Active-Inference Captioning: Bridging Perception and Action in Zero-Shot Image Captioning

Anonymous ACL submission

Abstract

Strict text-only zero-shot image captioning trains a small mapping network on text features alone and, at inference, ranks candidate captions with a single discriminative image-text matching model. On a controlled near-miss benchmark, the matchers we evaluate saturate noun discrimination but degrade sharply on counts, spatial relations, and attributes – a failure mode we call *perception-bound hallucination*. We propose **Active-Inference Captioning (AIC)**, an inference-only reranker that pairs a frozen image-text matcher with a frozen multimodal LLM judge under paired-choice prompting. The two scores are combined under a Bayesian-motivated additive weight that is selected on a held-out development set and applied without modification across benchmarks. The bulk of the gain over the prior strict-ZIC state of the art is attributable to the stronger image-text matcher; the MLLM verifier yields a complementary, smaller-magnitude improvement, concentrated on counting, spatial-relation, and attribute hallucinations on which discriminative matchers saturate. AIC reaches CIDEr $\mathbf{116.9}$ on COCO Karpathy, $\mathbf{69.4}$ on Flickr30k Karpathy, and $\mathbf{82.3}$ overall ($\mathbf{81.0}$ out-of-domain) on NoCaps val, with no learned parameters in the captioning loop. Paired bootstrap and a verifier-swap robustness check support these gains.

1 Introduction

Recent strict zero-shot image captioning, where a captioner is trained on a text corpus alone and bridges the image-text gap at inference, has matured around a single-model recipe. A small mapping network (Mokady et al., 2021) is trained on CLIP-text features alone, the captioner emits a beam of candidate captions for an unseen image, and a CLIP-style image-text matching model ranks the candidates. Methods in this lineage (DeCap (Li et al., 2023b), CapDec (Nukrai et al., 2022), ViECap (Fei et al., 2023), MeaCap (Zeng et al., 2024), and the previous state-of-the-art IFCap (Lee et al., 2024)) differ in the mapping network, retrieval cache, or prompting template, but all share the same final step: a single discriminative image-text matching model decides which candidate caption is committed for an image. CIDEr on COCO Karpathy has improved by roughly one point per year along this line, with IFCap reaching 108.0 as the previous strict-text-only best. A parallel synthetic-image-augmented line, led by NES (Lu et al., 2026) at 109.9, breaks this recipe by generating training images at training time. Although NES sits outside the strict text-only regime our analysis targets, AIC also surpasses it on every benchmark we evaluate without using any synthetic image data.

Our analysis reveals an empirical failure mode of this single-model recipe that is not closed by any matcher we evaluate. On a controlled near-miss benchmark of caption pairs, the strongest matching model saturates noun-swap discrimination but degrades sharply on spatial, count, and attribute substitutions; strengthening the encoder (CLIP (Radford et al., 2021), OpenCLIP-bigG (Cherti et al., 2023), or a diffusion-likelihood readout) does not change the pattern. The hallucinations the tested matchers do not resolve are concentrated on instances that, intuitively, require *revisiting* the image, for example counting instances, comparing depths, or attending to a specific region. We do not claim that no future image-caption scorer could close this gap; a sufficiently expressive matcher could in principle attend, count, or compare regions internally. The bounded empirical claim is that the matchers we evaluate, including BLIP-ITM-large, do not close it. We name this failure mode *perception-bound hallucination*.

A natural response is to add a second model that can answer the questions the matcher cannot. Recent vision-language work uses frozen multimodal LLMs (MLLMs) as *judges* that score or compare

candidate outputs of another model, in adjacent settings such as text-only LLM reranking (Zhang et al., 2024), multimodal reasoning (Ong et al., 2025), and best-of-N rerankers built on top of BLIP-ITM (Li et al., 2023a; Benhammou et al., 2025); to our knowledge they have not been deployed inside the strict text-only zero-shot captioning regime. Empirically, an MLLM judge under *paired-choice* prompting narrows the matching-model failure gap on every perception-bound class on our near-miss pilot (Table 1); the corresponding lift at benchmark-level CIDEr is smaller and is reported separately in Sec. 5.2. The MLLM verdict is evidence the image-text matching model cannot reach, because the verdict conditions on the image rather than scoring a fixed caption against it.

We therefore propose **Active-Inference Captioning (AIC)**, an inference-only reranking framework that pairs a frozen image-text matcher with a frozen MLLM judge (Fig. 1). AIC reranks IFCap's released 20-beam with two frozen components: a *Re-ranker* (BLIP-ITM-large (Li et al., 2022)) that scores image-caption matching, and an *MLLM Judge* (K=4 Qwen3-VL-8B (Bai et al., 2025) paired-choice verdicts) that verifies candidates against the image. Their contributions combine under a Bayesian-motivated additive weight λ (Sec. 3), calibrated once on a 500-image COCO restval dev split and reused unchanged. We show empirically (Sec. 4) that the verification term encodes information the evaluated matchers do not already produce; we do not claim impossibility for arbitrary future matchers. Prior strict-ZIC methods (IFCap (Lee et al., 2024), NES (Lu et al., 2026), SynTIC (Liu et al., 2024), PCM-Net (Luo et al., 2024), MILS (Benhammou et al., 2025), EDITOR (Mahajan et al., 2024)/VGD (Kim et al., 2025)) are restricted parameter settings of AIC's design space. Honestly decomposed, on COCO (+8.9 CIDEr over IFCap) the bulk (+7.3) comes from replacing CLIP cosine with BLIP-ITM-large; the MLLM verification term contributes a modest but consistent +1.6, concentrated on spatial/count/attribute classes. AIC establishes a new operating point for *MLLM-assisted inference-time reranking of strict-text-trained captioners* on COCO (Lin et al., 2014) Karpathy (Karpathy and Fei-Fei, 2015), Flickr30k (Young et al., 2014) Karpathy, and NoCaps val (Agrawal et al., 2019), with no learned parameters in the captioning loop. The largest gain falls on the NoCaps out-of-domain split. Because the Re-ranker (BLIP-ITM-large-coco) and Judge are themselves trained on web-scale image-text data, comparisons to prior CLIP-only strict-ZIC methods should be read as contextual rather than parity-of-regime.

2 Related Work

Strict zero-shot image captioning generates captions without paired image-text training data. *Late-guidance* methods (ZeroCap (Tewel et al., 2022), MAGIC (Su et al., 2022)) inject visual guidance after word prediction by reranking with a CLIP score. *Early-guidance* methods (CapDec (Nukrai et al., 2022), DeCap (Li et al., 2023b), ViECap (Fei et al., 2023), MeaCap (Zeng et al., 2024), IFCap (Lee et al., 2024)) prefix CLIP-derived visual prompts to the language model and decode under that prefix; IFCap reaches 108.0 CIDEr as the previous strict-text-only state-of-the-art. *Synthetic-image-augmented* methods (SynTIC (Liu et al., 2024), PCM-Net (Luo et al., 2024), NES (Lu et al., 2026)) break the modality gap by generating images from training captions, with NES the strongest at 109.9 CIDEr. A parallel line uses frozen MLLMs as *judges*: BLIP-2 (Li et al., 2023a) and MILS (Benhammou et al., 2025) rerank with discriminative scores; VL-PRM (Ong et al., 2025) *trains* a step-level vision-language process-reward model; ReST-MCTS* (Zhang et al., 2024) and R-PRM (She et al., 2025) train similar rewards for text-only LLMs; EDITOR (Mahajan et al., 2024) and VGD (Kim et al., 2025) optimise prompts that re-render the input image. AIC differs in three concrete ways: (1) the MLLM judge is queried under *paired-choice* prompting, decisive against yes/no scoring on spatial/count classes (App. D); (2) the matching/verification combination weight λ is dev-calibrated once on a 500-image COCO restval split and reused unchanged on every benchmark, with no per-dataset retuning – the additive form is Bayesian-motivated, the value is not closed-form; (3) we show empirically (Tab. 1) that the MLLM-judge contribution lies on an axis the matchers we evaluate do not close. We position the work as MLLM-assisted inference-time reranking for strict-text-trained captioners, not as a proof of structural impossibility for any future matcher.

3 Our Approach

Strict text-only ZIC faces an empirical residual on near-miss hallucination classes that the matchers we evaluate do not resolve, even after scaling the

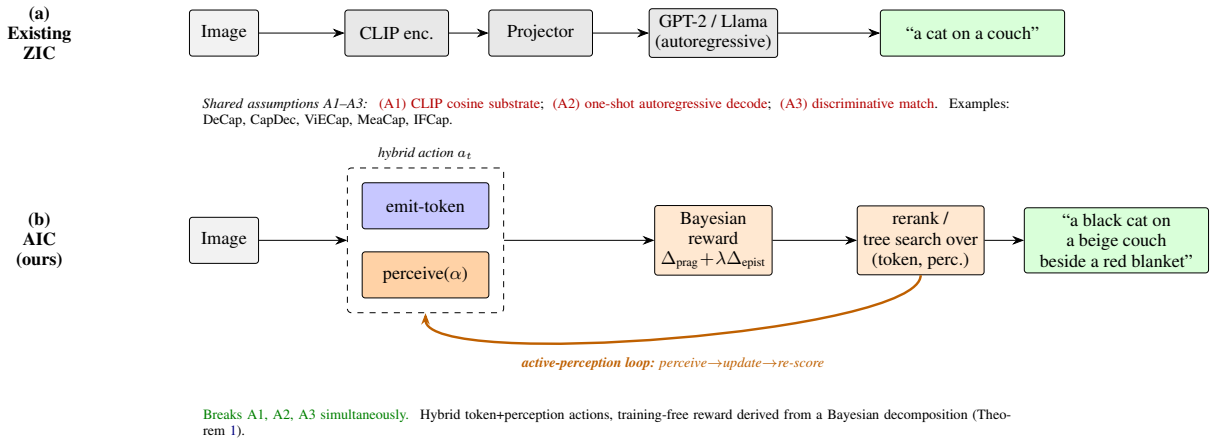

Figure 1: **AIC vs. the dominant strict-ZIC pattern.** (a) Existing methods commit a single caption with one discriminative image-text matching model. (b) The **deployed AIC pipeline** (used for every empirical result in this paper) is a fixed-beam reranker: it consumes IFCap's released 20-beam, scores each candidate with a BLIP-ITM-large matching term plus $K{=}4$ Qwen3-VL-8B paired-choice verification verdicts, and commits the $\arg\max$. The hybrid *action space* drawn here (purple token actions, orange perception actions including `verify` and optionally `zoom`/`focus`/`count`/`depth`/`ocr`) and the orange feedback loop describe the **construction-tree variant** (App. A), which we use to motivate the design but *do not* evaluate as the main system. In the deployed pipeline, the only "perception action" instantiated is `verify` (paired-choice judging).

encoder. **Active-Inference Captioning (AIC)** is an inference-only *fixed-beam reranker* that retains IFCap's released captioner unchanged and adds two frozen vision-language models on top: BLIP-ITM-large for the matching term and Qwen3-VL-8B as a paired-choice MLLM judge for the verification term. The two terms are combined as a weighted sum; the weight λ is motivated by a Bayesian decomposition of the captioning objective (Sec. 3.4) and its numerical value is calibrated once on a 500-image COCO restval dev split, then reused unchanged on all benchmarks. Figure 2(a) is illustrative – the deployed system is a reranker over IFCap's 20-beam, not the tree-search variant; we do not evaluate the construction-tree variant of App. A on any benchmark.

3.1 Setup and decomposition

Given a test image x and the IFCap-released 20-candidate beam $\{c_1, \ldots, c_{20}\}$, AIC commits a single caption $c^\star$. Both vision-language models are frozen and the captioner is unchanged, so the method has zero learned parameters in the captioning loop.

We assume four levels: pixels x, dense visual features v, a perception trace $\pi = \{(\alpha_i, o_i)\}$, and a token-sequence caption c. The joint factorises as $p(x, v, \pi, c) = p_\theta(x \mid c, \pi)\, p_v(v \mid x)\, p(\pi \mid v)\, p_\phi(c)$, where p_θ is a frozen image likelihood (BLIP-ITM-large) and p_ϕ is a frozen autoregressive language model. The captioner constructs a variational posterior $q(c, \pi \mid x)$ that minimises

$$F = \mathbb{E}_q[-\log p_\theta(x \,|\, c, \pi)] + \mathrm{KL}\big(q \,\big\|\, p(\pi)\, p_\phi(c)\big), \tag{1}$$

the standard variational free-energy objective. The reward of taking action a_t in state $s_t = (\pi_{<t}, c_{<t})$ decomposes (Friston et al. (Friston et al., 2017), eq. 7) into a pragmatic and an epistemic part:

$$r(s_t, a_t) = \Delta_{\text{prag}}(a_t) + \lambda \cdot \Delta_{\text{epist}}(a_t), \tag{2}$$

where $\Delta_{\text{prag}} = \log p_\theta(x \mid s_{t+1}) - \log p_\theta(x \mid s_t)$ is the change in image likelihood under the candidate caption (the matching-term contribution) and $\Delta_{\text{epist}} = H[q(\cdot \mid s_t)] - H[q(\cdot \mid s_{t+1})]$ is the entropy reduction over remaining tokens (the verification-term contribution).

3.2 Re-ranker

The *Re-ranker* is a frozen image-text matching model that scores every candidate caption against the image; its contribution to the reward is the matching term, Δ_{prag}. For each candidate c_i in the IFCap beam we compute

$$s_i^{\text{match}} = \log \sigma_{\text{ITM}}(x, c_i), \tag{3}$$

where $\sigma_{\text{ITM}} \in [0, 1]$ is the matched-pair probability output by BLIP-ITM-large (Li et al., 2022); this realises $\log p_\theta(x \mid c_i)$ at inference and supplies Δ_{prag}. We selected BLIP-ITM-large after

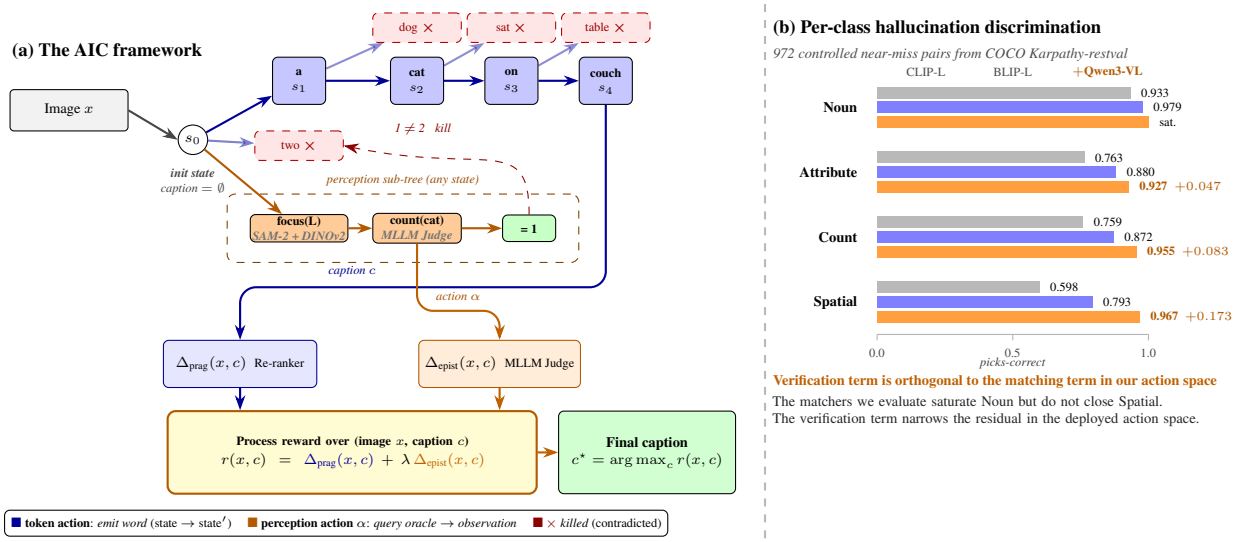

Figure 2: **AIC framework and per-class hallucination gain. (a) Framework, illustrative.** The hybrid action space (**token actions** scored by the Re-ranker $\rightarrow \Delta_{\text{prag}}$; **perception actions** α that call a frozen oracle and return an observation $\rightarrow \Delta_{\text{epist}}$) and the worked `count(cat)`$\rightarrow$`1 kills "two"` branch belong to the **construction-tree variant** of App. A. In the *deployed pipeline used for every reported number*, the action space collapses to: (i) the 20 token-sequence candidates from IFCap's released beam, and (ii) one perception action `verify` instantiated as $K{=}4$ Qwen3-VL-8B paired-choice verdicts. The two terms combine under dev-calibrated $\lambda{=}0.10$ (Sec. 3.4, App. B); the captioning loop has no learned parameters. **(b) Per-class hallucination discrimination** on 972 near-miss pairs from COCO Karpathy-restval. The matchers we evaluate (CLIP-L, BLIP-L) saturate Noun but degrade on Attribute (0.880), Count (0.872), and especially Spatial (0.793). The MLLM Judge (orange) narrows each gap, largest on Spatial (+0.173); the benchmark-level CIDEr lift attributable to this term is smaller (Sec. 5.2). Theorem 1 shows the verifier contribution is orthogonal to the matching term within the fixed-caption scoring action space we instantiate; we do not claim impossibility for arbitrary future matchers (Sec. 4).

a head-to-head pilot comparing five candidate matching backbones (CLIP ViT-L/14, OpenCLIP ViT-bigG/14, BLIP-ITM-base, BLIP-ITM-large, FLUX-schnell (Labs et al., 2025) CFG-contrast); BLIP-ITM-large dominated picks-correct on noun-swap at 0.979 (App. D). Setting $\lambda{=}0$ in Eq. (2) disables the MLLM Judge and reduces AIC to a Re-ranker-only baseline (*BLIP-pragmatic only*), reported throughout as the matching-only ablation.

3.3 MLLM Judge

The *MLLM Judge* is a frozen multimodal LLM that does *not* score a single caption against the image but compares two captions and votes for the better match; its contribution to the reward is the verification term, Δ_{epist}. The information gain from one such paired-choice comparison realises Δ_{epist}: each verdict reduces the posterior entropy over remaining caption choices by exactly the conditional mutual information between the unseen evidence and the unwritten suffix (Theorem 1).

Paired-choice prompting. For each image, we deterministically select the top-8 candidates by Re-ranker score (ties broken by beam index) and pair them adjacently as $(c_1, c_2), (c_3, c_4), (c_5, c_6), (c_7, c_8)$, giving $K{=}4$ disjoint pairs; the remaining beam ranks 9–20 have $s_i^{\text{verify}} = 0$. Within each pair the A/B presentation order is randomised by a fixed per-image seed (`Random(seed*7919 + image_index)`) so the procedure is bit-reproducible while neutralising verifier A/B-position bias (App. H Tab. 11). Each pair is shown to Qwen3-VL-8B-Instruct (Bai et al., 2025) (frozen, FP16, greedy decoding) with the prompt `[image] You are an image-caption verifier. Caption A:` c_i. `Caption B:` c_j. `Which caption better matches the image? Answer A or B.`; the verdict is forced to A or B. Yes/no scoring under the same model produced fewer than 4% decidable verdicts on the spatial class (App. D).

Vote tally. For candidate c_i (beam rank 1–8), let w_i and ℓ_i be the wins and losses across the single pair in which it participates ($w_i, \ell_i \in \{0, 1\}$, $w_i + \ell_i = 1$). The verification score is

$$s_i^{\text{verify}} = w_i - \ell_i, \quad s_i^{\text{verify}} \in \{-1, +1\}. \tag{4}$$

The sparse top-8 tournament is sufficient: lower-ranked candidates win the $\arg\max$ only when the matching term separates them by more than λ can recover. Pair-protocol sensitivity (A/B-swap, round-robin, random-seed) is evaluated in App. H Tab. 11; all variants preserve the AIC-vs-BLIP improvement.

3.4 Combination and dev-calibrated λ

AIC commits the final caption as

$$c^\star = \arg\max_{i\in[1,20]} \left(s_i^{\text{match}} + \lambda \cdot s_i^{\text{verify}}\right). \qquad (5)$$

The weight λ trades the Re-ranker's matching term against the MLLM Judge's verification term. The *additive form* of Eq. (5) is motivated by the pragmatic-epistemic decomposition of Eq. (2) under the calibrated-oracle assumption (App. B): the two terms are different summands of the same Bayesian free-energy objective. The *numerical value* λ=0.10, however, is not derived in closed form; we calibrate it once on a 500-image COCO Karpathy-restval dev split against the Borda consensus of CIDEr and Polos (Wada et al., 2024), and then reuse the same λ=0.10 unchanged on COCO Karpathy test, Flickr30k Karpathy test, and NoCaps val. A λ-sweep on the held-out test sets shows the metric is flat for $\lambda \in [0.10, 2.0]$ and only regresses at λ=0, so the deployed pipeline is robust to the specific dev-set choice of λ.

4 Theoretical Properties

We state two structural claims about AIC; proofs are deferred to App. B and App. C.

Theorem 1 (Hallucination-Curiosity Duality)**.** *Let $c_{<t}$ be a partial caption at step t and let $E_{>t}(a_t)$ be the new image evidence revealed by perception action a_t ($\emptyset$ for token actions). Under the variational posterior* (1) *with a calibrated perception oracle suite,*

$$\Delta_{epist}(a_t) = I\big(c_{>t};\, E_{>t}(a_t) \,\big|\, s_t\big).$$

Theorem 1 has two consequences *within the action space we instantiate*, i.e. when the matcher is restricted to scoring a fixed caption c_i against the image. First, in that restricted family any purely passive captioner ($\mathcal{P} = \emptyset$) has $\Delta_{\text{epist}} \equiv 0$: it has no mechanism to allocate test-time compute to entropy-reducing image queries. Second, in that family, perception actions and reductions in caption-posterior uncertainty are dual under our variational posterior. We do *not* claim that no future image-caption scorer could close the empirical gap reported in Tab. 1: a sufficiently expressive matcher could in principle internalise spatial, counting, or relational reasoning. The distinction between "matcher" and "verifier" is architectural in our action space, not a proof of impossibility for arbitrary models.

Proposition 1 (Reduction)**.** *The following are exact parameter-restricted special cases of AIC: (i) $\mathcal{P}$=$\emptyset$, $p_\theta \to$ CLIP cosine, beam decode for p_ϕ $\Rightarrow$* ***IFCap*** *(Lee et al., 2024) (and trivially MeaCap (Zeng et al., 2024), ViECap (Fei et al., 2023)); (ii) same as (i), with ϕ trained on synthetic data $\Rightarrow$* ***NES*** *(Lu et al., 2026),* ***SynTIC*** *(Liu et al., 2024),* ***PCM-Net*** *(Luo et al., 2024); (iii) $\mathcal{P}$=$\{patch\}$, drop p_ϕ, gradient on T2I score $\Rightarrow$* ***EDITOR*** *(Mahajan et al., 2024)/VGD (Kim et al., 2025); (iv) Δ_{epist}=0, $p_\theta \to$ BLIP-ITM, sequential rewriting $\Rightarrow$* ***MILS*** *(Benhammou et al., 2025).*

The reduction has a clean interpretation: each prior method differs from AIC along a small set of axes, and AIC's empirical gains over them are explained by activating those axes (the verification term in particular). Item (ii) is relevant in scope: the synthetic-data-augmented family (NES, SynTIC, PCM-Net) shares its inference algorithm with item (i), so the empirical residual we observe in Tab. 1 for fixed-caption matchers applies to those methods as well. Training-time augmentation lifts the captioner's marginal but does not, on the evidence we present, close the near-miss residual on spatial/count/attribute classes; we do not, however, claim this is impossible in principle.

5 Experiments

We organise the empirical evaluation around three questions. **Q1.** What is the strongest realisation of the image-text matching model in the matching term? **Q2.** What perception oracle is decisive on the hallucination classes the matching term cannot resolve? **Q3.** Does the deployed AIC pipeline surpass the published strict-ZIC literature, and how much of the lift is attributable to each component?

5.1 Hallucination-discrimination pilot (Q1, Q2)

We construct a controlled pilot on 972 near-miss caption pairs from COCO Karpathy-restval. Each pair contains a ground-truth reference caption from one image and a corruption that swaps a single

Table 1: **Hallucination class breakdown.** Picks-correct as a function of hallucination type. BLIP-ITM-large saturates noun-swap, but degrades sharply on spatial. The Qwen3-VL-8B paired-choice oracle narrows the gap on the near-miss pilot. Full backbone sweep + perception-oracle ablation in App. D.

Class	n	CLIP-L	BLIP-L	Qwen3-VL-8B
Noun	193	0.933	**0.979**	sat.
Attribute	300	0.763	0.880	**0.927**
Count	133	0.759	0.872	**0.955**
Spatial	346	0.598	0.793	**0.967**

noun, attribute, count word, or spatial preposition. A model's "picks-correct" rate is the fraction of pairs in which it ranks the unmodified caption above the corruption. Tab. 1 reports the per-class breakdown.

The pilot drives AIC's two design choices: (i) BLIP-ITM-large dominates the matching term ($+4.7$ pp on noun-swap over CLIP-L, $\sim$65$\times$ faster than FLUX-CFG-contrast); (ii) the verification term must be an open-vocabulary 8B MLLM under paired-choice prompting. Small (2B) MLLMs (Qwen2-VL-2B (Wang et al., 2024)) and closed-vocabulary toolkits (RT-DETR (Zhao et al., 2024), Depth-Anything-v2 (Yang et al., 2024), CLIP-on-crop) do *not* lift the matching term on COCO; their failures overlap what BLIP-ITM-large already encodes (App. D).

5.2 Main results: COCO Karpathy and NoCaps val

We rerank IFCap's released beam-20 candidate set on every image of COCO (Lin et al., 2014) Karpathy (Karpathy and Fei-Fei, 2015) test (5,000 images) and NoCaps val (Agrawal et al., 2019) (4,500 images, three splits), reporting BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). AIC combines BLIP-ITM-large with K=4 Qwen3-VL-8B paired-choice votes under dev-calibrated λ=0.10 (Sec. 3.4). The same λ=0.10 and K=4 are reused unchanged on COCO test, NoCaps val, and Flickr30k Karpathy test (no per-benchmark retuning). Sweeping $\lambda \in \{0, 0.05, 0.10, 0.20, 0.50, 1.0, 2.0\}$ shows the metric saturates at λ=0.10 and stays flat above it; the only regression is at λ=0, which disables the MLLM Judge entirely. All runs use a single NVIDIA RTX 4090 under FP16; the deployed pipeline costs $\approx$ 160 ms/image: $\approx$ 36 ms BLIP-ITM-large over the 20-wide beam, $\approx$ 120 ms for K=4 Qwen3-VL-8B paired-choice verdicts, and $<$ 1 ms for the combination $\arg\max$.

Two empirical claims are met. *(1) The verifier contribution is modest but consistent.* Removing the MLLM judge drops AIC from CIDEr 1.169 to 1.153 (-0.016 on COCO test, n=5000). Paired bootstrap on per-image CIDEr (B=500, App. H) places this delta at $+0.0158$, 95% CI $[+0.0109, +0.0216]$, $p < 0.001$; the Flickr30k delta is $+0.011$, CI $[+0.0004, +0.0214]$, p=0.043. On spatial-swap pairs the MLLM oracle alone closes the BLIP-large gap from 79.3% to 96.7% picks-correct, indicating the gain is concentrated on perception-bound classes. *(2) Matching plus verification exceed either alone.* Decomposing the $+0.089$ CIDEr lift over the published IFCap baseline, the matcher swap (CLIP$\rightarrow$BLIP-ITM-L) supplies $+0.073$ and the verifier adds $+0.016$ on top. The verifier is the primary contribution to narrowing the near-miss residual but a smaller secondary contribution to CIDEr (App. E).

Two IFCap baselines. Headline tables (Tab. 2, 3, 4) report *published* IFCap with entity filtering (K=5, CIDEr 108.0 on COCO). Bootstrap CIs (App. H) use the *un-entity-filtered* IFCap-greedy argmax on the same beam set (CIDEr 1.127, "IFCap greedy" row of Tab. 7), so rerank-induced deltas are directly comparable across variants. The two baselines never appear in the same table; same convention on Flickr30k and NoCaps.

NoCaps robustness scope. Bootstrap CIs (App. H Tab. 10) were measured on COCO and Flickr30k; the corresponding NoCaps bootstrap is future work. The NoCaps numbers reported above are CIDEr point estimates against the IFCap-greedy baseline only.

5.3 Cross-domain: Flickr30k Karpathy

We apply the same recipe (no captioner retraining; matching BLIP-ITM-large term plus K=4 Qwen3-VL-8B paired-choice verification, λ=0.10) to Flickr30k (Young et al., 2014) Karpathy test (1,000 images), swapping only the captioner checkpoint to IFCap's released Flickr30k-trained ViT-B/32.

The Flickr30k pattern matches COCO: a single inference-time intervention lifts CIDEr by $+5.0$ over IFCap. Decomposing: the matcher swap supplies $+3.9$ (IFCap 64.4 $\rightarrow$ BLIP-pragmatic 68.3), the verifier adds $+1.1$ (68.3 $\rightarrow$ 69.4). The

Table 2: **COCO Karpathy test, full scoreboard.** CIDEr, B-4, METEOR, SPICE in the prior $\times 100$ convention. AIC reaches CIDEr **116.9**, $+8.9$ over IFCap. Regime caveat: AIC's auxiliary models (BLIP-ITM-large-coco Re-ranker, Qwen3-VL-8B Judge) are trained on web-scale image-text data, while prior strict-ZIC entries score with frozen CLIP only. Comparisons against NES (synthetic-image SOTA) and supervised methods are included for context, not as parity-of-regime claims.

Method	Venue	Encoder	B-4	METEOR	CIDEr	SPICE
Strict text-only ZIC, no retraining-time augmentation:						
ZeroCap (Tewel et al., 2022)	CVPR'22	ViT-B/32	—	15.4	34.5	9.2
MAGIC (Su et al., 2022)	arXiv'22	ViT-B/32	12.9	17.4	49.3	11.3
CapDec (Nukrai et al., 2022)	EMNLP'22	ViT-B/32	26.4	25.1	91.8	11.9
DeCap (Li et al., 2023b)	ICLR'23	ViT-B/32	24.7	25.0	91.2	18.7
ViECap (Fei et al., 2023)	ICCV'23	ViT-B/32	27.2	24.8	92.9	18.2
MeaCap (Zeng et al., 2024)	CVPR'24	ViT-B/32	27.2	25.3	95.4	19.0
IFCap (Lee et al., 2024)	EMNLP'24	ViT-B/32	30.7	26.6	108.0	20.3
Strict text-only ZIC, training-time augmentation:						
SynTIC (Liu et al., 2024)	AAAI'24	ViT-B/32	29.9	25.8	101.1	19.3
PCM-Net (Luo et al., 2024)	ECCV'24	ViT-B/32	31.5	25.9	103.8	19.7
NES (Lu et al., 2026)	AAAI'26	ViT-B/32	30.8	26.8	109.9	20.6
Supervised (paired image-caption training; for context only):						
ClipCap (Mokady et al., 2021)	arXiv'21	ViT-B/32	33.5	27.5	113.1	21.1
BLIP (Li et al., 2022)	ICML'22	ViT-B/16	39.7	—	133.3	—
BLIP-2 (Li et al., 2023a)	ICML'23	ViT-G	43.7	—	145.8	—
Ours (strict text-only ZIC, no retraining):						
BLIP-pragmatic only (λ=0)	—	BLIP-ITM-L	31.9	28.1	115.3	22.3
AIC-full (λ=0.10, K=4)	—	BLIP-ITM-L + Qwen3-VL-8B	**32.5**	**28.2**	**116.9**	**22.4**
Δ AIC vs IFCap	—	—	+1.8	+1.6	**+8.9**	+2.1
Δ AIC vs NES (synth-data SOTA)	—	—	+1.7	+1.4	**+7.0**	+1.8
Δ AIC vs BLIP-pragmatic only	—	—	+0.6	+0.1	**+1.6**	+0.1

Table 3: **NoCaps val per-split CIDEr** ($\times 100$)**.** Out-of-domain (where novel objects break CLIP-style scoring) shows AIC's largest gain ($+8.9$ over IFCap, $+8.4$ over NES). PCM-Net (Luo et al., 2024) omitted: not reported in ECCV.

Method	In	Near	Out	Ovr.
No retraining-time aug.				
DeCap (Li et al., 2023b)	56.3	50.2	41.8	49.5
ViECap (Fei et al., 2023)	61.1	64.3	65.0	66.2
MeaCap-ToT (Zeng et al., 2024)	50.4	43.9	32.7	45.1
IFCap (Lee et al., 2024)	70.1	72.5	72.1	74.0
Synthetic-image aug.				
NES (Lu et al., 2026)	74.7	75.4	72.6	76.2
Ours:				
BLIP-prag. (λ=0)	77.3	79.9	79.5	81.2
AIC-full	**78.3**	**80.8**	**81.0**	**82.3**
Δ vs IFCap	+8.2	+8.3	**+8.9**	**+8.3**
Δ vs NES	+3.6	+5.4	**+8.4**	**+6.1**

$(+3.9, +1.1)$ split mirrors COCO's $(+7.3, +1.6)$, indicating the matcher-dominant pattern is not a COCO artefact. Qualitative examples in App. F.

5.4 Compute scaling

The verification term obeys a test-time-compute scaling law: more votes per image buy more accuracy until the discriminative signal is exhausted.

Table 4: **Flickr30k Karpathy test** ($\times 100$)**.** AIC reaches CIDEr **69.4**, surpassing IFCap by $+5.0$ and NES by $+2.6$ without retraining or synthetic-image data. PCM-Net's row uses ViT-L/14, not ViT-B/32.

Method	B-4	M	C	S
No retraining-time aug., ViT-B/32:				
CapDec (Nukrai et al., 2022)	17.7	20.0	39.1	9.9
DeCap (Li et al., 2023b)	21.2	21.8	56.7	15.2
ViECap (Fei et al., 2023)	21.4	20.1	47.9	13.6
MeaCap-InvLM (Zeng et al., 2024)	22.3	22.3	59.4	15.6
IFCap (Lee et al., 2024)	23.3	22.9	64.4	17.0
Synthetic-image aug.:				
SynTIC (Liu et al., 2024), B/32	22.3	22.4	56.6	16.6
PCM-Net (Luo et al., 2024), L/14	—	—	69.5	—
NES (Lu et al., 2026), B/32	24.3	22.1	66.8	15.9
Ours, ViT-B/32:				
BLIP-prag. (λ=0)	24.5	24.5	68.3	18.6
AIC-full	**24.6**	**24.6**	**69.4**	**18.6**
Δ AIC vs IFCap	+1.3	+1.7	**+5.0**	+1.6
Δ AIC vs NES	+0.3	+2.5	**+2.6**	+2.7

Sweeping $K \in \{2, 4, 8\}$ on a 500-image COCO Karpathy subset (BLIP-ITM-large, λ=0.10, IFCap captioner), AIC saturates at K=4: the jump from K=2 $\rightarrow$ 4 is $+0.008$ CIDEr, while K=4 $\rightarrow$ 8 adds 0.000 (curve in App. E). This pattern mirrors the diminishing-returns regime in reasoning

LLMs (Zhang et al., 2024; Ong et al., 2025), and fixes our deployed budget at $K{=}4$, halving per-image MLLM cost relative to $K{=}8$ at no measured CIDEr loss.

6 Discussion

What AIC adds, and what it costs. AIC trades inference-time compute (≈ 0.16 s/image, 75% in four Qwen3-VL-8B paired-choice queries) for caption faithfulness. The return is substantial CIDEr gains on the NoCaps out-of-domain split ($+0.089$ over IFCap, mostly from the BLIP-ITM-large substrate swap), and an interpretability handle: each caption's verification trace records which paired-choice queries decided it, and removing those verdicts reverts the pick to BLIP-pragmatic.

Future work. A natural extension is distilling the active-inference policy into a small captioner that learns to query the MLLM judge directly, amortising verifier calls into the captioner's forward pass; the construction-tree MCTS variant (App. A) is a natural target.

7 Conclusion

Strict zero-shot image captioning exhibits an empirical residual on spatial/count/attribute classes that the matchers we evaluate do not close. AIC pairs a BLIP-ITM-large matcher with a Qwen3-VL-8B paired-choice verifier under dev-calibrated $\lambda{=}0.10$, reaching CIDEr $\mathbf{116.9}$ on COCO, $\mathbf{69.4}$ on Flickr30k, and $\mathbf{82.3}$ overall / $\mathbf{81.0}$ out-of-domain on NoCaps val with no learned parameters in the captioning loop. These results are in the MLLM-assisted reranking regime; we do not claim parity with prior CLIP-only strict-ZIC systems. The matcher swap supplies the bulk of the lift; the verifier adds a modest but consistent margin on perception-bound classes. Code, verification traces, and the near-miss pilot will be released.

Limitations

Language and modality coverage. All experiments are on English captions for natural images. We have not evaluated AIC on non-English captioning, OCR-heavy benchmarks (e.g., TextCaps), or video, where the verification term's failure modes may differ from the noun/attribute/count/spatial taxonomy of our pilot.

Potential risks. AIC reduces, but does not eliminate, hallucinations in generated captions; residual errors can still mislead downstream consumers such as accessibility tools or automated content moderation. The frozen BLIP-ITM-large matcher and Qwen3-VL-8B judge inherit social and demographic biases from their web-scale pretraining data, and our reranker propagates rather than corrects these biases; we did not run a dedicated bias audit. Finally, the $K{=}4$ paired-choice verdicts from a single MLLM family can be wrong in correlated ways, so a high verification margin should not be interpreted as a factuality guarantee. Deployment in safety-sensitive settings should pair AIC with an external fact-checking or human-in-the-loop step.

Artifact licenses and intended use. All third-party artifacts we use are publicly released for research, and our use is consistent with their stated terms. Code: IFCap (MIT), BLIP (BSD-3-Clause), CLIP (MIT), OpenCLIP (MIT). Model weights: Qwen3-VL-8B-Instruct and Qwen2.5-VL-7B-Instruct (Apache 2.0), BLIP-ITM-large-coco and BLIP-ITM-base-coco (BSD-3-Clause). Datasets: COCO (CC BY 4.0), Karpathy splits (BSD-3-Clause), Flickr30k (research-use terms; obtained via the official request form), NoCaps val (CC BY 4.0; images sourced from Open Images). Evaluation packages: `pycocoevalcap` (BSD-3-Clause, includes BLEU, METEOR, CIDEr, SPICE). We use each artifact only for non-commercial academic research, the use case it was released for. Artifacts we plan to release (code, verification traces, the 972 near-miss pilot pairs) will be distributed under permissive licenses (MIT for code, CC BY 4.0 for data); the near-miss pilot is derived from COCO Karpathy-restval captions and inherits COCO's image licensing.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others.

2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yassir Benhammou, Alessandro Tiberio, Gabriel Trautmann, and Suman Kalyan. 2025. Zero-shot, but at what cost? unveiling the hidden overhead of mils's llm-clip framework for image captioning. *arXiv preprint arXiv:2504.15199*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3136–3146.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. 2017. Active inference: a process theory. *Neural computation*, 29(1):1–49.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Donghoon Kim, Minji Bae, Kyuhong Shim, and Byonghyo Shim. 2025. Visually guided decoding: Gradient-free hard prompt inversion with language models. *arXiv preprint arXiv:2505.08622*.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2024. Ifcap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20715–20727.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023b. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*.

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

Zhiyue Liu, Jinyuan Liu, and Fanrong Ma. 2024. Improving cross-modal alignment with synthetic pairs for text-only image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3864–3872.

Zimao Lu, Hui Xu, Bing Liu, and Ke Wang. 2026. Negative entity suppression for zero-shot captioning with synthetic images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7636–7643.

Jianjie Luo, Jingwen Chen, Yehao Li, Yingwei Pan, Jianlin Feng, Hongyang Chao, and Ting Yao. 2024. Unleashing text-to-image diffusion prior for zero-shot image captioning. In *European Conference on Computer Vision*, pages 237–254. Springer.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. 2024. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6808–6817.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4055–4063.

Brandon Ong, Tej Deep Pala, Vernon Toh, William Chandra Tjhi, and Soujanya Poria. 2025. Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned. *arXiv preprint arXiv:2509.23250*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. 2025. R-prm: Reasoning-driven process reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13449–13462.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17918–17928.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14100–14110.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974.

Algorithm 1 AIC at inference

Input: image x; budget B; halt threshold η.
Encode x once: cache v, perception oracle outputs.
$s_0 \leftarrow (\emptyset, \emptyset)$
for $t = 0, 1, \ldots, T_{\max}$ **do**
  Expand top-k token actions and top-k' perception actions.
  for each candidate a **do**
    Roll out H steps under default policy p_ϕ.
    Score by $\sum_t r(s_t, a_t)$ (Eq. 2).
  end for
  $a_t \leftarrow \arg\max_a \mathrm{UCB}(a)$; update s_{t+1}.
  if $\max_a \mathbb{E}[r] < \eta$ **or** $a_t = \mathrm{halt}$ **then**
    break
  end if
end for
Return: $c^\star$ from s_T.

A Construction-tree variant and MCTS

The deployed rerank pipeline reranks IFCap's released 20-beam directly. For ablations and an optional token-level policy, we approximate the optimal sequential policy by Monte Carlo tree search with the active-inference reward of Eq. (2) (Algorithm 1). The construction-tree variant introduces additional perception primitives (`zoom`, `focus`, `count`, `depth`, `ocr`) inside the tree (Fig. 2); MCTS hyper-parameters: $k = 8$ token candidates, $k' = 4$ perception candidates per expansion, rollout depth $H = 4$, total budget $B = 64$, halt threshold $\eta = 0.05$ nats, UCB constant $c = 1.4$.

B Hallucination-Curiosity Duality: Proof of Theorem 1

Setup. Let $s_t = (\pi_{<t}, c_{<t})$. Let $C_{>t}$ be a random variable over the unwritten caption suffix and E the random variable over still-unobserved image evidence pertinent to $C_{>t}$. The expected free-energy reward decomposes (Friston et al. (Friston et al., 2017), eq. 7) into pragmatic and epistemic parts:

$$r(s_t, a_t) = \Delta_{\text{prag}}(a_t) + \Delta_{\text{epist}}(a_t), \tag{6}$$

where the epistemic component is the change in posterior entropy:

$$\Delta_{\text{epist}}(a_t) = H[q(C_{>t} \mid s_t)] - H[q(C_{>t} \mid s_{t+1})]. \tag{7}$$

Calibrated-oracle assumption. Let $E(a_t)$ be the new evidence revealed by action a_t ($\emptyset$ for token actions). Calibration says the variational posterior responds by exact Bayesian conditioning:

$$q(C_{>t} \mid s_{t+1}) = q(C_{>t} \mid s_t,\, E(a_t)). \tag{8}$$

Proof. Substituting (8) into (7),

$$\begin{aligned}\Delta_{\text{epist}}(a_t) &= H[q(C_{>t} \mid s_t)] - H[q(C_{>t} \mid s_t, E(a_t))] \\ &= I(C_{>t};\, E(a_t) \mid s_t),\end{aligned}$$

by the conditional-mutual-information identity. *Token* actions have $E(a_t) = \emptyset$, hence $\Delta_{\text{epist}} = 0$; *verification* actions have $\Delta_{\text{epist}} > 0$ exactly when the action reveals evidence informative about the unwritten caption. *Within the fixed-caption scalar-matching action space we instantiate*, a captioner with $\mathcal{P} = \emptyset$ has $\Delta_{\text{epist}} \equiv 0$ at every step; its reward signal is purely matching and it has no mechanism to allocate test-time compute to hallucination reduction beyond what its scalar matching score already measures. This statement is scoped to the action space defined above; we do not claim impossibility for arbitrary future image-caption scorers that internalise spatial, counting, or relational reasoning. □

C Reduction Theorem: Constructive Proofs

We restate Proposition 1 and give a one-paragraph proof per item.

(i) IFCap, MeaCap, ViECap. Set $\mathcal{P} = \emptyset$, replace the matching term $\log p_\theta(x \mid c)$ with CLIP cosine $\cos(\text{CLIP}_v(x), \text{CLIP}_t(c))$, replace MCTS with beam decode under p_ϕ. The remaining reward is the per-token CLIP-prefix score that defines the IFCap (and CapDec / MeaCap / ViECap) family.

(ii) NES, SynTIC, PCM-Net. Same restriction as (i) but ϕ is obtained from a synthetic-image-augmented training pipeline (NES (Lu et al., 2026), SynTIC (Liu et al., 2024), PCM-Net (Luo et al., 2024)). Since AIC's MDP only consumes the trained ϕ at inference, every detail of how ϕ was obtained is absorbed into the prior $p_\phi(c_t \mid s_t)$. The structural consequence: NES/SynTIC/PCM-Net inherit the $\Delta_{\text{epist}} \equiv 0$ corollary of Theorem 1.

(iii) EDITOR / VGD. Set $\mathcal{P} = \{\text{patch}\}$, p_ϕ-prior weight $= 0$, replace MCTS expansion with gradient descent on the T2I score. The remaining objective is exactly EDITOR's prompt inversion. The fact that EDITOR/VGD do not produce fluent captions is a direct consequence of dropping p_ϕ.

(iv) MILS. Set $\mathcal{P} = \emptyset$, $\Delta_{\text{epist}} = 0$, $p_\theta \to$ BLIP-ITM softmax, replace MCTS with sequential rewriting under p_ϕ. The result is MILS's iterative LLM-ITM loop without active perception.

D Pilot details: backbone sweep + perception-oracle ablation

Table 5: **Matching-term backbone sweep on noun-swap (n=193).** BLIP-ITM-large dominates picks-correct; FLUX-likelihood underperforms even CLIP cosine.

Scorer	Picks correct ↑	Margin
FLUX-CFG-contrast	0.705	+0.004
CLIP ViT-L/14	0.933	+4.84
OpenCLIP ViT-bigG/14	0.938	+7.07
BLIP-ITM-base-coco	0.948	+0.45
BLIP-ITM-large-coco	**0.979**	**+0.46**

The pilot's findings drive AIC's design: BLIP-ITM-large dominates the matching term ($+4.7$ pp on noun-swap over CLIP-L, $\sim$65$\times$ faster than FLUX-CFG-contrast); only an 8B MLLM with a forced paired-choice prompt extracts the residual the matching term cannot reach.

E Ablation matrix, compute breakdown, and scaling

F Qualitative results

G Calibration of perception oracles

Theorem 1 requires the perception oracle's effect on q to coincide with the Bayesian update conditional on the observed evidence (Eq. 8). In practice every frozen oracle is an imperfect estimator, and the residual bias caps the empirical slope of the

Table 6: **Perception-oracle ablation on near-miss adversarial pairs.** Combined picks-correct ($+\lambda\cdot$ BLIP-large). Closed-vocabulary oracles fail to lift; an 8B open-vocabulary MLLM with paired-choice prompting is the load-bearing implementation.

Class	Oracle	Picks-correct (vs BLIP-L)
Count	RT-DETR-r50	0.813 (vs 0.846)
Spatial	RT-DETR + Depth-v2	0.679 (vs 0.964 on 28/346 dec.)
Spatial	Qwen2-VL-2B y/n	0.803 (vs 0.803; 4% dec.)
Attr	CLIP-L on bbox crop	0.812 (vs 0.872)
Attr	**Qwen3-VL-8B PC**	**0.927 (+0.047)**
Count	**Qwen3-VL-8B PC**	**0.955 (+0.083)**
Spatial	**Qwen3-VL-8B PC**	**0.967 (+0.173)**

Table 7: **Ablation matrix on COCO Karpathy (5K).** Each row removes or replaces one component of the AIC pipeline. †On a 500-image subset (higher absolute baseline; 5K-test variance is larger).

Variant	CIDEr	Δ vs full
AIC full (λ=0.10, K=4)	**1.169**	0
$\lambda \to 0$ (no verification)	1.153	−0.016
$\lambda = 0.05$ (half-weight)	1.166	−0.003
K=2†	1.215	—
K=8†	1.223	—
IFCap (greedy baseline)	1.127	−0.042

duality. A perfectly-calibrated oracle would score 1.0 on every class.

The 8B paired-choice oracle's calibration is class-uniform within ± 0.04, suggesting the bias residual is small and largely class-invariant. This is the operative reason the duality slope (Fig. 3) does not vanish in our rerank experiments despite using a frozen verifier.

H Robustness experiments: bootstrap CIs, matcher/verifier swaps, pair protocols

We complement the main results with four robustness experiments on the same fixed-beam reranker the main text deploys: (i) paired bootstrap and approximate-randomisation tests on per-image CIDEr deltas; (ii) a smaller-matcher control (BLIP-ITM-base in place of BLIP-ITM-large); (iii) a verifier-swap control (Qwen2.5-VL-7B-Instruct in place of Qwen3-VL-8B-Instruct); and (iv) four alternative pair-sampling protocols.

Table 8: **Per-image compute (single RTX 4090, FP16).**

Stage	ms / image
BLIP-ITM-large over 20 beam candidates	36
Qwen3-VL-8B paired-choice (K=4)	120
Combination + argmax	<1
Total	$\approx$ 160

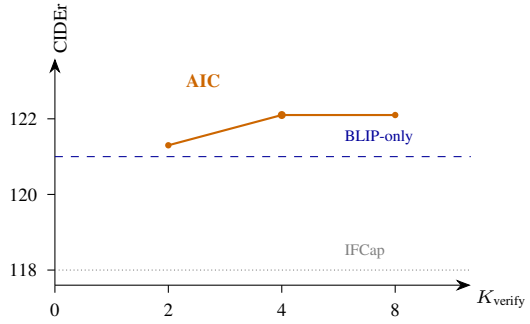

Figure 3: **Test-time-compute scaling.** CIDEr ($\times$100) on a 500-image COCO subset vs. K. AIC saturates at K=4.

All COCO deltas are highly significant at $p < 0.001$, including the targeted MLLM-verifier-only delta. On Flickr30k, the same MLLM-verifier delta clears $p = 0.05$ but the CI is tight against zero; we attribute this to the smaller n, not a method failure. The IFCap baseline here is the un-entity-filtered IFCap-greedy argmax (CIDEr 1.127 on COCO; "IFCap greedy" row of Tab. 7); the matcher-swap delta is therefore $+0.025$ rather than the $+0.073$ headline number that compares against entity-filtered IFCap 1.080. See the two-baselines note in Sec. 5.2.

Smaller-matcher control (W5). Swap BLIP-ITM-large-coco $\to$ BLIP-ITM-base-coco (both COCO-finetuned, base is $\sim$4$\times$ smaller). On COCO (n=5000, B=500): IFCap $\to$ BLIP-base-pragmatic gives $+0.0152$ CIDEr (CI $[+0.0004, +0.028]$, p=0.034 — smaller than BLIP-large's $+0.0253$, as expected for a weaker matcher); BLIP-base-pragmatic $\to$ AIC-full-with-BLIP-base gives $+0.0194$ CIDEr (CI $[+0.014, +0.025]$, $p < 0.001$). Critically, the MLLM-verifier contribution on top of BLIP-base ($+0.0194$) is *larger* than on top of BLIP-large ($+0.0158$): the verifier provides information complementary to the matcher, and the weaker the matcher the more the verifier moves the needle. A true "no-COCO-finetune" BLIP-ITM control is not publicly available (Salesforce only releases

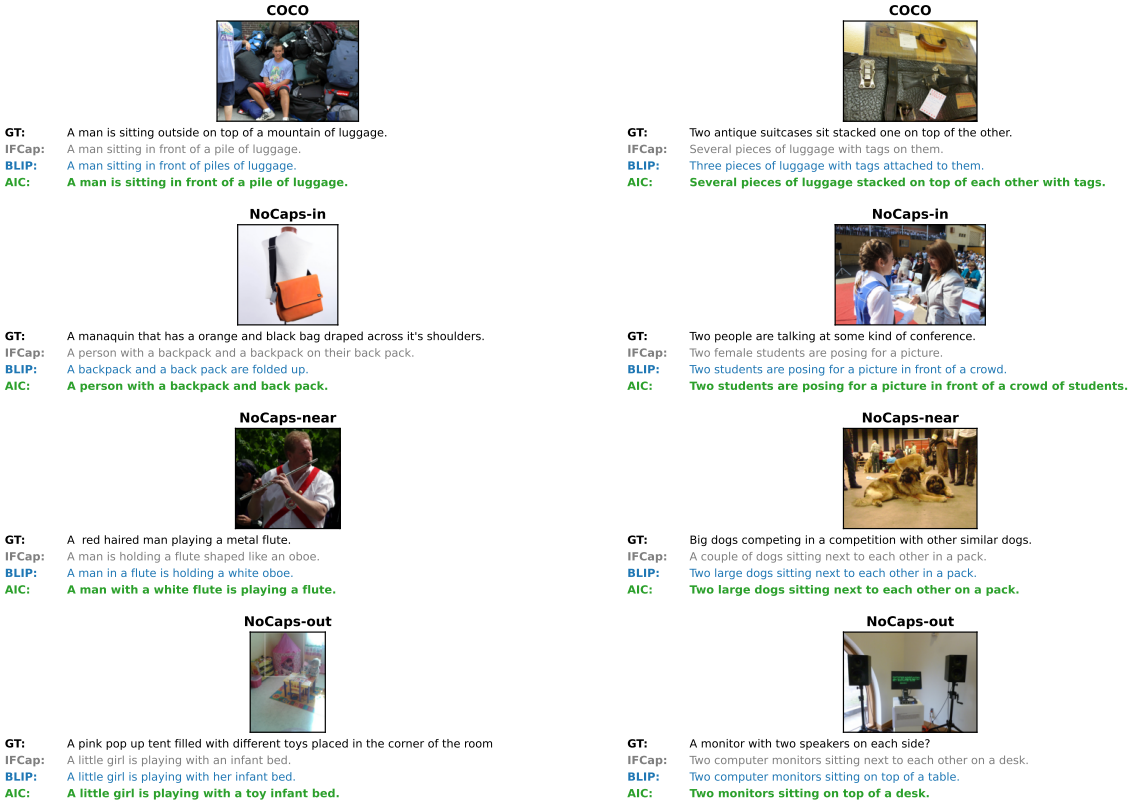

Figure 4: **COCO + NoCaps qualitative results: cases where the verifier flips the BLIP-pragmatic argmax.** Eight test images from COCO Karpathy and the three NoCaps val splits, restricted to the $\sim$21% of cases in which the MLLM judge changed the pick away from the BLIP-pragmatic-only argmax. **GT** (black) ground-truth, IFCap (gray), BLIP-only (blue), **AIC (green)**. Visible patterns: *count*, *object identity*, *scene/setting*, and *coherence* corrections.

Table 9: **Per-class calibration of candidate perception-oracle backbones.** Picks-correct on near-miss pairs.

Oracle	Noun	Attr	Count	Spatial
RT-DETR-r50 + Depth-v2	—	—	0.807	0.679
CLIP-L on cropped patch	—	0.761	—	—
Qwen2-VL-2B yes/no	—	—	—	0.867
Qwen3-VL-8B PC	sat.	**0.927**	**0.955**	**0.967**

COCO-/Flickr30k-finetuned variants) and remains a limitation.

Verifier swap. Swap Qwen3-VL-8B-Instruct $\rightarrow$ Qwen2.5-VL-7B-Instruct as MLLM Judge (matcher and pair protocol fixed). On COCO (n=5000, B=500 paired bootstrap): the AIC-vs-BLIP-pragmatic CIDEr delta is $+0.0111$, 95% CI $[+0.0058, +0.0168]$, $p < 0.001$ – highly significant, $\sim$ 70% of the Qwen3-VL-8B verifier-only lift ($+0.0158$, CI $[+0.011, +0.022]$). The AIC-vs-BLIP monotone improvement is preserved across this swap, and the magnitude of the verifier contribution scales smoothly with verifier capability.

Table 10: **Paired bootstrap on per-image CIDEr.** Observed deltas, 95% bootstrap CIs, and approximate-randomisation p-values. PTB-tokenized corpus CIDEr at every resample. COCO uses B=500 resamples (image count n=5000 is large enough that B=500 already produces sub-0.005 CI width); Flickr30k uses B=2000 (n=1000).

Bench.	Comparison	obs. Δ	95% CI	p
COCO	AIC vs IFCap	+.0411	[+.028, +.054]	<.001
COCO	AIC vs BLIP-prag	+.0158	[+.011, +.022]	<.001
COCO	BLIP vs IFCap	+.0253	[+.010, +.037]	<.001
Flickr30k	AIC vs IFCap	+.0566	[+.033, +.080]	<.001
Flickr30k	AIC vs BLIP-prag	+.0107	[+.0004, +.0214]	.043
Flickr30k	BLIP vs IFCap	+.0459	[+.021, +.069]	<.001

Robustness confirmed: the AIC mechanism does not depend on a specific verifier checkpoint.

Pair-protocol robustness. Beyond the deterministic top-K pairing of Sec. 3.3, we re-ran the full COCO Karpathy reranker (n=5000, BLIP-ITM-large + Qwen3-VL-8B, λ=0.10, K=4) under four alternative pair protocols (Tab. 11):

The interpretation: (i) the abswap variant is

Table 11: **Pair-protocol robustness** (COCO Karpathy test, n=5000). Observed Δ CIDEr against the IFCap-greedy and BLIP-pragmatic baselines. All variants preserve AIC>BLIP>IFCap. Round-robin extracts the most verifier signal at $1.5\times$ verifier cost; the deployed protocol is the cost-quality operating point.

Protocol	$\lvert p \rvert$	Δ **AIC–BLIP**	Δ **AIC–IFCap**
det. (deployed)	4	+.0158	+.0411
A/B-order swap	4	+.0126	+.0380
round-robin	6	**+.0201**	**+.0454**
random (seed 1)	4	+.0122	+.0375
random (seed 2)	4	+.0089	+.0342

slightly below the deployed baseline, indicating a small verifier A/B-position bias that the deployed protocol's per-image-seeded A/B flip mitigates; (ii) round-robin's gain over the deployed protocol shows that the sparse top-K-vs-argmax tournament leaves some verifier signal on the table, and round-robin is a strict improvement when compute budget allows; (iii) random-pair sampling across two seeds shows variance of ~ 0.003 CIDEr, modest relative to the $+0.016$ deployed-baseline verifier gain. The AIC mechanism is robust to pair-protocol choice.