MANY CAPTIONERS OR ONE? THE PRICE AND PAYOUT OF MIXING FROZEN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recaptioning pipelines often mix several vision–language models (VLMs), on the premise that different models see different things. We test this premise against its obvious alternative, resampling one good captioner, at a matched number of captions. To this end, we build a controlled corpus in which 30 frozen VLMs caption the same 1.28M ImageNet training and 50k validation images under one prompt and one decoding specification, with resamples at identical settings. We find that mixing buys protection, not quality. In six of seven uses, from training a text encoder to auditing labels, the pre-registered comparisons find that the mix never reliably beats resampling the captioner that is best for that use, and it is worse in three, including text-encoder training, where much of the gap goes with caption length. Against a typical captioner, however, the mix gains in the same six uses: one captioner drawn at random per image trains a better text encoder than 29 of the 30 models used alone. This protection matters because the best captioner is hard to know in advance. It changes with the use and the prompt, its zero-shot accuracy is an unreliable guide to its value as a training source, and in the seventh use, a concept bottleneck, the captioner we declared best in advance is the weakest source we tested, so the mix wins. Mixing is therefore insurance against a poor choice rather than an improvement on a good one, at about twice the cost of generation. Code: https://anonymous.4open.science/r/ICRL_2027_34944_code.

1 INTRODUCTION

Recaptioning pipelines regenerate a dataset's captions with vision–language models (VLMs) and train on the result (Fan et al., 2023; Li et al., 2024c). Several of these pipelines draw on more than one captioner, on the premise that different models notice different things (Zheng et al., 2024a). This premise has an obvious alternative, namely to sample one good captioner several times; yet, to the best of our knowledge, the two have not been compared for captions at a matched number of captions. Such a comparison requires a corpus in which the images, the prompt and the decoding are held fixed, only the captioner varies, and the same captioner is also resampled. Towards this end, we built one: 30 frozen VLMs caption the same 1.28M ImageNet (Deng et al., 2009) training images and 50k validation images under one prompt and one decoding specification, with resamples at identical settings.

We frame mixing as a trade-off with two sides, both measured as the mix minus a single captioner. The *premium* is the mix minus the captioner that is best for the use, resampled to the same number of captions; a negative premium is what mixing costs. The *payout* is the mix minus a typical captioner; a positive payout is what mixing protects against. Consider, for example, training a text encoder. A random captioner per image trains 12.2 points better than the median captioner alone, which is the payout, whereas six different captioners per image train 3.0 points worse than the best captioner's release caption plus up to five of its samples, which is the premium (pre-registered). The same shape appears in six uses of the captions: training a text encoder, zero-shot classification, reading a test image's captions, and three label audits (Figure 1, Table 5). The premium is significant in training, in primary-label identification (-6.1; -3.8 with a learned aggregator) and in test-time reading with three captions (-2.8), and it is not significant after correction elsewhere among these six. In the pre-registered comparisons, the mix never significantly beats the captioner that is best for the use, resampled, after multiplicity correction; a secondary prototype-averaging construction favours it by

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

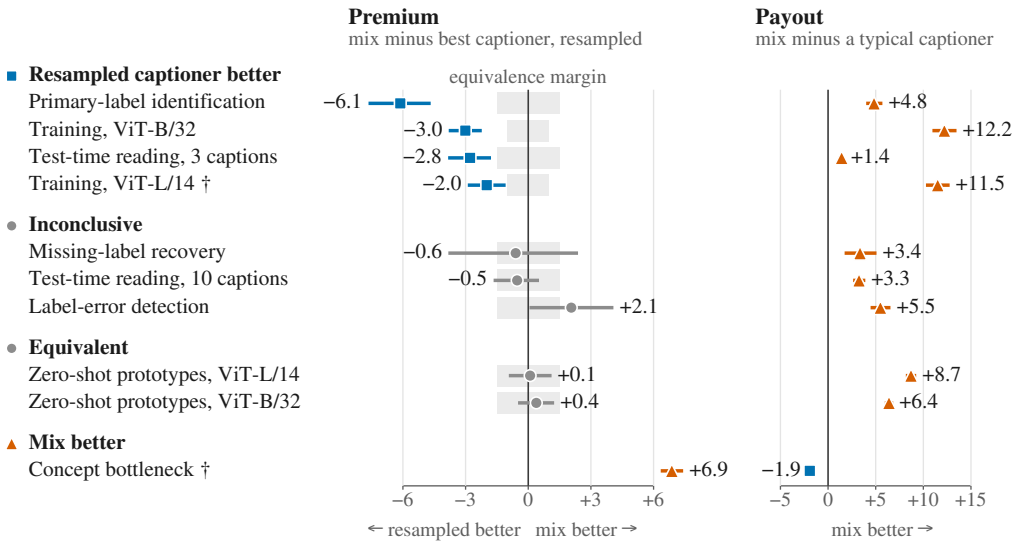


Figure 1: The price and the payout of mixing captioners, per use, at a matched caption count. *Premium*: the mix minus the captioner that is best for the use, resampled (negative = what mixing costs). *Payout*: the mix minus a typical captioner (positive = what mixing protects against). Rows are grouped by the verdict after multiplicity correction; the points are estimates with 95% class-bootstrap CIs, not multiplicity-adjusted. AUROC and precision $\times 100$; grey bands, the equivalence margin fixed before computing. †: not pre-registered. In the concept bottleneck, the mix is compared with the captioner declared best in advance, which turned out to be the weakest source we tested for that use, so there the mix wins. Budgets and references per row are in Table 5.

1.0 point. The mix wins clearly only in a seventh use, a concept bottleneck, where the resampled captioner is the wrong one for the use. Moreover, the mix is also the more expensive option to generate (Section 8).

Beyond this trade-off, three findings qualify the usual ensemble argument. First, captioners differ far more as training sources than their prototype accuracy suggests. Trained alone, the thirty span 18.4–41.5% zero-shot accuracy, whereas a captioner's prototype accuracy is only loosely associated with that value (Spearman 0.23 on ViT-B/32, not significant, and 0.60 on ViT-L/14); in fact, a mix of the six next-best captioners by this proxy trains worse than a random six. Second, in training, extra captions per image help about as much whether they come from resampling or from other models (an interaction of only $+0.4$ points, not significant). This is consistent with the gap between the best captioner resampled and a mix coming from the source rather than from diversity, although much of the source's advantage goes with caption length: with every caption capped at 64 or 48 tokens, the lead shrinks to 1.4 or 0.8 points. Third, which captioner is best depends on the use. The best contrastive-training source is the weakest of four sources for a concept bottleneck, and the best zero-shot captioner ranks 22nd of 30 for label-error detection. Resampling the second-best training source instead does not beat a mix, so in training, resampling pays only if a pilot finds the best source. Taken together, in our ImageNet setting a mix reduces the downside of an unpiloted choice, whereas resampling the right captioner is better once it has been found.

Overall, our contributions can be summarized as follows:

- A controlled corpus of 30 frozen VLMs captioning all of ImageNet under one prompt and one decoding specification, with resamples at identical settings and a quality gate that needs no reference captions.
- A matched-caption-count comparison of many captioners against one captioner, resampled, across seven uses, reported as a premium and a payout, with pre-registered primary comparisons: mixing insures against a weak choice but does not reliably improve on the best captioner.

- A practical rule for when to mix and when to resample: the best captioner is specific to the use and only loosely predicted by its prototype accuracy, and a small labeled pilot finds it for the evaluation uses but not for training.

2 Related Work

Recaptioning with one or several captioners: LaCLIP rewrites captions with a large language model (Fan et al., 2023), VeCLIP enriches them with visual detail (Lai et al., 2023), and Recap-DataComp recaptions billions of web images with a single captioner (Li et al., 2024c). Captioning-benchmark scores do not predict a captioner's value for CLIP training (Nguyen et al., 2023), and downstream models prefer different caption formats (Lai et al., 2025). Other pipelines combine several captioners: DreamLIP draws long captions from three multimodal models and merges them for pre-training (Zheng et al., 2024a), while IC3 summarises a committee of captions into one (Chan et al., 2023). Liu et al. (2023b) combine four captioners, but against a single-model baseline with one caption per image. These works treat several captioners as a construction choice and do not separate the benefit of diversity from that of more stochastic captions; we compare several captioners with one captioner, resampled, at a matched number of captions.

Ensembles and repeated sampling: Self-consistency marginalises several sampled outputs of one model (Wang et al., 2022), and deep ensembles combine independently trained networks (Lakshminarayanan et al., 2017). From this perspective, a mix of captioners should beat a typical member but not a much stronger one. For language models, Self-MoA finds that aggregating samples of the best LLM usually beats mixing different ones (Wang et al., 2025a; Li et al., 2025), whereas voting over LLMs of comparable strength favours the mix (Chen et al., 2026a); we ask the question for captions used as data, across several uses. Caption utility has also been studied directly, by CAPEval, which varies the captioner and relates caption properties to downstream performance (Liu et al., 2026), and by CaptionQA (Yang et al., 2026). CAPEval does not resample a captioner, so it cannot separate the choice of source from the diversity of its captions, which is the comparison we make.

Descriptions derived from captions or language: Menon & Vondrick (2022) and Pratt et al. (2022) improve CLIP (Radford et al., 2021) classifiers by prompting them with generated class descriptions, whereas Roth et al. (2023) show that random descriptors are a strong baseline. Similarly, concept bottlenecks (Koh et al., 2020) have been built from language-derived concepts (Yang et al., 2022; Oikarinen et al., 2023) and from captions (Cagliyan et al., 2026). When such descriptions are derived from captions of a class's own images, they can restate the class name in forms that a token filter misses; in our corpus, this inflated zero-shot accuracy by about 7 points (Appendix E). Our label-audit uses of the captions relate to ReaL (Beyer et al., 2020) and to work on label errors (Northcutt et al., 2021b); Appendix H discusses these and further related lines.

3 Corpus and Matched-Caption-Count Design

Corpus: Thirty vision–language models of diverse architectures (projector-based, early-fusion, encoder-free, perceiver-resampler, cross-attention and visual-expert designs, from about 4B to 72B parameters; Table 3) caption the *same* 1,281,167 ImageNet training images and the same 50,000 validation images. All receive one prompt and decode under one specification: temperature 0.6, top-p 1.0, no top-k truncation, no repetition penalty and at most 300 tokens. Every caption passes a nine-check quality gate, which screens for incomplete or duplicated output, empty captions, instruction leakage, refusals, degenerate text, captions of the wrong image, and captions that describe the class rather than the picture (Appendix B). A captioner enters the study only when both of its splits reach 99.99% coverage; failing captions are repaired or regenerated, and a small residual is dropped. The gate needs no reference captions, but it is not label-free: two checks use the ImageNet class label. Both act on whole captioners, and neither removed an individual caption from any label-audit input.

Because images, prompt and decoding are held constant, a difference between two caption sets is attributable to the captioner. Throughout, a *captioner* therefore denotes the whole system: checkpoint, serving configuration (Appendix I), common prompt and common gate. We call the zero-shot top-1 of class prototypes built from a captioner's training captions its *prototype accuracy* (Section 5). Notably, the ordering of the captioners is stable across temperature, top-p and top-k (mean rank

Table 1: Matched-caption-count design. "Resample": how the one-captioner arm gets several captions per image. "Best": how the resampled captioner is chosen. "Reg.": what was committed before computing (yes: comparison and margin; margin: the equivalence margin only; metric only: the metric only).

Use	Images	Resample	Best (chosen on)	Reg.
Text-encoder training (Sec. 4)	300k train	release + up to 5 samples	prototype acc., val calib.	yes
Concept bottleneck (Sec. 4)	200k train	release + up to 5 samples	same captioner	metric only
Zero-shot prototypes (Sec. 5)	300k train	release + up to 5 samples	val calib.	yes
Label audits (Sec. 5)	2k val	30 true samples	held-out classes	margin
Test-time reading (Sec. 6)	2k val	30 true samples	val calib. half	yes

correlation 0.995 over the 14 non-release cells of a 42-configuration re-captioning of 2,000 images), but not across generation budgets (0.53 at a 100-token cap) or prompts (Appendix C).

Resamples: Two sets of resamples at identical settings make "one captioner, resampled" measurable. First, ten captioners caption $2{,}000$ validation images (two per class) 30 times each. Second, the captioner declared best in advance, Gemma 4 12B (the strongest leak-free prototype captioner on a validation calibration split), captions a class-stratified subset of 300k training images five more times. Only the sampling seed changes. Resamples that fail the gate were not regenerated, so each of the five training passes covers about 94% of the 300k images, and an image has 5.7 of its six captions on average.

Matched caption count: Each premium comparison gives the mix and the resampled captioner the same nominal number of captions per image (Table 1). In training, the resampled arm has only 5.7 of its six captions on average, which works against it. The premium is the mix minus the captioner that is best for the use, resampled, and the payout is the mix minus a typical captioner; Table 5 gives each row's budget and reference. Label audits are compared at the same number of predicted labels, not at the same agreement threshold: one captioner's samples are correlated, so at a given threshold it predicts more labels, and a threshold-matched comparison favours the mix. Where we test equivalence (two one-sided tests (Schuirmann, 1987)), the margin was fixed before computing: ± 1.5 points, about a tenth of the 13.8-point spread across captioners in zero-shot accuracy, or $\max(1.0, 2 \times \text{seed SD})$ for training. Superiority is tested in both directions, p-values are Holm-corrected (Holm, 1979) within each pre-registered family, and uncertainty comes from a class-level bootstrap. The registered comparisons and their amendments were committed to version control before the runs they govern, with two exceptions (Appendix I). Every other analysis is marked † in figures and tables and is reported with its unadjusted 95% interval.

4 Captions as Training Data

Protocol: Following locked-image tuning (Zhai et al., 2021), we keep a CLIP image encoder frozen, train a text encoder from scratch on captions of ImageNet training images, and classify the ImageNet validation images zero-shot with class-name prompts through the trained encoder. All arms share the images (the same 300k, 300 per class; the complete-case checks below use a 225k subset), the architecture and the schedule, and all 300k-image arms share the number of optimiser steps; only the caption source changes. With one fixed caption per image, a run memorises its captions and declines after a few epochs; we therefore choose each run's epoch on 10k held-out *training* images, a rule committed before the runs it governs. The pre-registered comparison uses six captions per image: (E) the best captioner's release caption plus its five new samples that passed the gate (5.7 captions per image on average), against (D_6) six different captioners drawn independently for each image from all thirty (a per-image policy, not a fixed ensemble; the three seeds use independent draws). Two one-caption arms complete the design: (B) the best captioner's release caption alone, and (D_1) the caption of one captioner drawn at random per image (details in Appendix D).

Main comparison: E reaches 46.4% against 43.4% for D_6 on ViT-B/32, a lead of 3.0 points (95% CI $[2.2, 3.8]$; all three seeds positive; Table 6). The lead holds under every check we ran

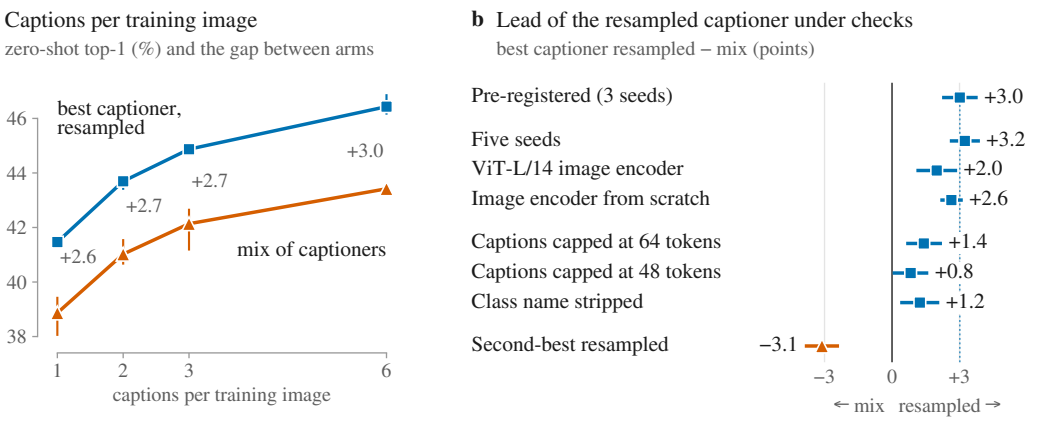


Figure 2: Training a text encoder. (a) Against a locked CLIP ViT-B/32 image encoder on 300k ImageNet training images: the best captioner resampled (E and its one- to three-caption versions) and a mix of captioners drawn per image (D), by captions per image; three seeds, bars show the seed range, numbers give the gap. Both gain about equally from more captions, consistent with the gap coming from the source rather than from a lack of diversity. (b) The lead of the best captioner resampled over the mix ($E - D_6$; the opposite sign of the premium in Figure 1) under a selection of checks (blue: the resampled captioner leads; orange: the mix leads), with 95% class-bootstrap CIs (five seeds: seed-level interval); image encoder and image set as labelled, ViT-B/32 and 300k images otherwise. The dotted line marks the pre-registered estimate; the other rows are follow-up checks. The last row resamples the second-best training source instead (50k images, 720 steps). Every check is in Tables 6 and 7.

(Figure 2b; Tables 6 and 7; most specified before running, Appendix I): five paired seeds ($+3.2$), the first 200k images ($+3.1$), a ViT-L/14 image encoder ($+2.0$), the validation halves not used to choose the captioner ($+3.0$), exactly six captions per image ($+2.8$), the best captioner's raw, ungated captions ($+2.7$), its repetition-flagged images dropped from both arms ($+2.9$), and a ResNet-50 image encoder trained from scratch together with the text encoder ($+2.6$, 9.6% vs. 7.0%; $+5.0$ when trained three times longer, one seed). However, the mix's captions name the image's class more often (60.3% vs. 57.5%) and, within the encoder's 128-token context, are longer (108 vs. 96 tokens). Length carries much of the lead. With every caption capped at the same 64 or 48 tokens, both arms improve, the mix much more ($+2.2$ and $+3.3$ points, against $+0.6$ and $+1.1$), and E leads by 1.4 ($[0.6, 2.2]$) and 0.8 ($[0.0, 1.6]$). Both intervals lie above zero on the class bootstrap, the registered reading, whereas the seed-level intervals include zero. Class names carry part of the lead, but not all of it: with every name form of each image's own class stripped from its captions, both arms collapse, yet E still leads, 17.7% vs. 16.5% ($+1.2$, $[0.4, 2.1]$). Furthermore, the best captioner's captions of an image are as varied in wording as those of six captioners (mean pairwise word overlap 0.235 vs. 0.233), although one captioner's samples agree more on the classes they name. It is also the best of the thirty as a single training source, on held-out training images (43.0%, three seeds) as on validation.

Source versus diversity: E, B, D_6 and D_1 form a 2×2 design (source: best or random; captions per image: one or six). Going from one to six captions per image helps about equally whether the extra captions are samples of the same captioner ($E - B$, $+5.0$ points) or come from other captioners ($D_6 - D_1$, $+4.6$). The interaction, $+0.4$ ($[-0.3, +1.2]$; D_1 was not pre-registered), is not significant. This is consistent with most of the advantage of E over D_6 being the advantage of the best source over a random one, which at one caption per image is $+2.6$. The scaling runs agree: with two and three captions per image the gap is $+2.7$ and $+2.7$, so both arms gain about equally from more captions (Table 7). At the resampled best captioner's generation cost, a random mix affords about three captions per image (Appendix D); with three, it reaches 42.1%, 4.3 points below it.

Single sources and the prototype proxy: Trained alone on the same images (one caption per image), the thirty captioners span 18.4–41.5%, with a median of 26.7% (Figure 3). This is where

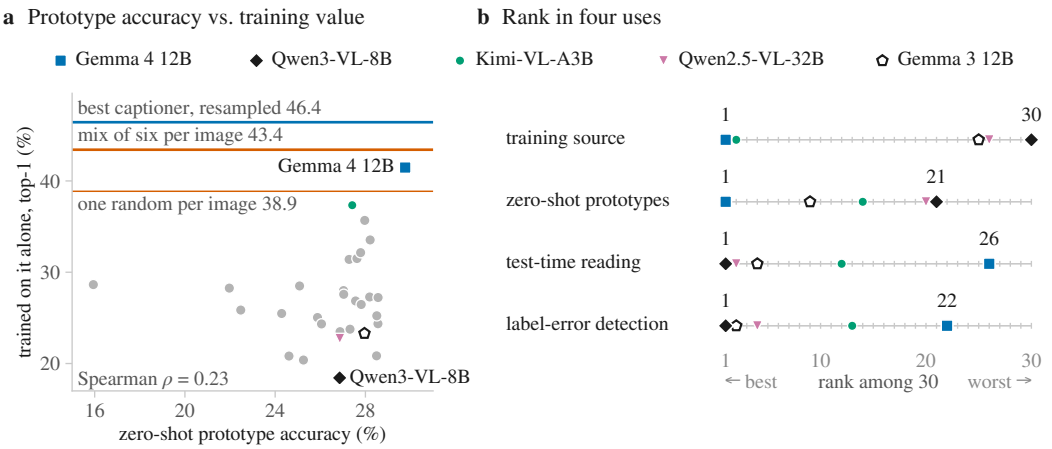


Figure 3: Single captioners as training sources, and their ranks across uses. (a) Each of the 30 captioners trained alone (ViT-B/32, one caption per image; single seed except the best captioner, three), against its prototype accuracy (Spearman 0.23 here, 0.60 on ViT-L/14, Section 4); lines: the best captioner's release caption plus up to five samples, six captioners per image, and one random captioner per image. (b) Rank among the 30 captioners for four uses (one caption each) of the best training source (Gemma 4 12B), the best reader (Qwen3-VL-8B) and the runner-up of three uses (Kimi-VL-A3B, training; Qwen2.5-VL-32B, reading; Gemma 3 12B, label errors). The best training source is 26th for reading; the best reader is the worst training source.

the insurance pays off: one captioner drawn at random per image (D_1) reaches 38.9%, 12.2 points above the median captioner and above 29 of the 30, yet 2.6 points below the best one. On ViT-L/14 the pattern repeats: D_1 reaches 51.8%, 11.5 points above the median captioner ($[10.3, 12.8]$), again above 29 of the 30 and 2.2 points below the best, and the captioners rank similarly as training sources on the two image encoders (Spearman 0.76). A captioner's prototype accuracy, in contrast, is only loosely associated with its value as a training source (Spearman 0.23, $[-0.15, 0.56]$ over captioners; 0.60, $[0.28, 0.81]$, on ViT-L/14), and on ViT-B/32 seed noise cannot explain this, since the single-seed reliability is 0.99; nevertheless, the best captioner is first on both. Mixing the six next-best captioners by that proxy trains *worse* than a random six (35.6% vs. 43.4%; ViT-L/14: 48.7% vs. 56.1%), since as single sources they rank between 4th and 27th. Replacing one of these six with the best captioner recovers only the random-six level (43.2%), still 3.3 points below its own samples. Ranking by training value itself, which requires a training pilot, still falls short: the six best single training sources, ranked on held-out training images, reach 44.3%, 0.9 points above a random six but 2.1 below the resampled best captioner, and without the best captioner the same ranking is 1.5 points below a random six. Moreover, drawing six captioners once and keeping them is risky: three such draws (one seed each) reach 40.2, 36.7 and 44.9% (only the last includes the best captioner), against 43.4% when the six are drawn per image. Finally, resampling did not pay for the second-best training source (Kimi-VL-A3B, chosen post hoc; 50k images). In the 120-step block, specified before running, where every run was still improving, it ties six captioners ($+0.1$, $[-0.5, +0.7]$), and in the declared 720-step block it trains 3.1 points *worse* ($[2.4, 3.9]$), whereas the best captioner resampled on the same images leads by 5.5 and 2.7. Kimi-VL-A3B alone (37.3%) is below one random captioner per image (38.9%).

Concept bottleneck: When a concept bottleneck (Koh et al., 2020) is supervised on frozen CLIP features with name-free WordNet (Miller, 1995) concepts from the captions (Appendix G), the same best captioner is the weakest of the four single sources we tested (39.0% against 48.7–53.2%); after class names are removed, its captions keep the fewest concepts. Resampling it does not help (-0.1 points), and six different captioners beat its release caption plus up to five samples by 6.9 points. Thus, where the resampled captioner is the wrong one for the use, the mix is the better choice, which is the insurance case.

5 Zero-Shot Prototypes and Label Audits

Zero-shot prototypes: For each captioner, we build class prototypes from the nouns its captions use for a class's training images, with every form of the class name removed (a word-level filter misses compounds such as "bald eagle" and inflates accuracy by 7.2 points; Appendix E); we then encode them with CLIP's text encoder and classify the 50k validation images. The captioners span 13.8 points of top-1, and which captioner performs well depends strongly on the class: in the 30×1000 captioner-by-class accuracy matrix, a global captioner ranking explains 1.0% of the variance and the captioner-by-class interaction 28.9% (34.6% on ViT-L/14), and the per-class best captioner is the same on the two backbones for only 16% of classes (Appendix E). An unweighted vote over the thirty, which uses no labels to weight or choose captioners, improves on the best single captioner by 4.1 points on ViT-B/32 and 4.8 on ViT-L/14. The decisive comparison, registered before running, uses true resamples and the same construction for both arms. On the 300k training images with resamples, the best captioner's release caption plus its five samples give six prototype sets, against six sets from six different captioners (100 random draws); all sets are built by the same procedure and combined by the same vote. The two arms are equivalent within ± 1.5 points on both backbones: the mix is ahead by 0.4 ($[-0.5, +1.2]$) on ViT-B/32 and by 0.1 ($[-0.9, +1.1]$) on ViT-L/14, and equivalence holds after Holm correction within the registered zero-shot family (TOST $p = 0.008$ on both backbones). The payout, in contrast, is large: the same diverse vote beats the mean single captioner by 6.4 and 8.7 points, whereas resampling the best captioner adds only 2.1 and 3.3 points over its single prototype set. When the six prototype sets are averaged into one instead of voted, a secondary construction, the mix is ahead by 1.0 point on ViT-B/32 and inconclusive on ViT-L/14. Comparisons against caption subsets of the best captioner are in Appendix E.

Label audits: Counting how many captioners name each concept in an image yields audit signals for ImageNet's labels. Concepts that many captioners mention but the label set lacks recover missing labels (precision 91.7% at unanimity against human ReaL labels, versus a co-occurrence null of 23.4%, at a recall of 1.3%). Conversely, images whose label few captioners mention flag label errors: against the human re-annotation of ReImageNet (Volkov et al., 2026), this signal reaches an AUROC of 0.738, ahead of a CLIP margin baseline but below a class-error-rate prior (0.780); added to per-class intercepts, it raises the AUROC from 0.825 to 0.870 (Appendix F). With ten captioners resampled 30 times each on 2,000 validation images, we compare different captioners against one captioner's samples at the same number of captions and predictions: ten captioners with three samples each (for label errors, all thirty captioners' release captions) against one captioner's thirty samples. Against the captioner best for each audit, chosen on held-out classes, the mix identifies an image's primary label worse (-6.1 points, $[-7.6, -4.7]$, significant after correction; -3.8 when both arms use a Dawid–Skene aggregator fitted without labels on one half of the images and applied to the other; post hoc, Appendix F). It is inconclusive on missing labels (-0.6) and on label errors ($+2.1$ AUROC points, $[+0.1, +4.1]$ before correction, not significant after it). Against the average captioner resampled, however, the mix is better on all three audits (primary label $+4.8$, missing labels $+3.4$, label errors $+5.5$ points).

6 Reading the Captions at Test Time

Protocol: Unlike prototypes, which use *training* captions only, this use reads a test image's own captions. The vote proposes the ten most likely classes of each validation image, and a small combiner, fitted on a calibration half, reranks them by the word overlap between the image's captions and a per-class noun profile built from the training captions, with every caption word that names any candidate masked first. This supervised caption classifier raises top-1 from 33.9% to 49.0% on ViT-B/32 and from 41.1% to 52.5% on ViT-L/14, whereas giving each image another image's captions returns exactly the vote (Table 2).

Matched caption count: At a matched number of captions, diversity does not help this reading; since all arms rerank the same candidate list, the comparison isolates the reading step. On the 2,000 images resampled 30 times, three captions from three different captioners score 2.8 points below three samples of the captioner chosen on the calibration half ($[-3.8, -1.8]$, significant after correction), and ten from ten score 0.5 below ten of its samples ($[-1.7, +0.5]$; equivalence is

Table 2: Reading the image's own captions at test time. **Top:** the vote's top-10 candidates are reranked by word overlap between the image's captions (every name form of every candidate masked) and a per-class noun profile; top-1 (%), held-out halves of the ImageNet validation set. **Bottom:** the same rerank with test captions from different sources; matched-count rows report diverse minus resampled (points, 95% class-bootstrap CI). The best captioner here (Qwen3-VL-8B) was chosen on the calibration half, on both backbones. †Not pre-registered (ViT-L/14 replication, specified before running; the pre-registered comparison is on ViT-B/32). Differences are computed before rounding.

Method	ViT-B/32	ViT-L/14
Unweighted vote, 30 captioners (candidates)	33.9	41.1
+ training-caption profile	49.0	52.5
+ training-caption profile, strictest name mask	47.0	51.4
control: another image's captions	33.9	41.1†
Matched caption count, 2,000 validation images resampled 30 times		
3 different captioners vs. best captioner ×3	44.2 vs. 47.0	48.3 vs. 51.6
difference, 95% CI	−2.8 [-3.8, -1.8]	−3.3 [-4.3, -2.2]†
10 different captioners vs. best captioner ×10	47.1 vs. 47.7	50.9 vs. 52.4
difference, 95% CI	−0.5 [-1.7, +0.5]	−1.5 [-2.6, -0.4]†

not established after correction). On ViT-L/14 (not pre-registered, specified before running), the resampled captioner leads at both counts, by 3.3 and 1.5 points, with both intervals excluding zero.

7 The Price and the Payout

The payout of mixing (Figure 1, Table 5) is positive in every use where we measured it but one: 1.4–12.2 points over a typical captioner, largest in training. The exception is the concept bottleneck, where a random captioner per image falls 1.9 points below the mean of three randomly drawn single captioners. The premium, in turn, is not significant for zero-shot prototypes (where equivalence holds), missing labels, label errors (after correction) and test-time reading with ten captions on ViT-B/32. It is a significant cost in training, primary-label identification and test-time reading with three captions, and it turns into a gain only in the concept bottleneck, where the resampled captioner is the wrong one for the use.

The cost of finding the best captioner: Whether the premium is worth paying depends on this cost. For zero-shot prototypes, few labels suffice: with five labeled images per class, selection on a calibration split picks the best zero-shot captioner in 95% of draws when the pilot covers the test classes, and in 71% (89% with 25 images) when it does not. Test-time reading and label-error detection need a similar pilot: on classes disjoint from the pilot, 25 labeled images per class pick the best test-time reader in 88% of draws and the best label-error detector in 89%, with mean regret of 0.07 and 0.14 points (Appendix D). However, the choice does not transfer across uses. The best zero-shot captioner ranks 22nd of 30 for label-error detection and 16th for missing-label recovery and primary-label identification (rank correlation 0.15 between zero-shot accuracy and label-error AUROC), and prototype accuracy predicts training value only at 0.23 (0.60 on ViT-L/14). Conversely, the captioner that is best for both test-time reading and label-error detection, Qwen3-VL-8B, is the worst of the thirty as a training source (18.4%). Measured as regret against the better of mixing and resampling in each use, always mixing risks at most 6.1 points (primary label; 3.8 with the learned aggregator), always resampling each use's reference captioner 6.9 (concept bottleneck), and piloting with five labeled images per class at most 2.5 wherever selection was measured.

In practice: These numbers suggest a simple rule. Where the use can be piloted with labels, as for zero-shot prototypes, test-time reading and label-error detection, select the captioner on 5–25 labeled images per class and resample it; in our measurements this forgoes the premium at a regret of at most 2.5 points. Training a text encoder is different, because only training runs reveal the best source, and mixing the next-best captioners by the prototype proxy backfires. With such a pilot, resampling the best source still pays (3.0 points on ViT-B/32). Without one, mix: one captioner

drawn at random per image comes within 2.6 points of the best single source on ViT-B/32 (2.2 on ViT-L/14) without any selection, at about twice the generation cost of resampling one captioner.

8 LIMITATIONS

Scope of the training evidence: The main training experiments keep the CLIP image encoder frozen. The one check in which the captions also train an image encoder uses a ResNet-50 trained from scratch at 128 pixels on 300k images, which reaches at most 18.7% zero-shot accuracy; image-encoder training at the scale of recaptioning pipelines therefore remains untested. The experiments use 300k ImageNet training images, class-name prompts at evaluation and one captioner resampled at that scale; the single-source runs use one seed each (the best captioner and the random-per-image arm use three), and the concept-bottleneck comparison covers four single sources. The epoch rule for training was adopted after single-caption pilots showed memorisation, but before any run it governs. Each of the best captioner's five training resample passes covers about 94% of the 300k images, because samples that failed the quality gate were not regenerated; restricting training to the images with all six captions leaves the primary difference at $+2.8$. The zero-shot result also depends mildly on how prototype sets are combined (voting or averaging; Section 5).

Equivalence only for zero-shot prototypes: Only there are the mix and the best captioner resampled equivalent, on both backbones. The label-audit comparisons are mostly inconclusive in both directions, and the design could not have settled them: with bootstrap standard errors of 1.0 and 1.6 points, they could not establish equivalence at ± 1.5 even if the true difference were zero (about $4\times$ and $10\times$ more classes would be needed for 80% power). Our claim is therefore "no reliable advantage for the mix over the best captioner resampled", not "interchangeable".

Matched captions, not matched cost: Our mixes pool the captions of different captioners; pipelines that merge several captions into one (Zheng et al., 2024a; Chan et al., 2023) were not tested. We match captions per image, not compute. Costing generation with the throughput measured in our own production runs (Appendix D), six captioners drawn at random cost on average 1.8–$2.1\times$ the GPU time of the best captioner resampled six times, so in training the resampled captioner is both cheaper and better. A mix also means serving several models, from about 4B to 72B parameters, each with its own serving recipe. Nor is the insurance free: a caption from a random captioner costs on average $2.7\times$ one from the captioner of median training value. These are production rates, not tuned optima (Appendix D).

One dataset, one prompt: All experiments use ImageNet-1k, with WordNet's label set and one captioning prompt. A small check with two other prompts (five captioners, 2,000 images; Appendix C) largely reorders the five captioners' caption lengths ($\rho = 0.3$ and -0.3) and partly reorders their test-time reading ($\rho = 0.8$ and 0.4); the best zero-shot and training captioner at our prompt reads worst of the five under both alternatives, although training and zero-shot were not re-run. Every "best captioner" in this paper is thus best at our prompt, which, if anything, strengthens the case for a mix when the use cannot be piloted. Finally, caption-derived prototypes measure caption content and fall far below a 25-shot linear probe (at most 35.2% vs. 68.1% on ViT-B/32).

9 CONCLUSION

This work compared many captioners with one captioner, resampled, at a matched caption count, on a corpus in which thirty frozen VLMs caption all of ImageNet under identical conditions. Across seven uses, mixing usually gains more over a typical captioner than it costs against the captioner best for the use, resampled; the exceptions are primary-label identification with a plurality vote and three-caption test-time reading. Moreover, the best captioner changes with the use and the prompt: for zero-shot prototypes, test-time reading and label-error detection, a pilot of 5–25 labeled images per class finds it, or one a fraction of a point worse on average, whereas for training only training runs do. Hence, in the ImageNet uses we measured, with frozen or small-scale image encoders, selecting and resampling a captioner for the use generally avoids the premium of mixing, whereas mixing reduces the risk of a weak choice when no such selection is possible.

AI USE STATEMENT

An AI coding assistant helped implement and run the experiments (model serving, caption quality control and analysis scripts) under the authors' direction, and language models were used to refine the style and check the grammar of the text. All reported numbers are produced by scripts from the stored results, and every reference was checked against its source. The authors take full responsibility for the content of the paper.

ETHICS STATEMENT

The study involves no human subjects and collects no personal data. It uses the public ImageNet-1k images, which we do not redistribute, and the public ReaL and ReImageNet relabelings. The captions are produced by publicly released models and may reproduce their biases, including about people in ImageNet images; the captions and resamples will be made available upon publication, for research use under the captioning models' licences.

REPRODUCIBILITY STATEMENT

The corpus, prompt, decoding specification and roster are in Appendix A, the quality gate in Appendix B, training and evaluation settings and cost measurements in Appendix D, and serving and the pre-registrations in Appendix I. Every table is generated from stored results by script, and the pre-registrations were committed before the runs they govern (two exceptions, Appendix I). The code is at https://anonymous.4open.science/r/ICRL_2027_34944_code.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2026-09-26.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Fuyu-8B: A multimodal architecture for AI agents. Adept AI blog, https://www.adept.ai/blog/fuyu-8b, 2023.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Seref Baris Cagliyan, Umut Ozdemir, Merve Tapli, and Emre Akbas. Caption bottleneck models. In *European Conference on Computer Vision*, pp. 445–461. Springer, 2026.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

David M Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. Ic3: Image captioning by committee consensus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8975–9003, 2023.

Jianhao Chen, Zishuo Xun, Bocheng Zhou, Han Qi, Hangfan Zhang, Qiaosheng Zhang, Yang Chen, Wei Hu, Yuzhong Qu, Wanli Ouyang, and Shuyue Hu. Do we truly need so many samples? multi-llm repeated sampling efficiently scales test-time compute. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026a.

Junyu Chen, Md Yousuf Harun, and Christopher Kanan. Unlocking imagenet's multi-object nature: Automated large-scale multilabel annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2284–2293, 2026b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, et al. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, et al. Molmo2: open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, et al. SenseNova-U1: unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

ERNIE Team, Baidu. ERNIE 4.5 technical report. https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf, 2025.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

Gautam Rajendrakumar Gare, Jia Shi, Zhiqiu Lin, Deepak Pathak, John Galeotti, and Deva Ramanan. Attributes should come from images, not class names: Distribution-conditioned attribute selection for vision-language models. *arXiv preprint arXiv:2607.18695*, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, et al. CogVLM2: visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing structural regularities of labeled data in overparameterized models. *arXiv preprint arXiv:2002.03206*, 2020.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *arXiv preprint arXiv:2007.04612*, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. Veclip: Improving clip training via visual-enriched captions, 2023.

Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Juan Lao Tebar, Wenze Hu, Zhe Gan, Peter Grasch, Meng Cao, and Yinfei Yang. Revisit large-scale image-caption data in pre-training multimodal foundation models. In *International Conference on Learning Representations (ICLR)*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*, 2025.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024.

Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. Mllms-augmented visual-language representation learning. *arXiv preprint arXiv:2311.18765*, 2023b.

Zhipeng Liu, Haochen Wang, and Zhaoxiang Zhang. Capeval: A decoupled caption evaluation across understanding and generation. *arXiv preprint arXiv:2608.02589*, 2026.

Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. Trivial or impossible–dichotomous data difficulty masks model differences (on imagenet and beyond). *arXiv preprint arXiv:2110.05922*, 2021.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021a.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021b.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0/, 2024.

Guillermo Ortiz-Jimenez, Mark Collier, Anant Nawalgaria, Alexander Nicholas D'Amour, Jesse Berent, Rodolphe Jenatton, and Efi Kokiopoulou. When does privileged information explain away label noise? In *International Conference on Machine Learning*, pp. 26646–26669. PMLR, 2023.

Zirui Pang, Haosheng Tan, Yuhan Pu, Zhijie Deng, Zhouan Shen, Keyu Hu, and Jiaheng Wei. When vlms meet image classification: Test sets renovation via missing label identification. *arXiv preprint arXiv:2505.16149*, 2025.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB endowment. International conference on very large data bases*, volume 11, pp. 269, 2017.

Karsten Roth, Jae Myung Kim, A Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15700–15711. IEEE, 2023.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 8634–8644, 2020.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Illia Volkov, Nikita Kisel, Tetiana Mishkina, Klara Janouskova, and Jiri Matas. Doomed to re-annotate, forever: The imagenet story. *arXiv preprint arXiv:2608.13783*, 2026.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations (ICLR)*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, et al. InternVL3.5: advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, et al. DeepSeek-VL2: mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. Captionqa: Is your caption as useful as the image itself? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23741–23750, 2026.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv preprint arXiv:2211.11158*, 2022.

Moran Yanuka, Morris Alper, Hadar Averbuch-Elor, and Raja Giryes. Icc: Quantifying image caption concreteness for multimodal dataset curation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11048–11064, 2024.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, et al. MiniCPM-V: a GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, et al. MiniCPM-V 4.5: cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels, 2021. URL https://arxiv.org/abs/2101.05022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *arXiv preprint arXiv:2111.07991*, 2021.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pp. 73–90. Springer, 2024a.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

A Corpus Details

Images: All captioners describe the same $1{,}281{,}167$ training and $50{,}000$ validation images of ImageNet-1k. Images that are not RGB are converted to RGB before captioning.

Captioners: Table 3 lists the 30 captioners. We chose them for architectural breadth rather than leaderboard rank: the roster includes an early-fusion token model (Chameleon-7B), an encoder-free model (Fuyu-8B), a perceiver-resampler design (Idefics2-8B), a cross-attention design (Llama-3.2-11B-Vision), mixture-of-experts backbones (DeepSeek-VL2-Small, Aria, ERNIE-4.5-VL), and scales from about 4B (Phi-3.5-Vision) to 72B (Qwen2.5-VL-72B) parameters. A captioner is admitted only if its captions pass the quality gate of Appendix B on at least 99.99% of the images of both splits; all 30 do.

Prompt: Every captioner receives the same prompt on both splits:

> Describe the image using 3 to 5 sentences. Your whole answer should be about 160 words in total. Do not provide any preamble or summary/ending.

Three captioners need a change to the request. The Phi models require their own image tag in the prompt, and Gemma 4 requires its chat template to begin with the start-of-sequence token; neither changes the instruction. SenseNova-U1 requires an appended instruction to answer in English and its reasoning mode switched off; this is the only change to the instruction itself.

Decoding: All captioners decode with temperature 0.6, top-p 1.0, no top-k truncation, no repetition penalty and at most 300 new tokens. The temperature is set explicitly in every request, so a checkpoint's default generation settings cannot override it. The 300-token limit is a ceiling that accommodates the requested length of about 160 words, not a length target.

B The Caption-Quality Gate

A caption corpus is only as trustworthy as its captions. A fluent caption that leaks the prompt, refuses the task, degenerates into a loop or describes the wrong image contaminates every agreement statistic, yet passes any fluency or grounding score. We therefore admit a captioner only after its captions clear a nine-check gate on both splits. The gate needs no reference captions, and each check was added because a real captioner failed it silently.

Structural checks:

1. **Completeness:** every image of the split has a caption.
2. **Uniqueness:** every image has exactly one caption.
3. **Coverage:** no image is missing; duplicated and missing captions can occur in the same output.
4. **Non-empty:** no empty caption and none shorter than five words.

Content checks:

5. **No instruction leakage:** the caption does not echo the prompt. instruction fragment followed by a real caption ("*Write directly into the text field.* A man holding a fish").

Table 3: The 30 captioners. All receive the same images, prompt and decoding; the captioner and its serving configuration (Appendix I) are what vary. Parameters as stated by each model's source (total / activated for mixture-of-experts models). Checkpoint: the Hugging Face repository served; Llama-3.2-11B-Vision and Pixtral-12B were served from unsloth's re-uploads of the official weights. Every captioner passes the quality gate (Appendix B) on at least 99.99% of the images of both splits.

Captioner	Parameters	Checkpoint
Aria (Li et al., 2024b)	25.3B / 3.9B act.	`rhymes-ai/Aria`
Chameleon-7B (Chameleon Team, 2024)	7B	`facebook/chameleon-7b`
DeepSeek-VL2-Small (Wu et al., 2024)	2.8B act.	`deepseek-ai/deepseek-vl2-small`
ERNIE-4.5-VL (ERNIE Team, Baidu, 2025)	28B / 3B act.	`baidu/ERNIE-4.5-VL-28B-A3B-PT`
Fuyu-8B (Bavishi et al., 2023)	8B	`adept/fuyu-8b`
Gemma 3 12B (Gemma Team et al., 2025)	12B	`google/gemma-3-12b-it`
Gemma 4 12B (Gemma Team et al., 2026)	11.95B	`google/gemma-4-12b`
GLM-4V-9B (Hong et al., 2024; Team GLM et al., 2024)	9B	`THUDM/glm-4v-9b`
Idefics2-8B (Laurençon et al., 2024)	8B	`HuggingFaceM4/idefics2-8b-chatty`
InternVL2-8B (OpenGVLab Team, 2024; Chen et al., 2024)	8B	`OpenGVLab/InternVL2-8B`
InternVL3.5-8B (Wang et al., 2025b)	8.5B	`OpenGVLab/InternVL3_5-8B`
Janus-Pro-7B (Chen et al., 2025)	7B	`deepseek-ai/Janus-Pro-7B`
Kimi-VL-A3B (Kimi Team et al., 2025)	16B / 3B act.	`moonshotai/Kimi-VL-A3B-Instruct`
Llama-3.2-11B-Vision (AI, 2024)	11B	`unsloth/Llama-3.2-11B-Vision-Instruct`
LLaVA-1.5-7B (Liu et al., 2023a)	7B	`llava-hf/llava-1.5-7b-hf`
LLaVA-1.6-13B (Liu et al., 2024)	13B	`llava-hf/llava-v1.6-vicuna-13b-hf`
LLaVA-OneVision-7B (Li et al., 2024a)	7B	`llava-hf/llava-onevision-qwen2-7b-ov-hf`
MiniCPM-V 2.6 (Yao et al., 2024)	8B	`openbmb/MiniCPM-V-2_6`
MiniCPM-V 4.5 (Yu et al., 2025)	8B	`openbmb/MiniCPM-V-4_5`
Molmo-7B-O (Deitke et al., 2024)	7B	`allenai/Molmo-7B-O-0924`
Molmo2-8B (Clark et al., 2026)	8B	`allenai/Molmo2-8B`
Phi-3.5-Vision (Abdin et al., 2024)	4.2B	`microsoft/Phi-3.5-vision-instruct`
Phi-4-Multimodal (Microsoft et al., 2025)	5.6B	`microsoft/Phi-4-multimodal-instruct`
Pixtral-12B (Agrawal et al., 2024)	12B + 0.4B vision	`unsloth/Pixtral-12B-2409`
Qwen2-VL-7B (Wang et al., 2024)	7B	`Qwen/Qwen2-VL-7B-Instruct`
Qwen2.5-VL-32B (Bai et al., 2025b)	32B	`Qwen/Qwen2.5-VL-32B-Instruct`
Qwen2.5-VL-72B (Bai et al., 2025b)	72B	`Qwen/Qwen2.5-VL-72B-Instruct`
Qwen2.5-VL-7B (Bai et al., 2025b)	7B	`Qwen/Qwen2.5-VL-7B-Instruct`
Qwen3-VL-8B (Bai et al., 2025a)	8B	`Qwen/Qwen3-VL-8B-Instruct`
SenseNova-U1 (Diao et al., 2026)	8B	`SenseNova-U1/SenseNova-U1-8B-MoT-SFT`

6. **No refusals:** no "I cannot see the image"-type answer.
7. **No degenerate text:** (a) no token salad (more than 20% non-ASCII characters) and (b) no repetition loop (a sentence repeated three or more times, or any 8-gram occurring three or more times).

Image checks:

8. **Image–caption alignment:** the caption must belong to its own image. A scrambled or image-blind output passes checks 1–7 (complete, unique, non-empty, no leakage), yet pairs captions with the wrong images or invents scenes. We detect it by per-class label relevance, which lies between 0.3 and 0.7 for a healthy captioner and falls to the shuffled-caption floor (0.01–0.04) for an image-blind one.
9. **Image specificity:** the caption must describe the picture, not the class ("A shark is a fish whose skeleton is made of cartilage…"). Such captions are on topic and pass label relevance, but carry nothing specific to the image. We detect them by within-class retrieval: a caption that is the same for every image of a class cannot retrieve its own image among them, so retrieval falls to chance.

Checks 8 and 9 use the class label and act on whole captioners only: they admit or reject a captioner, and they removed or altered no individual caption in any evaluation set.

Caption length: Caption length and format are properties of a captioner, not defects: a terse captioner (about 42 words) and a verbose one (about 182 words) are both admitted. We check instead that each captioner's length distribution is the same on the training and validation splits.

Relation to prior checks: Prior work covers these failure modes separately: object-hallucination rates against annotations, reference-free consistency checks for faithfulness, measures of degenerate text, constraint checks for format, and CLIP-score filtering of web image–text pairs. Two of our checks have no standard counterpart. Check 8 detects image-blind and scrambled output without

Table 4: Stability of the captioner ordering across decoding configurations. Spearman ρ of the captioner ordering against the release cell (30 captioners, 2,000 paired images per cell). The ordering is invariant to the sampling knobs but not to the generation budget. The last column is the *per-image* correlation, whose own noise floor is the re-run row: individual captions are not reproducible even at identical settings, so every claim here is about aggregates.

cell varies	cells	ρ words	ρ TTR	per-image ρ
re-run at the release settings	1	0.998	0.998	0.266
temperature / top-p / top-k only	14	0.995	0.976	0.242
generation cap 600	12	0.991	0.975	0.238
repetition penalty 1.15	1	0.805	0.424	0.198
frequency penalty 1.15	1	0.933	0.727	–
generation cap 100	12	0.529	0.755	0.138

reference captions and with encoders separate from the captioner, whereas CLIP-score filtering uses a single shared embedding space. Check 9 rejects captions that are on topic but describe the class rather than the image. We combine the nine checks into one admission gate for a generated corpus of 30 captioners and 1.28M images, and release it with the corpus so that new captioners can be admitted under the same standard.

Repair: A captioner that fails a check is repaired where the failure is mechanical (pruning repetition, stripping leaked prefixes); the remaining failures are regenerated, and a small residual that still fails is dropped. A captioner enters the paper only when both of its splits reach the 99.99% coverage threshold.

C Robustness to Decoding and Prompt

An instrument is only useful if its readings do not depend on an arbitrary setting. We fixed one decoding specification for the corpus, and this appendix asks what that choice cost. To this end, we re-captioned a fixed set of $2{,}000$ validation images, two from each of the $1{,}000$ classes, with all 30 captioners under 42 decoding configurations: the $3 \times 2 \times 2$ temperature $\times$ top-p $\times$ top-k grid, the same grid at generation caps of 600 and 100 tokens, greedy decoding, wider and narrower nucleus and top-k settings, a repetition penalty, a frequency penalty, and a re-run at the release settings as a noise control. This amounts to $1{,}260$ cells and roughly 2.5 million captions.

Sampling settings: Table 4 reports the rank correlation of the captioners against the release ordering. Across every cell that varies only temperature, top-p or top-k, the mean ρ is 0.995 (minimum 0.978); whichever of these settings we had chosen, the captioners would appear in the same order.

Generation budget: Halving the cap to 100 tokens drops the correlation to 0.53. The budget is thus the one decoding choice that changes which captioner looks better, because truncation removes exactly the material on which long captioners differ. Captioners compared under different token budgets are therefore not compared on the same terms. A repetition penalty also moves the ordering, though less, and the quality gate absorbs much of its effect.

Prompt: The grid holds the prompt fixed, so we ran a small prompt-crossed check afterwards (specified before generation). Five captioners spanning the prototype ranking (Gemma 4 12B, InternVL3.5-8B, Qwen2-VL-7B, Qwen3-VL-8B and MiniCPM-V 2.6) captioned the 2,000 validation images of the decoding grid at release decoding under two other prompts, P1 "Describe this image in detail." and P2 "Write one or two sentences describing what is in this image.", with the same quality gate plus a prompt-echo check; 1,995 images are scored, and a second release-prompt sample gives the noise floor. The five captioners' ordering by mean caption length, which the second release-prompt sample reproduces exactly ($\rho = 1.0$), changes to $\rho = 0.3$ under P1 and -0.3 under P2. On a downstream use, test-time caption reading (Section 6, same combiner and candidates), the ordering is more stable: $\rho = 0.8$ under P1 and 0.4 under P2. Here each captioner contributes one caption per image, read with the combiner calibrated for three-caption pools; the class-bootstrap intervals, which reflect image sampling only, are $[0.5, 0.9]$ and $[0.3, 0.7]$, and with five captioners only

Table 5: The price and the payout of mixing captioners, per use, both as the mix minus a single captioner. *Premium*: the mix minus the captioner that is best for the use, resampled to the same number of captions (negative = what mixing costs). *Payout*: the mix minus a typical captioner (positive = what mixing protects against). Budget columns give captions (or prototype sets) per image for mix v reference; "avg." is the average of the ten resampled captioners; PR = pre-registered comparison; †not pre-registered. How "best" is chosen per use is in Table 1. Points with 95% class-bootstrap CIs; AUROC and precision $\times 100$; differences are computed before rounding. In the last row the pre-declared best captioner is the wrong one for that use, so its premium is against a mis-chosen captioner; the payout there is on the full training set.

Use		Premium	budget	Payout	budget, reference
Zero-shot prototypes	ViT-B/32	$+0.4$ [-0.5, +1.2]	6 v 6 sets (PR)	$+6.4$ [+6.0, +6.8]	6 sets v 1, mean single
Zero-shot prototypes	ViT-L/14	$+0.1$ [-0.9, +1.1]	6 v 6 sets (PR)	$+8.7$ [+8.2, +9.2]	6 sets v 1, mean single
Label errors (AUROC)		$+2.1$ [+0.1, +4.1]	30×1 v 1×30 (PR margin)	$+5.5$ [+4.5, +6.5]	10×3 v avg. 1×30
Missing labels (precision)		-0.6 [-3.8, +2.4]	10×3 v 1×30 (PR margin)	$+3.4$ [+1.7, +5.1]	10×3 v avg. 1×30
Primary label		-6.1 [-7.6, -4.7]	10×3 v 1×30 (PR margin)	$+4.8$ [+4.0, +5.7]	10×3 v avg. 1×30
Test-time caption reading	3 captions	-2.8 [-3.8, -1.8]	3 v 3 (PR)	$+1.4$ [+1.1, +1.8]	3 v avg. 1×3
Test-time caption reading	10 captions	-0.5 [-1.7, +0.5]	10 v 10 (PR)	$+3.3$ [+2.6, +3.9]	10 v avg. 1×10
Training a text encoder	ViT-B/32	-3.0 [-3.8, -2.2]	6 v 5.7 (PR)	$+12.2$ [+11.0, +13.5]	1 v 1, median single†
Training a text encoder	ViT-L/14	-2.0 [-2.9, -1.1]	6 v 5.7†	$+11.5$ [+10.3, +12.8]	1 v 1, median single†
Concept bottleneck	mis-chosen best	$+6.9$ [+6.3, +7.4]	6 v 5.7, 200k	-1.9 [-2.2, -1.6]	1 v 1, mean of 3 singles

$\rho \geq 0.9$ would reach one-sided $p < 0.05$. Qwen3-VL-8B is the best reader under every prompt, whereas Gemma 4 12B, the best zero-shot and training source at the release prompt, follows the other prompts poorly (16% and 20% of its captions fail the gate) and falls to last, 3.0 and 4.3 points below its own release-prompt resample. Which captioner is best therefore depends on the prompt as well as on the use.

Per-caption noise: The last column of Table 4 correlates *per-image* caption length between a cell and the release cell. Two runs at *identical* settings agree at only $\rho = 0.266$, because sampling at temperature 0.6 makes each caption a fresh draw. Every claim in this paper is therefore about aggregates over thousands of images, and the captioner-level stability (mean $\rho = 0.995$) does not imply any per-caption determinism. Cells at the 100-token cap fall below even this noise floor ($\rho = 0.138$), consistent with truncation being a real effect rather than resampling noise.

Serving versions: The corpus was produced with two vLLM versions, three minor versions apart. We measured the consequence on 20 paired cells with the same captioner, images and sampling, one copy per version: every paired word-count interval contains zero, and the largest absolute difference is 0.50 words on means of 79–120. At the resolution of our claims, the two versions are interchangeable.

What this appendix does not settle: The decoding grid measures caption *properties*, not what the captions are worth downstream, and the prompt check covers five captioners, two prompts and one downstream use; neither re-runs the training or zero-shot experiments under another prompt.

D Experimental Setup

Contrastive text-tower training: The image tower is a frozen OpenAI CLIP ViT-B/32 (or ViT-L/14) whose L2-normalised image embeddings are computed once. The text tower follows the CLIP text-transformer architecture (width 512, 12 layers, 8 heads, context 128 BPE tokens, causal attention, end-of-text pooling, a linear projection to the image-embedding dimension) and is trained from random initialisation with the symmetric InfoNCE loss and a learnable temperature, AdamW (learning rate 5×10^{-4}, weight decay 0.2, 10% linear warm-up then cosine decay), batch $4{,}096$, 10 epochs. On the 300k-image subset this is 730 optimiser steps for every arm (the complete-case arms of Table 7 run the same ten epochs on 225k images, 540 steps); each epoch samples one caption per image uniformly from that image's pool (one caption for B and D_1, six for D_6, up to six for E). Evaluation is ImageNet-validation zero-shot top-1 with the prompt "a photo of a {class name}." (first WordNet synonym). The reported epoch is the one with the best top-1 on 10,000 held-out training images (ten per class, disjoint from every training subset). Class-name mention rates count captions containing any name form of the image's class under the same rule as the class-name filter.

Table 6: Captions as contrastive training data. A text encoder is trained from scratch against a locked CLIP image encoder on the same 300k class-stratified ImageNet training images, up to six captions per image in every multi-caption arm (E: the release caption plus up to five samples, 5.7 on average), identical schedule; ImageNet zero-shot top-1 (%) with class-name prompts, mean of 3 seeds, epoch chosen on 10k held-out *training* images. *Names*: share of training captions that name the image's class (ViT-B/32 run). Bottom: differences in points with 95% class-bootstrap CIs of the seed-mean paired difference, computed before rounding; seed-level intervals are in Table 7. †Not pre-registered.

Training captions per image	ViT-B/32	ViT-L/14	names
(E) best captioner: release caption + up to 5 own resamples	46.4	58.1	57.5
(D_6) six captioners drawn from all 30	43.4	56.1	60.3
(D_6^{top}) the six next-best by prototype accuracy	35.6	48.7	61.3
(D_6^{top+}) the best + the next five by prototype accuracy	43.2	55.6	59.8
(B) best captioner, release caption only (1 per image)	41.5	54.0	57.5
1 caption per image† *:*			
(D_1) one random captioner per image	38.9	51.8	–
single captioner alone, median of 30 (range; single seed except the best)	26.7 (18.4–41.5)	40.3 (28.9–54.0)	–
E − D_6 (pre-registered on ViT-B/32; ViT-L/14†)	+3.0 [+2.2, +3.8]	+2.0 [+1.1, +2.9]	
B − D_6	−2.0 [-2.8, -1.1]	−2.2 [-3.1, -1.2]	
E − D_6^{top}	+10.8 [+9.6, +12.1]	+9.4 [+8.1, +10.8]	
E − D_6^{top+}	+3.3 [+2.6, +4.0]	+2.5 [+1.7, +3.3]	
D_1 − median single captioner†	+12.2 [+11.0, +13.5]	+11.5 [+10.3, +12.8]	
D_1 − best captioner alone†	−2.6 [-3.4, -1.8]	−2.2 [-3.1, -1.2]	

Test-time caption reading: For each validation image the unweighted vote supplies ten candidate classes. For each candidate we compute the word overlap between the image's captions (all thirty captioners, or the subset under test) and the candidate's profile of 47 characteristic nouns from the training captions, after removing every caption and profile noun that names any candidate. A conditional-logit model over the candidates, with the vote's log-probability and the overlap features as inputs, is fitted on a class-stratified half of the validation images and evaluated on the other half, over five splits.

Generation cost: Generation cost is taken from the logs of the release captioning runs, as captions per second per GPU over the whole run; tensor-parallel models count every GPU. Seven captioners whose training-run logs are lost were re-measured with their production recipes on five random samples of 600 training images, and scaled by the median ratio of production to benchmark rate of four captioners with full-run logs (0.93, range 0.81–1.05). Rates span 0.8 to 29.7 captions per second per GPU; the best captioner, Gemma 4 12B, runs at 8.5. Resampling it six times costs 0.70 GPU-seconds per image, six captioners drawn per image 1.51 (2.1×; 1.8× if the resamples are costed at their own measured rate), and only 2.3% of the 593,775 fixed six-captioner subsets are cheaper than the resampled captioner. Generation cost does not track training value (Spearman -0.27, $p = 0.15$). At the resampled captioner's cost a random mix affords 2.8 captions per image; one random caption per image (D_1, 0.36× that cost) already beats the median captioner by 12.2 points, but the median captioner resampled to the same cost (7.5 captions per image) was not trained, so a cost-matched insurance comparison is not measured.

Selection cost: Selection cost for test-time reading and label-error detection follows the zero-shot protocol: pick the captioner with the best score on n labeled calibration images per class and score it on disjoint images, over 500 draws. For reading, the selection uses only images outside the half on which the combiner is fitted. On classes disjoint from the calibration classes, $n = 25$ picks the best reader in 88.2% of draws (mean regret 0.07 points) and the best label-error detector in 89.2% (mean regret 0.14 AUROC points), so neither reaches 90% in the tested range, although the pick is within one point of the best in 97.8% and 90.8% of draws; within the same classes $n = 10$ suffices for 91.8% and 92.4%.

Missing resamples: The best captioner's five resample passes miss 6.1% of image slots (samples that failed the gate were not regenerated). The rate varies by class (median 5.8%, maximum 17.5%) but is unrelated to the difference between the arms (per-class Spearman -0.03, $[-0.09, 0.04]$; classes with high vs. low missingness -0.4 points, $[-2.0, +1.2]$).

Table 7: Training follow-ups (not pre-registered unless marked, most specified before running; ViT-B/32, 3 seeds unless stated, same schedule and held-out-train epoch rule as Table 6). *cc*: complete case, the images on which the best captioner has its release caption and all five resamples, so every arm has exactly six captions per image. *train*: captioners ranked by their accuracy as single training sources on held-out *training* images. Differences in points with 95% class-bootstrap CIs, computed before rounding.

Training captions per image	top-1
(E) best captioner, release + up to 5 resamples, all 300k images	46.4
(D_6) six captioners drawn from all 30, all 300k images	43.4
(E^{cc}) E on the 225k images with all six captions	45.0
(D_6^{cc}) D_6 on the same 225k images	42.2
(B^{cc}) release caption only, same 225k images	39.9
(D_6^{train}) the six best single training sources	44.3
(D_6^{train-}) the same ranking without the best	41.9
$E^{cc} - D_6^{cc}$	$+2.8$ [+2.0, +3.6]
$E - D_6^{train}$	$+2.1$ [+1.4, +2.8]
$E - D_6^{train-}$	$+4.6$ [+3.6, +5.5]
$D_6^{train} - D_6$	$+0.9$ [+0.4, +1.5]
$D_6^{train-} - D_6$	-1.5 [-2.2, -0.9]
further checks (amendments 5, 7–13):	
$E - D_6$, 5 paired seeds, seed-level t interval	$+3.2$ [+2.6, +3.9]
$E - D_6$, first 200k images (pre-registered replicate)	$+3.1$ [+2.3, +3.9]
$E - D_6$ on validation halves not used to choose the captioner (mean of 5)	$+3.0$
E − D at 1 / 2 / 3 / 6 captions per image (amendment 7)	$+2.6$ / $+2.7$ / $+2.7$ / $+3.0$
E with raw, ungated captions $- D_6$ (amendment 9)	$+2.7$ [+1.9, +3.5]
E raw − E gated	-0.3 [-0.6, -0.1]
$E - D_6$, image encoder trained from scratch, 3 seeds (amendment 8)	$+2.6$ [+2.1, +3.1]
$E - D_6$, flagged images dropped from both arms (amendment 11)	$+2.9$ [+2.1, +3.6]
$E - D_6$, each image's own class name stripped (amendment 11)	$+1.2$ [+0.4, +2.1]
$E - D_6$, captions capped at 64 tokens (amendment 13)	$+1.4$ [+0.6, +2.2]
$E - D_6$, captions capped at 48 tokens (amendment 13)	$+0.8$ [+0.0, +1.6]
$E - D_6$, image encoder trained 30 epochs, one seed (amendment 10)	$+5.0$ [+4.3, +5.7]
second-best captioner resampled $- D_6$ (50k, 120 steps, specified before running)	$+0.1$ [-0.5, +0.7]
$E - D_6$ on the same 50k, 120 steps, specified before running	$+5.5$ [+4.7, +6.3]
second-best captioner resampled $- D_6$ (50k, declared 720 steps; amendment 12)	-3.1 [-3.9, -2.4]
$E - D_6$ on the same 50k, 720 steps	$+2.7$ [+1.8, +3.5]
six captioners drawn once, fixed (three draws, seed 0)	40.2, 36.7, 44.9

E ZERO-SHOT PROTOTYPES

The main-text zero-shot comparison is the pre-registered symmetric control with true resamples (Section 5). This appendix reports the earlier analyses on the full training set, in which the one-captioner arm is built from ten disjoint caption subsets, together with their controls.

The class-name filter: A caption noun is removed from a class's prototype if it restates any synonym of that class (all WordNet lemmas and the ImageNet synonym list) under three rules: (i) it equals a synonym or any word of a multi-word synonym, including the plural forms *-s*, *-es* and *-y→-ies*; (ii) after lower-casing and deleting every non-letter, it equals a synonym treated the same way ("chain mail" and "chainmail", "lawn mower" and "lawnmower"), again with plurals; and (iii) any contiguous run of its words, joined in this way, equals such a synonym or its agentive form *-er*/*-ers* ("Leonberger", "snorkeler"). The first filter we used applied only rule (i) to single words and so missed compounds: 94–266 of the 1000 prototypes per captioner still contained the full name ("bald eagle", "tree frog", "chainmail"), and the leak count predicted a captioner's accuracy (Spearman 0.80). Rules (ii) and (iii) removed the remaining leaks, which cut accuracy by 7.2 points on average and changed the captioner ranking (rank correlation 0.75 between the old and the final accuracies); and an independent audit of the final prototypes found no class-name form in any of the 30 sets. The filter is a few dozen lines of code and is part of our code release.

Table 8: Downstream zero-shot accuracy by captioner: class-name-free prototypes built from that captioner's own training captions (every noun matching any synonym of the class, including compounds, removed), evaluated on the full 50k validation set with CLIP ViT-B/32. Identical images, identical protocol; accuracy in %. Spread best–worst 13.8 points (95% CI [12.3, 15.4], class-level bootstrap of the paired per-class difference between the two endpoint captioners, which are selected in-sample). The L/14 column re-scores the same prototypes with CLIP ViT-L/14: spread 16.1 points [14.2, 17.9] (Gemma 4 12B 36.5 to MiniCPM-V 2.6 20.4); captioner ordering Spearman 0.69 across backbones.

captioner	Top-1	Top-5	Top-1 (L/14)
Gemma 4 12B	29.8	58.3	36.5
InternVL2-8B	28.6	57.6	31.9
Aria	28.6	56.1	30.6
ERNIE-4.5-VL	28.5	55.6	32.8
InternVL3.5-8B	28.5	57.3	30.4
Janus-Pro-7B	28.2	56.0	33.5
DeepSeek-VL2-Small	28.2	55.2	31.1
LLaVA-OneVision-7B	28.0	55.4	32.2
Gemma 3 12B	28.0	55.9	33.9
Qwen2.5-VL-7B	27.8	55.3	32.7
Qwen2-VL-7B	27.8	55.3	33.1
Phi-3.5-Vision	27.6	56.3	31.9
Phi-4-Multimodal	27.6	55.5	32.3
Kimi-VL-A3B	27.4	54.1	31.9
SenseNova-U1	27.3	54.8	29.9
LLaVA-1.6-13B	27.3	54.3	31.9
Pixtral-12B	27.0	54.4	31.6
MiniCPM-V 4.5	27.0	54.4	30.5
GLM-4V-9B	26.9	54.9	32.9
Qwen2.5-VL-32B	26.9	54.6	32.0
Qwen3-VL-8B	26.9	54.1	33.4
Molmo-7B-O	26.1	53.6	30.0
Molmo2-8B	25.9	53.1	29.8
Fuyu-8B	25.3	52.7	29.3
LLaVA-1.5-7B	25.1	51.4	29.0
Qwen2.5-VL-72B	24.6	50.7	29.4
Idefics2-8B	24.3	48.9	26.9
Chameleon-7B	22.5	50.7	24.8
Llama-3.2-11B-Vision	22.0	47.3	27.6
MiniCPM-V 2.6	15.9	33.4	20.4

Table 9: Decomposition of captioner-by-class accuracy. Two-way variance decomposition of the 30×1000 captioner-by-class Top-1 accuracy matrix (50 images per cell, leak-free prototypes). The captioner main effect, which is what a global leaderboard captures, is 1.0% of variance, while the class×captioner *interaction* is 28.9%, about $30\times$ larger; subtracting the expected binomial noise leaves 26.4%. With CLIP ViT-L/14 the interaction is 34.6% (noise-corrected 32.8%). The per-class oracle's gain over the best single captioner is $+25.2$ points in-sample and $+21.7$ when the per-class choice is made on 25 images per class and scored on the other 25 (mean of 5 splits); a purely additive world with the same noise would produce $+8.6$. The oracle uses labels and is an upper bound, not a method.

variance component	ViT-B/32	ViT-L/14
captioner main effect (global ranking)	1.0%	1.0%
class main effect (difficulty)	70.2%	64.4%
class×captioner interaction (competence)	28.9%	34.6%
interaction, binomial-noise corrected	26.4%	32.8%

Captioner and class effects: Top-1 spans 13.8 points across the 30 captioners (Table 8). When the captioner×class accuracy matrix is decomposed, a global captioner ranking explains 1.0% of the variance and the class×captioner interaction 28.9% on ViT-B/32 (34.6% on ViT-L/14; Table 9); in

Table 10: Votes of different captioners against caption subsets of the best captioner. Unweighted soft vote (mean of per-captioner softmax, T=100) over leak-free prototypes, full 50k validation set. *Self-ensemble*: the best captioner's training captions split into 10 disjoint folds, one prototype set per fold, soft-voted, i.e. the same vote over one captioner's captions. Top-1 in %, differences in points. *10 random VLMs*: mean over 100 draws of 10 distinct captioners (draw sd in parentheses). Differences are paired per class; 95% CIs resample the 1,000 classes. The best single captioner is chosen with all labels (a reference, not a deployable choice). Panel B (ViT-B/32) repeats the comparison for six anchors across the ranking: for weaker anchors, 10 other VLMs beat their own self-ensemble by a wide margin.

A. Top-1 by CLIP backbone

	B/32	L/14
best single (Gemma 4 12B)	29.8	36.5
Gemma 4 12B, 10-fold self-ensemble	33.2	41.4
10 random VLMs	33.2 (0.4)	40.1 (0.7)
all 30 VLMs	33.9	41.2
all 30 − best single	+4.1	+4.8
95% CI (best re-selected per replicate)	[+3.0, +5.0]	[+3.4, +6.1]
10 random − self-ensemble	+0.0	-1.4
95% CI	[-0.8, +0.9]	[-2.4, -0.3]
all 30 − self-ensemble	+0.7	-0.2
95% CI	[-0.2, +1.6]	[-1.3, +0.9]

B. Six anchors, CLIP ViT-B/32 (500 random draws of 10)

anchor (rank of 30)	single	self-ens.	10 random	Δ	95% CI
Gemma 4 12B (1)	29.8	33.2	33.2	+0.0	[-0.8, +0.9]
ERNIE-4.5-VL (4)	28.5	30.3	33.2	+2.9	[+2.1, +3.7]
InternVL3.5-8B (5)	28.5	30.7	33.2	+2.5	[+1.8, +3.2]
Qwen2.5-VL-7B (10)	27.8	29.3	33.2	+3.9	[+3.2, +4.7]
LLaVA-1.5-7B (25)	25.1	27.3	33.2	+5.9	[+4.8, +7.0]
MiniCPM-V 2.6 (30)	15.9	18.7	33.2	+14.5	[+13.0, +16.0]

Table 11: Equivalence and superiority tests of the mix against the best captioner resampled. For each comparison: two one-sided tests for equivalence (margin ± 1.5 points, fixed before computing) and one-sided superiority tests in both directions, all 21 Holm-corrected together; class bootstrap. $\Delta > 0$ favours the mix; 95% CI. Zero-shot rows compare against the best captioner's ten caption-subset prototypes; the other rows use true resamples on 2,000 validation images, with the best captioner chosen per task on held-out classes.

			Holm p			
Comparison	Δ	95% CI	equiv.	div. >	res. >	reading
Zero-shot B/32: 10 random captioners vs. best, 10 subsets	+0.0	[-0.8, +0.9]	0.012	1.000	1.000	equivalent
Zero-shot B/32: all 30 vs. best, 10 subsets	+0.7	[-0.2, +1.6]	0.646	1.000	1.000	inconclusive
Zero-shot L/14: 10 random captioners vs. best, 10 subsets	−1.4	[-2.4, -0.3]	1.000	1.000	0.122	inconclusive
Zero-shot L/14: all 30 vs. best, 10 subsets	−0.2	[-1.3, +0.9]	0.202	1.000	1.000	inconclusive
Label errors (AUROC×100): 30 captioners vs. best×30	+2.1	[+0.1, +4.1]	1.000	0.357	1.000	inconclusive
Missing-label recovery: diverse vs. best×30	−0.6	[-3.8, +2.4]	1.000	1.000	1.000	inconclusive
Primary label: diverse vs. best×30	−6.1	[-7.6, -4.7]	1.000	1.000	0.010	resampled better

other words, which captioner performs well depends on the class. However, this per-class structure is not a stable property of the captioner: the per-class best captioner agrees across the two backbones for only 16% of classes, and, as the routing analysis below shows, the structure is hard to exploit.

Ensembles versus caption subsets of the best captioner: An unweighted vote, which averages each captioner's softmax over classes and takes the argmax, improves on the best single captioner by 4.1 points on ViT-B/32 (95% CI $[3.0, 5.0]$) and 4.8 on ViT-L/14 ($[3.4, 6.1]$). The earlier control splits the best captioner's own training captions into ten disjoint subsets, builds ten prototype sets, and votes over them. At matched size, this self-ensemble ties ten random *different* captioners on ViT-B/32 ($+0.04$ points, $[-0.8, +0.9]$), is ahead of them on ViT-L/14 (-1.4, $[-2.4, -0.3]$, not significant

Table 12: Per-class routing: usable complementarity. Top-1 (%) on held-out validation images, 25 calibration / 25 test images per class, 5 seeds, 95% class-bootstrap CI. The last row routes test images by their true class, which uses the test label; it is an upper bound, not a method.

Method	labels	ViT-B/32	ViT-L/14
Unweighted vote, 30 captioners	no	33.8 [32.0, 35.6]	41.3 [39.4, 43.3]
Best single captioner (chosen on calibration)	yes	29.7 [28.0, 31.4]	36.5 [34.6, 38.4]
Per-class router, routed by predicted class	yes	33.9 [32.2, 35.6]	42.1 [40.2, 44.0]
Per-class router, chosen on predicted-class buckets	yes	35.2 [33.6, 36.8]	44.0 [42.3, 45.8]
Same router applied by *true* class (not deployable)	yes	51.4 [49.7, 53.1]	61.7 [60.0, 63.4]

after correction), and all thirty do not beat it on either backbone (Table 10). The advantage of the subset vote over the captioner's full-data prototype is a prototype-averaging effect rather than an ensembling one: a single prototype set formed by averaging the ten subset embeddings scores the same as voting over them. Two further controls bracket the comparison. At a matched caption budget, i.e. ten different captioners each using a tenth of their captions, the difference is -0.4 points ($[-1.3, +0.5]$; -0.5, $[-1.6, +0.5]$ on ViT-L/14). Against ten bootstrap prototype sets built from all of the best captioner's captions, which recover less of the averaging gain, ten different captioners are ahead by 2.0 ($[1.0, 2.9]$) and 1.6 ($[0.5, 2.7]$) points, at ten times the caption budget. For the five anchors below the best, different captioners beat the anchor's own self-ensemble by 2.5–14.5 points, but mostly because the other captioners are simply better, not because they are different. Diversity does roughly double the ensembling gain over the average member: $+6.7$ points for ten different captioners, against $+2.9$ to $+4.6$ for one captioner's caption subsets, whose members each use a tenth of the captions. The consistent reading is that an ensemble protects against choosing a poor captioner, but does not supply what the best captioner, resampled, lacks.

Equivalence tests: A confidence interval that covers zero is not evidence of equivalence. We therefore fixed a margin of ± 1.5 points before computing and, for each of seven comparisons, tested equivalence and superiority in both directions, Holm-correcting all 21 tests together (Table 11). Equivalence is established for ten random captioners against the best captioner's subsets on ViT-B/32. One captioner, resampled, is better at identifying the primary label (-6.1 points, $[-7.6, -4.7]$). The other five comparisons are inconclusive, including ViT-L/14, where the resampled captioner is ahead but not significantly after correction, and label errors, where diversity is ahead by 2.1 AUROC points ($[+0.1, +4.1]$); the unadjusted interval excludes zero, but not after correction (Holm $p = 0.36$). After correction, no comparison favours diversity. The accurate statement for this family is therefore that diversity *never reliably beats* the best captioner resampled, not that the two are interchangeable everywhere.

Per-class routing: A natural way to exploit per-class differences is to route each class to its best captioner. Evaluated by grouping test images by their *true* class, which is impossible at inference, where the class is what is being predicted, such a router appears to reach 51.4% on ViT-B/32 (Table 12, last row), 17.6 points above the vote, but this is test-label leakage. Routing on the vote's *predicted* class, with per-class choices made on 25 labeled calibration images per class, recovers almost none of it: $+0.1$ points over the vote on ViT-B/32. Choosing each captioner on the calibration images that the vote *predicts* as that class does better, $+1.4$ and $+2.7$ points on the two backbones, a small but real gain that spends labels. For scale, a linear probe trained on the same labels directly on CLIP image features reaches 68.1%, far above every classifier built from caption prototypes (at most 35.2% on ViT-B/32, Table 12). The large apparent complementarity is therefore mostly the value of knowing the true class.

F Label Audits

Definitions: A *missing label* is a class that the human ReaL annotators added to an image beyond its ImageNet label; a proposal is a class named by at least k captions. A *label error* is an image whose original ImageNet label is absent from ReImageNet's re-annotation; it is scored by how few captions name the assigned label. The *primary label* of an image is the class named by the most

Table 13: Missing-label recovery by caption agreement, scored against human ReaL *added* labels only (the original ImageNet label is excluded, so proposing it earns nothing). The informative baseline is the *co-occurrence* null ('co-occ.'): decoys drawn per image from classes that actually co-occur with its own label. The ensemble is *below* this null at k=1 and only clears it from k=5 upward. The 'preds' column is the number of predictions the method makes at each threshold (TP+FP): precision rises because the method fires on ever fewer images, from 71,038 at k=1 to 168 at k=30. Precision, null, recall and F1 in %.

k of 30	preds	precision	co-occ. null	recall	F1	TP
1	71,038	9.6	17.6	58.1	16.4	6,785
5	17,947	25.1	21.0	38.6	30.4	4,506
9	10,539	34.0	21.7	30.6	32.2	3,578
10	9,472	35.8	22.0	29.0	32.0	3,387
15	5,781	45.2	22.4	22.4	29.9	2,613
20	3,551	54.6	23.1	16.6	25.5	1,939
25	1,873	67.5	23.0	10.8	18.7	1,265
29	504	86.5	23.6	3.7	7.2	436
30	168	91.7	23.4	1.3	2.6	154

Table 14: Class-level confound audit of the agreement label-error detector. AUROC for flagging the validation images whose original label ReImageNet's annotators reject (6,082 of 49,852 images, base rate 12.2%). *Agreement*: 1 minus the fraction of the 30 captioners that mention the label; *CLIP*: the negated image–label margin (ViT-B/32). **(a)** Per-image signals. *Within class* ranks an image only against images of the same class; *class mean only* replaces each image's signal by its class mean. Differences carry 95% class-cluster bootstrap CIs. **(b)** Class-level models: the class error rate alone, and a per-class intercept (fixed-effect, FE) logit with and without the signals. *Cross-fitted (2-fold for class means and rates, 5-fold for FE).

(a) Signal	Agreement	CLIP	Agreement − CLIP
Pooled AUROC	0.738	0.712	+0.027 [+0.007, +0.046]
Within-class AUROC	0.740	0.705	+0.035 [+0.011, +0.060]
Class mean only*	0.609	0.621	
(b) Class-level model	AUROC	Gain over FE only	
Class error rate only*	0.780		
FE only*	0.825		
FE + agreement*	0.870	+0.045	
FE + CLIP*	0.856	+0.031	
FE + agreement + CLIP*	0.880	+0.056	

captions (ties split), and it is correct when it equals the ImageNet label. Pooling different captioners spreads mentions over the classes that each captioner prefers, whereas one captioner's samples concentrate on the same class, which plausibly explains why the resampled captioner identifies the primary label better.

Missing labels and label errors: Scored against the human ReaL re-annotation on *added* labels only (the original label excluded), precision rises monotonically with the agreement threshold, to 91.7% at unanimity, against an informative co-occurrence null of 23.4% (Table 13). Ranking images by how few captioners name the assigned label predicts the labels that ReImageNet's annotators reject at an AUROC of 0.738, ahead of a CLIP image–label margin (0.712; difference $+0.027$, 95% CI $[0.007, 0.046]$ with classes as the resampling unit). The signal is a property of the image, not only of its class: within a class, it ranks wrong images above right ones at 0.740 (Table 14). On its own, however, it does not beat a class prior. ReImageNet's flags cluster strongly by class, and a class-error-rate prior estimated on held-out images alone reaches 0.780. Added to per-class intercepts (fixed effects), agreement raises the cross-fitted AUROC from 0.825 to 0.870 ($+0.045$), and it is blind in classes whose names captioners never use.

Comparison with one captioner, resampled: With ten captioners resampled 30 times each on the same images, we compare B different captioners against one captioner's B samples at the same budget. For label errors, 30 captioners score an AUROC of 0.747 against 0.726 for the best captioner's 30 samples, chosen on held-out classes ($+2.1$ points, $[+0.1, +4.1]$), which is not significant after correction (Holm $p = 0.36$). For missing labels, the diverse ensemble looks better at a matched agreement *threshold*, but that comparison is biased: one captioner's samples agree far more on the classes they name, so the same threshold is a looser filter for it, and it predicts more labels. At a matched number of *predictions*, against a captioner chosen on held-out classes, the diverse ensemble is not significantly different after correction on missing labels or label errors, although equivalence is not established either (Table 11), and it is worse at identifying the primary label (-6.1 points, $[-7.6, -4.7]$). Against an *average* captioner, it wins on all three. Diversity thus protects against picking a weak captioner, but does not reliably outperform a well-chosen one.

A learned aggregator (post hoc): The vote treats the class mentions of every caption alike. A Dawid–Skene aggregator (Dawid & Skene, 1979), fitted by EM without labels on the same captions with a per-captioner, class-conditional false-mention rate and applied to both arms at the same budget, raises the mix from the vote's 35.4% to 45.4% and the best captioner resampled from 41.5% to 48.4%. The premium therefore shrinks from -6.1 to -3.0 points ($[-4.4, -1.6]$; a one-coin model gives -5.8). This fit is transductive; fitted on one half of each class's images and applied to the other, the premium is -3.8 ($[-5.3, -2.4]$), so fitting on the evaluation images made it look 0.8 points smaller, and this is the number we quote in the main text. The largest learned false-mention rates belong to classes matched through ambiguous words (*light* for lighter, *orange* the colour for the fruit), noise that costs a mix of captioners more than one captioner. With the transductive fit, the payout over the average captioner grows from 4.8 to 8.8 points.

G Concept Bottleneck

Concepts and bottleneck: Captions are mapped to WordNet concepts (the nouns of the grounding pipeline, lemmatised). Every name form of *any* of the $1{,}000$ classes is removed from the vocabulary (compounds, fused compounds and plurals; 30% of concept mentions, including generic words such as *dog* and *water* that occur inside class names), and the $1{,}500$ most frequent remaining concepts form the vocabulary; none of them restates any class name. A concept bottleneck (Koh et al., 2020) is trained on frozen CLIP B/32 features, as a linear map from image to concepts, supervised by the captions' concepts, followed by a linear map from concepts to class; we use three seeds and identical settings in every arm. When the class head reads concept *probabilities*, the bottleneck leaks: concepts shuffled across all images still give 48.4% top-1, so that accuracy is mostly not concept information. Our primary head therefore reads *binary* concepts (p>0.2, set on held-out training images), for which the same shuffle falls to chance (0.3%; Table 15).

Mixing and concept supervision: Mixing captioners does not improve concept supervision. One random captioner per image reaches 48.7%, 1.9 points $[1.6, 2.2]$ *below* the mean of three single randomly drawn captioners (50.6%), and on the subset, six different captioners per image equal one random captioner (45.2% vs. 45.3%). The pre-declared best captioner is the weakest supervisor (39.0%; 9.7 points $[9.1, 10.2]$ below the mixture), but not because it is one model: after the class-name filter, its captions keep 7.9 vocabulary concepts per training image, against 12–15 for the others, and accuracy follows concept density. Ranking captioners by prototype zero-shot accuracy therefore does not transfer to supervising a name-free bottleneck. Shuffling concepts among images of the same class costs 6.6 points $[5.9, 7.2]$, so the bottleneck uses image-specific concepts, not only class-level concept statistics.

Resampling the best captioner: Resampling the pre-declared best captioner does not close its gap. On the same 200k images, its release caption plus up to five resamples per image that pass the gate reaches 38.4%, indistinguishable from the release caption alone (-0.1 points $[-0.4, +0.1]$) and 6.9 points $[6.3, 7.4]$ below six different captioners. The resamples keep the same 7.9 vocabulary concepts per caption (six captioners: 12.6), and the bottleneck activates 11.1 concepts per validation image against 21.1, so the gap again tracks concept density.

Table 15: Concept-bottleneck training on frozen CLIP ViT-B/32 features, the 50k ImageNet validation images, mean of 3 seeds. Concepts: WordNet nouns of the captions, every class-name form of all 1,000 classes removed (1,500 kept). *Hard*: the class head sees binary concepts ($p>0.2$), the primary metric; *soft*: it sees concept probabilities, which leak (concepts shuffled across all images still reach 48.4%). *Concepts/img*: active concepts per validation image (hard). Top-1 in %. Differences in the text are computed before rounding. Subset rows: 200k class-stratified training images, same number of optimiser steps.

concept supervision	hard top-1	soft top-1	concepts/img
linear probe, no bottleneck (ceiling)	–	72.0	–
(A) one random captioner per image	48.7	67.2	19.0
(B) Gemma 4 12B (pre-declared best)	39.0	64.3	9.9
(C) Kimi-VL-A3B	49.8	68.0	20.2
(C) Idefics2-8B	48.7	66.7	24.5
(C) InternVL3.5-8B	53.2	69.0	27.3
(A), concepts shuffled within class	42.1	65.0	12.4
(A), concepts shuffled across all images	0.3	48.4	3.1
subset: one random captioner	45.3	66.2	21.2
subset: Gemma 4 12B release caption	38.5	63.8	11.2
subset: six different captioners	45.2	65.6	21.1
subset: Gemma 4 12B release + up to 5 resamples	38.4	62.9	11.1

H Extended Related Work

Aggregating noisy labellers into labels: Treating the agreement of many imperfect annotators as a label is classical. Dawid & Skene (1979) estimate each observer's error rate and the latent true label by EM, and Snorkel denoises many programmatic labelling functions into training labels without ground truth (Ratner et al., 2017). Our captioners play the role of their observers, as frozen VLMs rather than crowd workers or heuristics. Unlike these methods, we deliberately use the simplest aggregator, a vote threshold, so that the reported curve is a property of the agreement signal rather than of a learned denoiser; Appendix F also reports a Dawid–Skene aggregator.

Caption utility and captions as supervision: Whether captions are useful downstream is an active question. CAPEval varies the caption source across ten captioners and correlates caption properties with downstream performance (Liu et al., 2026); CaptionQA asks whether a caption can substitute for the image (Yang et al., 2026); Caption Bottleneck Models build interpretable classifiers from captions of the actual images (Cagliyan et al., 2026); and, concurrently with us, Gare et al. (2026) argue that attributes should come from images rather than from class names. IC3 fuses multiple caption perspectives by committee consensus (Chan et al., 2023), and ICC scores caption concreteness without references (Yanuka et al., 2024). These works establish that the choice of captioner matters and that captions carry utility. By contrast, we ask whether *combining* captioners buys more than resampling the best one. One caution applies to this whole line of work, ours included: descriptions derived from captions of a class's own images can restate the class name in forms that a token-level filter misses (compounds such as "bald eagle", or fused forms such as "chainmail"). In our corpus, this inflated zero-shot accuracy by about 7 points and reordered the captioners (Appendix E). We have not re-run the pipelines of other papers and make no claim about them, but the check is cheap, and it is part of our code release.

Relabelling ImageNet, label errors and difficulty: ReaL re-annotated the ImageNet *validation* set with human multi-label ground truth and explicitly left the 1.28M training images unassessed (Beyer et al., 2020); Shankar et al. (2020) likewise multi-label a large validation sample, and work on label errors shows that the original labels are themselves imperfect (Northcutt et al., 2021b;a). Two lines of work automate the relabelling of the *training* set: ReLabel distils a strong classifier into dense multi-labels (Yun et al., 2021), and, concurrently, Chen et al. (2026b) make the full training set multi-label by unsupervised object discovery and report downstream *gains*. REVEAL renovates classification *test* sets with VLM consensus (Pang et al., 2025). Unlike these works, our label audits rely on the agreement between captions and an ontology, not on a classifier

or an object-discovery head, and we use them to audit the existing labels, not as training targets. Finally, our caption agreement is not intended as an example-difficulty score; that role belongs to classifier-agreement signals (Meding et al., 2021; Jiang et al., 2020; Ortiz-Jimenez et al., 2023).

I REPRODUCIBILITY

Serving: In this paper a captioner is a checkpoint together with the configuration that serves it. Most captioners were served with vLLM (Kwon et al., 2023) (version 0.26.0); those that SGLang serves more reliably were served with SGLang (Zheng et al., 2024b), and SenseNova-U1, which neither supports, with LightLLM. All captions of a captioner, including its resamples and its decoding-grid cells, come from one serving stack. The container images and the serving configuration of every captioner will be released with the corpus.

Serving versions: Part of the decoding grid was generated with vLLM 0.23.0 instead of 0.26.0; on 20 cells regenerated under both versions, word counts agree (largest difference 0.5 words), so we pool the two.

Pre-registration: Every pre-registered comparison and every amendment was committed to version control before the runs it governs, and we will release these files with their commit times. There are two exceptions. In the first training registration, one seed of the full-set arms had been seen before the file was committed, which the file discloses; and the re-measurement of seven captioners' throughput was committed 24 minutes after it started. The ViT-L/14 training comparisons are outside the pre-registered families.