.
Evaluation: 5-way matching accuracy (20% chance); 95% Wilson CIs; exact paired McNemar tests.
Research Question: Is the bottleneck representation (faithfully embedding cartoon and captions) or reasoning (working out why a caption resolves the incongruity)? We probe encoders, fine-tuning, cross-attention enrichment, and training-free reasoning to decompose the gap.

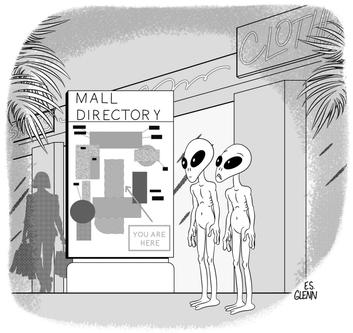


Figure 1: Contest 895. Winning caption: "Do you think death rays would be considered electronics or sporting goods?"

Dataset

New Yorker Magazine Cartoon Contest Dataset
(https://huggingface.co/datasets/jmhessel/newyorker_caption_contest)

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

From_description(textual mode):
An empty room description: There are a bunch of men that look identical coming into a room. A woman is on the phone staring at them. uncanny: All the men look the same. This is not possible.

Figure 2: Contest 584

Label(winning caption): A (human selected)

Dataset Organization

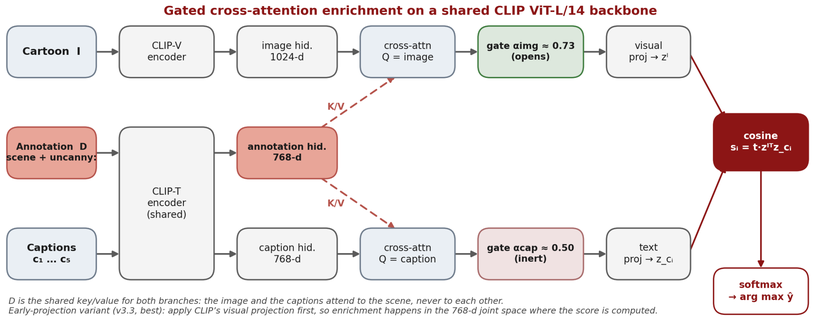
Fold 0 of the benchmark's five-fold split:
9,792 training examples
531 validation examples
528 test examples

Built from only ~700 unique cartoons (one contest per week). No amount of training adds new visual jokes; this scarcity, not architecture, drives the overfitting we document.

Data Preprocessing

Images: resized to each encoder's native resolution (224; 336 for the high-res reproduction), CLIP normalization.
Text: annotation and captions tokenized with the CLIP tokenizer and encoded in CLIP's native 768-d embedding space.

Methods



D is the shared key/value for both branches: the image and the captions attend to the scene, never to each other.
Early-projection variant (v3.3, best): apply CLIP's visual projection first, so enrichment happens in the 768-d joint space where the score is computed.

Principle: preserve CLIP's geometry. A shared CLIP ViT-L/14 encodes both modalities; scoring is temperature-scaled cosine in CLIP's aligned space, $s_i = t \cdot z'^{T} z\_c_i$, with softmax + label-smoothed cross-entropy (ε = 0.1). Each candidate is scored against the same (I, D). We rejected a learned ITM head: it would overfit ~700 cartoons and discard the geometry that yields ~61% for free.

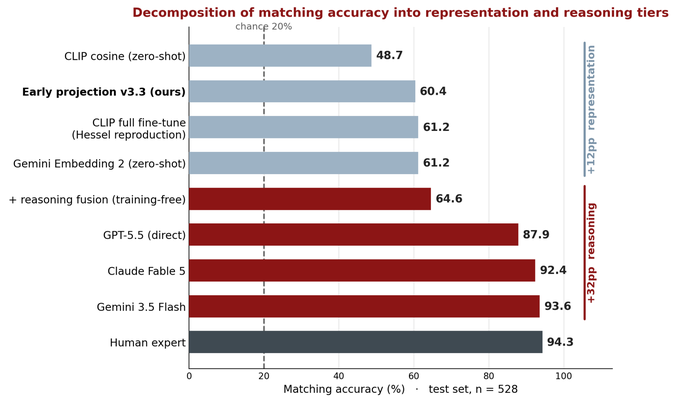
Gated enrichment. The annotation D is the shared key/value for both 2-layer, 8-head cross-attention branches; the image and the captions attend to the scene, never to each other. Learned scalar gates blend enriched with original, and they are diagnostic: αimg ≈ 0.73 opens, αcap ≈ 0.50 stays inert. The best variant (v3.3) enriches after visual projection, in the 768-d space where the score is computed.

Reasoning probes (training-free, scoring rule fixed): (i) LLM caption expansion + late fusion; (ii) a knowledge-rich encoder (Gemini Embedding 2) scored zero-shot; (iii) frontier VLMs in a single forced choice, with blank-image and shuffled-cartoon leakage controls.

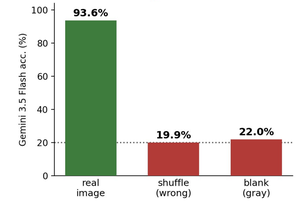
Training. Partial fine-tune (last 4 vision + 2 text blocks) + new modules; AdamW 3e-5, weight decay 0.05, dropout 0.15, early stopping (patience 5), batch 16, one L4 GPU. Overfitting is structural: 100% train accuracy within 2 epochs while validation plateaus near 60%; a frozen-backbone control (92M → 26M params) leaves test accuracy unchanged.

Statistics. 95% Wilson intervals on every accuracy; exact paired McNemar tests for same-example comparisons. Our systems cluster tightly, so this guards against reading seed noise as a real effect.

Results



Model	Test acc. (%)	Released
Random chance	20.0	-
Text-only DistilBERT	48.7	-
CLIP cosine (zero-shot)	48.7	-
Cross-attn v3.1 (image-only)	58.3	-
Dual cross-attn v3.2	59.3	-
Early proj. v3.3 (ours, best)	**60.4**	**-**
Gemini Embedding 2	61.2	-
GPT-5.5	87.9	Apr 2026
Claude Fable 5	92.4	Jun 2026
Gemini 3.5 Flash	93.6	May 2026
Human expert [2]	94.3	-



- **Representation gains saturate near 61%; reasoning accounts for the remaining ~32 points:** every axis we varied (scale, resolution, MaxSim, SigLIP 2, fine-tuning, our cross-attention) hits the same ceiling, which a knowledge-rich encoder reaches zero-shot; frontier VLMs jump to 87.9–93.6% vs the 94.3% human ceiling. Controls (right) confirm genuine grounding.
- **Access ≠ ability:** the gold human rationale doesn't help (within ±1.3pp seed noise), the caption gate stays inert (α≈0.50), and chain-of-thought doesn't help GPT-5.5 (−1.3pp, p=0.35). What's missing is internal, decision-level reasoning over the five candidates.

Conclusions and Future Work

We asked whether caption matching is bottlenecked by representation or by reasoning. Representation is necessary but saturates early: every axis we varied plateaus near 61%, and a knowledge-rich encoder reaches the same ceiling with no task-specific training. Reasoning supplies the rest: frontier models reach ~94%, near the human ceiling, while reasoning injected into our pipeline helps only marginally. The bottleneck is reasoning capacity, not access to scene information.
Limitations: closed-API strongest encoder; overfitting to ~700 unique cartoons despite regularization; residual exposure cannot be excluded.
Future work: decision-level reasoning over a strong open-weight VLM (analyzable, reproducible); and broader humor data, which remains scarce.
We did try explaining the jokes to our models. As with people, it did not help.

References
[1] Hessel et al. Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker caption contest. ACL 2023. [2] Zhang et al. Humor in AI: massive scale crowd-sourced preferences and benchmarks for cartoon captioning. NeurIPS D&B 2024. [3] Radford et al. Learning transferable visual models from natural language supervision (CLIP). ICML 2021.
[4] Vural et al. Learning to think like a cartoon captionist: incongruity-resolution supervision for multimodal humor understanding. 2026. [5] Google DeepMind. Gemini Embedding 2: a native multimodal embedding model from Gemini. 2026.