GENERATION AS EVIDENCE: IMPROVING VISUAL UNDERSTANDING IN UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models support visual understanding and image generation, but whether generation provides useful evidence beyond strong understanding-only methods remains unclear. We study this question through caption discrimination: selecting the correct description of an image from two candidates. We score the observed image conditioned on each caption using the generation pathway, then combine generation and understanding scores with a linear combiner calibrated on a small labeled set. The pretrained model remains frozen, and no new images are generated. We compare against four progressively stronger understanding-only baselines, with the strongest using the same understanding scores, calibration data, and combiner. Across six unified models and a cross-model control on five compositional benchmarks, generation significantly improves 10 of 27 evaluated pairs by 1.2–8.1 percentage points over this baseline. Eight gains replicate in an independent implementation; uninformative replacements of generation scores yield no significant gains. The median gain falls from 4.8 to 1.1 points as understanding baselines are strengthened, showing why generation's contribution must be measured against strong controls. Localization, occlusion, and image-shuffling analyses show that generation scores depend on image content and concentrate on regions that distinguish the captions. These findings show that generation pathways can provide complementary visual evidence for caption discrimination within frozen unified models.

1 INTRODUCTION

Visual understanding models are widely used to evaluate and guide image generation (Lin et al., 2024b; Xu et al., 2023). The reverse direction has also attracted interest: generative scoring has been explored for image–text matching, self-supervision, and retrieval (Krojer et al., 2023; Hong et al., 2025; Ko et al., 2025). Unified multimodal models bring both capabilities into one set of weights, so both pathways can evaluate the same image–text pair. We ask: **can the generation pathway improve visual understanding beyond what a strong understanding-only system already achieves?**

We study this question through **caption discrimination** (Figure 1): given an image and two candidate captions, choose the one that matches. Consider an image of a red cube and a blue ball, and a foil caption describing a blue cube and a red ball, as in ColorSwap. The two captions contain the same objects and colors but bind them differently, so an understanding scorer that does not reliably associate each color with its object may score them similarly. Generation offers a second test of the same descriptions: conditioned on the color-swapped caption, the model may assign a lower score to the observed image, where the colors are associated with different objects. This intuition motivates testing whether generation scores resolve errors that understanding scores make.

Our method, U+G, evaluates each candidate in both directions. The understanding pathway scores the caption given the image, $\log p(c \mid x)$, and the generation pathway scores the observed image given the caption, $\log p(x \mid c)$. A linear combiner fitted on a small labeled calibration set combines the two kinds of scores; the pretrained model stays frozen, and no new image is generated. Because the two pathways use different prediction heads, token spaces, and training objectives, their scores need not agree, and we test whether this disagreement carries useful evidence.

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.

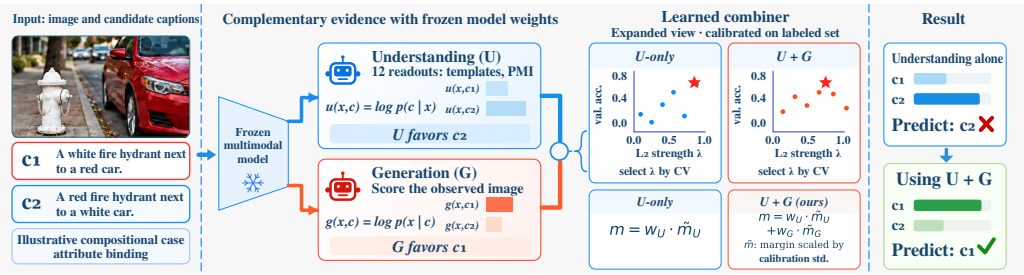


Figure 1: Generation as evidence for caption discrimination. A frozen unified model scores each candidate caption with its understanding pathway (U, $\log p(c \mid x)$) and its generation pathway (G, $\log p(x \mid c)$, scoring the observed image). A linear combiner fit on labeled calibration pairs combines the scaled margins; the understanding-only baseline uses the same combiner without G (§3.3). The example, score bars, and calibration plots are schematic.

Measuring that evidence requires an equally strong understanding-only comparison. Better prompts and combinations of understanding scores can improve caption discrimination without any generation, so a comparison against a single caption score can overstate the benefit attributed to generation. We therefore build four progressively stronger understanding-only baselines, ending with a learned combination of twelve understanding scores that uses the same understanding scores, calibration data, and combiner as U+G and differs only in lacking generation scores. Our main results measure what generation adds beyond this strongest baseline.

Generation adds evidence beyond the strongest understanding baseline. Across six unified models and a cross-model control on five compositional benchmarks, generation significantly improves 10 of 27 evaluated pairs by 1.2–8.1 percentage points over the strongest baseline, and no pair loses more than 0.6. On ColorSwap, the setting of our example, it improves three of the four unified backbones, by up to 4.8 points. Eight of the ten gains are reproduced by an independent implementation, and replacing the generation scores with shuffled scores or noise yields no significantly improved pair. Strengthening the understanding baseline reduces the median gain from 4.8 to 1.1 percentage points, so these are gains beyond the understanding methods tested here.

Gains follow two measurable conditions. Generation helps where the understanding baseline still makes errors and generation alone distinguishes correct from incorrect captions; across the 27 pairs, the two conditions together rank the gains (Spearman $\rho = 0.86$, descriptive). Improvements are largest for object, attribute, and spatial errors.

The generation scores depend on the image. On three autoregressive backbones, score differences between competing captions concentrate on the image regions that distinguish them; on Chameleon, occluding the named objects lowers generation accuracy more than occluding control regions of the same area; and replacing each image with a different one drives three tested generation configurations toward chance. These analyses support the image dependence of the tested scores, though they do not by themselves attribute the combined gains to visual content.

Our main contributions are as follows:

- **A controlled protocol for measuring generation evidence.** Understanding-only baselines that share the method's scores, labels, and combiner, and differ only in lacking generation, show how much of an apparent generation gain is recoverable by understanding alone (§3.3, §4.1).
- **Generation evidence beyond strong understanding baselines.** Significant gains over the strongest baseline that hold under an independent implementation and zero-information controls, together with two measurable conditions that the gains follow (§4.1–§4.3).
- **Visual grounding of generation evidence.** Localization, occlusion, and image-shuffling analyses that tie generation scores to the image regions distinguishing competing captions (§5).

2 Related work

Understanding as a critic of generation. VQA-based alignment evaluation (Lin et al., 2024b), human-preference rewards (Xu et al., 2023), and verifier-guided re-sampling use understanding to assess generation. We ask whether generation can supply evidence for understanding.

Generative classifiers. Generative classification is classical (Ng & Jordan, 2002); noisy-channel prompting scores text classes by $p(\text{input} \mid \text{label})$ alongside direct $p(\text{label} \mid \text{input})$ (Min et al., 2022); text-to-image models implement it by scoring $p(\text{image} \mid \text{class prompt})$ (Li et al., 2023a; Clark & Jaini, 2023; Jaini et al., 2024). DiffusionITM studies generative image–text matching and finds no evidence that its correct predictions complement those of CLIP or BLIP (Krojer et al., 2023), captioning-pretrained models score candidate captions by likelihood on benchmarks including SugarCrepe (Tschannen et al., 2023), and SelfEval scores real images under text-to-image models to evaluate their fine-grained text alignment (Rambhatla & Misra, 2024). SUDER uses reverse conditional likelihoods within unified models as training rewards (Hong et al., 2025), while BLiM combines text and video-feature likelihoods for retrieval (Ko et al., 2025). VisualGPTScore uses caption likelihood with a language-prior correction (Lin et al., 2024a). We measure what generative scoring adds within one frozen unified model, against understanding-only baselines built from the same readouts, labels, and combiner. These baselines include a PMI-style language-prior correction motivated by VisualGPTScore.

Ensembles, cascades, and weak-to-strong transfer. Unlike classifier ensembles (Dietterich, 2000), cost-based cascades and routers (Chen et al., 2023; Ong et al., 2024), and weak-to-strong training (Burns et al., 2023), we test which pathway of one frozen model supplies complementary evidence.

Does generation training help understanding? Unified models train understanding and generation jointly (Wu et al., 2025; Zhou et al., 2025), and MetaMorph asks whether generation *training* improves understanding (Tong et al., 2025); we ask whether the resulting pathway supplies evidence at *inference* time, which need not appear as a training-time gain. Mustafar fuses CLIP and conditional diffusion features and analyzes question-relevant regions (Agarwal et al., 2025), UniG2U-Bench evaluates generation-assisted understanding, including generating intermediate images before answering (Wen et al., 2026), and Unison scores generation-guided understanding among other synergies (Liu et al., 2026); we instead score observed images and test the additional value and visual grounding of those scores. Recent benchmarks find that the understanding and generation directions of unified models often disagree (Wang et al., 2026a;b); we treat this disagreement as a source of evidence and measure what it adds.

3 Method

Given an image and two candidate captions, we compute from frozen weights an understanding-side score, which evaluates the caption conditioned on the image, and a generation-side score, which evaluates the observed image conditioned on the caption. Our method, U+G, combines all understanding and generation margins with a learned linear combiner (§3.3). To measure what generation adds, we compare it with a ladder of understanding-only baselines of increasing strength that differ from U+G only in the absence of generation scores. Weights remain fixed throughout; target-task labels serve only to select readouts and fit combination parameters.

3.1 Understanding and Generation Readouts

For an image x and a candidate caption c, understanding (U) and generation (G) score the two conditional directions:

$$u(x, c) = \log p(c \mid x), \qquad g(x, c) = \log p(x \mid c). \tag{1}$$

Both pathways score each candidate independently. For candidates c_1 and c_2, a readout's signed margin is the score of c_1 minus that of c_2: its sign gives the prediction, its magnitude the strength of the preference.

G is not a rescaled U in practice. For an exact joint model, Bayes' rule gives $\log p(x \mid c) = \log p(c \mid x) + \log p(x) - \log p(c)$; since $\log p(x)$ is shared by both captions, the G margin would

equal the U margin minus the caption-prior difference, which is the PMI-corrected U margin already in our understanding pool (§3.2). A unified model estimates the two directions through different heads, token spaces, and objectives, so they need not agree (Wang et al., 2026a;b); a gain over the learned U-only combination measures evidence in their disagreement.

The generation-side estimator is architecture specific: autoregressive models use image-token likelihoods, Show-o a masked-denoising proxy, BAGEL a flow-matching proxy. Generation scores are therefore compared only within a backbone.

3.2 Understanding Readouts

The understanding side uses a pool of twelve U configurations: four caption forms—the original and three templates—each with two score aggregations, plus pointwise mutual information (PMI) correction of two caption forms under both aggregations. PMI subtracts the no-image caption likelihood to reduce language priors (Holtzman et al., 2021; Lin et al., 2024a). The generation side uses the four caption forms, giving four G configurations. The *selected ensemble* is the understanding-only system chosen on calibration data among the best single U readout, U–U gating (which follows a second U readout when its margin exceeds a calibrated threshold), two-readout fusions (a weighted sum of two scaled margins), and twelve-readout mean or majority votes (Appendix A.4.1). Prompted understanding formats are evaluated separately because prompt-format bias can confound U–G comparisons (Appendix B.5.4).

3.3 Learned Combination and Understanding-Only Baselines

Our method, U+G, fits one linear combiner over all sixteen margins. Let $\mathbf{m} \in \mathbb{R}^{16}$ collect the twelve U and four G margins of a caption pair, each divided by its standard deviation on the calibration set. U+G predicts c_1 when $\mathbf{w}^\top \mathbf{m} > 0$. We fit $\mathbf{w}$ by L_2-regularized logistic regression without intercept on the symmetrized calibration set $\{(\mathbf{m}, 1), (-\mathbf{m}, 0)\}$, so swapping the two captions flips the prediction; the regularization strength is chosen from $\{10^{-3}, \ldots, 10^{2}\}$ by five-fold grouped cross-validation within the calibration set, and the combiner is refit on all calibration pairs and frozen.

We admit a baseline only if it differs from our method solely in the absence of generation scores. An admissible baseline uses only the understanding readouts, calibration labels, data splits, and combiner available to U+G. We construct four increasingly strong understanding-only baselines: (i) the original-caption likelihood; (ii) the best single readout among the twelve; (iii) the selected ensemble of §3.2; and (iv) the *learned U-only combination*, the combiner above fit on the twelve understanding margins alone. At each level, adding generation means appending the four G margins to that level's understanding margin(s) and fitting the same combiner. Level (iv) is the strongest understanding-only system we obtain on aggregate (Appendix B.3.3) and shares the exact combiner of our method, so any stronger combiner would apply equally to both; we report our main results against it.

3.4 Evaluation Setup and Protocol

We evaluate six unified multimodal backbones: Show-o (1.4B) (Xie et al., 2025), Janus-Pro (7B) (Wu et al., 2024; Chen et al., 2025), Emu3 (8B) (Wang et al., 2024), BAGEL (7B active) (Deng et al., 2025), Chameleon (7B) (Chameleon Team, 2024), for which we use Anole-7B (Chern et al., 2024), a continued-training checkpoint with the same architecture that restores image generation, and Emu3.5-Image (34B) (Cui et al., 2025), a generation-specialized checkpoint. Emu3 uses its official Chat/Gen checkpoint pair; the other backbones share weights across U and G (Appendix A.1). We also pair Qwen3.8-27B (Qwen Team, 2026), which has no generation pathway, with Show-o as its generation model.

We evaluate two-candidate caption discrimination on SugarCrepe (Hsieh et al., 2023) (4,757 length-matched pairs), SugarCrepe++ (Dumpala et al., 2024), where we pair each image's human-validated second positive caption with the SugarCrepe hard negative, ColorSwap (Burapacheep et al., 2024), SPEC (Peng et al., 2024), and a 5,000-item subset of ARO (Yuksekgonul et al., 2023), giving 27 model–benchmark pairs. Within this five-benchmark panel, models and methods share five image-

or scene-grouped calibration/test splits; we report mean test accuracy. Dataset and diagnostic protocols are detailed in Appendix A.

Our main comparison is U+G against the learned U-only combination (§3.3). Simpler rules that add a single G readout to the selected ensemble by margin-based gating or score fusion, tested against every understanding-only addition and against G alone, are reported in Appendix B.2.

For every comparison, a split counts as *significant* only if, against *every* specified control, the G-augmented system has a positive paired accuracy difference, an exact two-sided McNemar $p < 0.05$, and a 95% whole-group bootstrap confidence interval for the difference entirely above zero. A model–benchmark pair is *significantly improved* (it *passes*) only when the same four or more of the five splits are significant. Because the splits overlap, this criterion measures stability, not experiment-wide error control (Appendix A.4).

4 Results against understanding controls

4.1 Generation beyond the strongest understanding baseline

Generation significantly improves 10 of 27 model–benchmark pairs over the strongest understanding baseline. Table 1 shows gains of 1.2–8.1 points: $+2.0$ (Show-o) and $+8.1$ (Emu3.5) on SugarCrepe; $+2.5$, $+1.2$ and $+4.5$ (Show-o, Emu3, Emu3.5) on SugarCrepe++; $+1.7$, $+3.0$ and $+4.8$ (Show-o, Emu3, Chameleon) on ColorSwap; $+7.3$ (Janus-Pro) on SPEC; and $+1.3$ (Janus-Pro) on ARO. Across all 27 pairs, no pair loses more than 0.6 points on average.

Table 1: Five-split mean accuracy (%) for learned U-only (U) and U+G; Δ: gain in points before rounding. Upper panel: shaded bold entries are significant on at least four splits; ‡: not significant in a second implementation. Bottom: four baseline levels (§3.3), with the strongest highlighted. G-alone scores and split counts: Appendix Table 7; all levels: Table 6.

Model	SugarCrepe			SugarCrepe++			ColorSwap			SPEC			ARO		
	U	U+G	Δ	U	U+G	Δ	U	U+G	Δ	U	U+G	Δ	U	U+G	Δ
Show-o	88.2	**90.2**	**+2.0**	86.7	**89.2**	**+2.5**	92.2	**93.9**	**+1.7**‡	76.2	75.7	−0.5	81.2	81.4	+0.2
Janus-Pro	93.9	94.3	+0.4	89.3	90.2	+0.9	95.6	96.7	+1.1	79.6	**86.9**	**+7.3**	81.6	**82.9**	**+1.3**‡
Emu3	91.2	91.8	+0.6	87.2	**88.4**	**+1.2**	91.9	**94.8**	**+3.0**	71.5	76.1	+4.6	81.5	81.4	−0.1
Chameleon	87.1	88.3	+1.2	81.9	83.1	+1.1	86.2	**91.1**	**+4.8**	51.4	54.7	+3.3	75.3	76.0	+0.8
BAGEL	96.5	96.9	+0.3	92.0	92.5	+0.5		—			—		86.4	86.5	+0.1
Emu3.5	74.7	**82.8**	**+8.1**	80.9	**85.4**	**+4.5**		—			—			—	
Qwen × Show-o	97.2	97.3	+0.0		—		97.8	97.9	+0.1		—			—	

Understanding baseline	(i) Original	(ii) Best readout	(iii) Ensemble	(iv) Learned
Significantly improved pairs	20/27	15/27	13/27	**10/27**
Median gain (points)	+4.8	+1.8	+1.4	**+1.1**

The reported gains are net of what the understanding side can recover. Moving from the original caption to the learned combination (Table 1, bottom rows), the number of improved pairs falls from 20 to 10 and the median gain from 4.8 to 1.1 points. Much of this change comes from the understanding side: on SugarCrepe, the template "A photo of {c}." alone raises Show-o's understanding accuracy from 0.740 (level i) to 0.852 (level ii), and the generation gain falls from $+9.3$ to $+3.1$. The template places the candidate in an explicit caption context and redistributes the model's probability over its wording, a known sensitivity of string-probability scoring (Holtzman et al., 2021). Beyond the best single readout the gain changes much less: on 7 of the 10 improved pairs, the gain at the strongest level retains at least 60% of its size over the best single readout, for example $+5.7 \rightarrow +4.8$ on ColorSwap/Chameleon and $+10.0 \rightarrow +7.3$ on SPEC/Janus-Pro.

Four checks support the 10 gains (Table 2). First, a second, independent implementation of the learned combination reproduces the count of 10 improved pairs, and 8 of them are improved under both implementations (Appendix B.3.3). Second, replacing the four G margins with zero-information columns of the same shape, permuted G margins or Gaussian noise, significantly improves 0 of 27 pairs in each of ten draws; these controls do not reproduce the observed gains (Appendix B.1). Third, generation changes answers in both directions, but on every improved pair it

Table 2: All 10 gains stay clear of the zero-information controls and fix more items than they break; 8 hold in a second implementation. Gain: points over learned U-only. Second impl.: improved in an independent implementation. Null: most significant splits in ten zero-information G replacements (four would count as improved). Fixed / broken: test items corrected / newly wrong (%).

Benchmark	Model	Gain	Second impl.	Null	Fixed / broken
SugarCrepe	Show-o	+2.0	✓	1/5	4.2 / 2.2
	Emu3.5	+8.1	✓	0/5	15.1 / 7.0
SugarCrepe++	Show-o	+2.5	✓	1/5	4.7 / 2.3
	Emu3	+1.2	✓	0/5	3.2 / 2.1
	Emu3.5	+4.5	✓	0/5	9.1 / 4.6
ColorSwap	Show-o	+1.7	×	1/5	2.7 / 1.0
	Emu3	+3.0	✓	1/5	3.9 / 1.0
	Chameleon	+4.8	✓	0/5	7.6 / 2.8
SPEC	Janus-Pro	+7.3	✓	1/5	11.4 / 4.2
ARO	Janus-Pro	+1.3	×	1/5	3.6 / 2.3
Total / mean			8/10	≤1/5	6.6 / 2.9

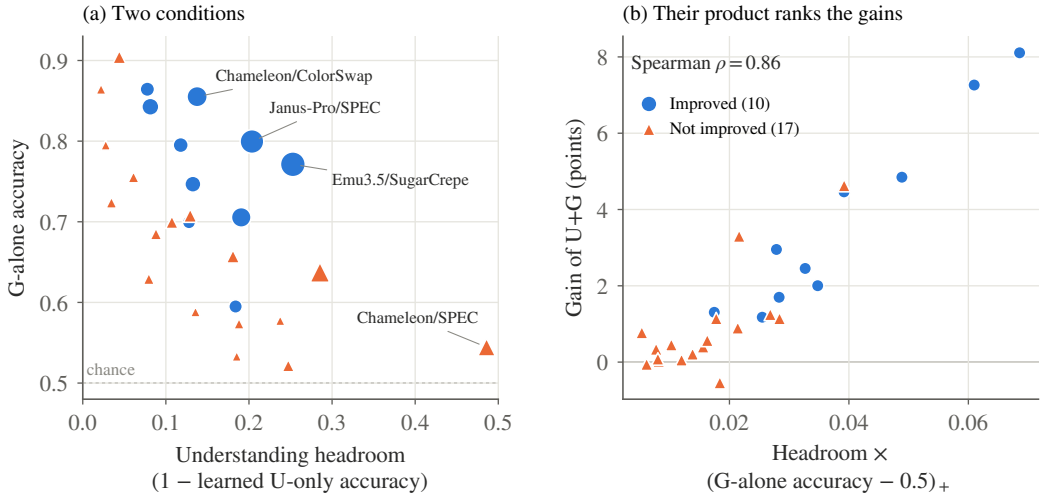


Figure 2: Where generation helps (27 pairs). (a) Understanding headroom versus G-alone accuracy; marker size encodes the gain over learned U-only. (b) Headroom times above-chance G accuracy ranks the gains ($\rho = 0.86$). Post hoc, descriptive analysis.

corrects more items than it breaks: 6.6% versus 2.9% on average (2.2:1), and corrections also outnumber new errors on 25 of all 27 pairs (Appendix B.4). Fourth, with caption length controlled, 2 of the 3 improved SugarCrepe++ pairs remain improved (§4.3).

The learned combiner is the stronger rule on both sides. Simpler rules that add one G readout to the selected ensemble also find gains: one-parameter score fusion, a weighted sum of the two scaled margins, significantly improves 10 of the 27 pairs, and margin-based gating, which follows G only when its margin exceeds a calibrated threshold, improves 6, each against the best understanding-only system built the same way and G alone. The learned U-only combination, however, beats that understanding-only system on 9 of 27 pairs by up to 20 points, and U+G exceeds one-parameter fusion on 13 of 27 pairs and is never significantly worse, so we report U+G against the learned baseline (Appendices B.2, B.3.3).

4.2 Where generation adds evidence

Two conditions, measurable on calibration data, rank where generation helps: understanding headroom and informative generation. Neither condition alone ranks the gains well across the 27 pairs (Spearman $\rho = 0.39$ for understanding headroom, 0.35 for the accuracy of generation alone), but their product does ($\rho = 0.86$, $p < 10^{-8}$; Figure 2), and the same holds when both conditions are measured on calibration data only ($\rho = 0.86$). The largest gains combine both: Emu3.5-Image,

a generation-specialized checkpoint whose learned U-only accuracy is the lowest on both of its benchmarks (also with prompting on SugarCrepe; Appendix B.5.5), and Janus-Pro on SPEC. None of the five pairs above 93.8% improves significantly, and a prospective Show-o test on a two-caption adaptation of CREPE, frozen before scoring, shows no gain at 96.65% (Appendix F). ARO leaves understanding headroom but its generation readouts reach only 0.52–0.60 alone. The product form was chosen after the two conditions were identified, so this analysis is descriptive (Appendix B.4).

Generation helps most on replaced objects and attributes and on spatial positions. Table 3 splits the gain by foil type. It is positive on at least half of the models for every type and largest for replaced objects and attributes on both SugarCrepe variants ($+1.6$ to $+3.0$ points) and for spatial positions on SPEC ($+3.7$). Relation replacements gain less ($+1.0$ to $+1.2$), and swaps of attributes between objects gain no more than replacements ($+1.0$ to $+1.4$), although ColorSwap, whose foils all rebind colors, gains $+2.1$ on average. Types differ in understanding accuracy and size, and no per-type test was run, so these differences are descriptive.

Table 3: Generation helps on every foil type for at least half of the models. Gain of U+G over learned U-only (points), averaged over models (five-split means); accuracies in %. Items: test items per split. Positive: models with a positive mean gain. Descriptive; no per-type test.

Benchmark	Foil type	Items	U-only	G alone	Gain	Positive
SugarCrepe	Replace attribute	389	89.4	73.9	+2.20	6/7
	Replace object	831	92.5	83.6	+2.33	7/7
	Replace relation	697	88.0	67.0	+1.15	5/7
	Swap attribute	326	90.2	72.9	+1.41	7/7
	Swap object	121	82.7	66.1	+1.87	5/7
SugarCrepe++	Replace attribute	389	88.3	67.8	+1.57	6/6
	Replace object	831	90.1	79.3	+2.99	5/6
	Replace relation	697	79.2	60.3	+1.02	6/6
	Swap attribute	326	89.5	66.1	+0.96	5/6
	Swap object	121	87.2	59.3	+0.33	3/6
ColorSwap	Color binding	718	92.7	86.6	+2.14	5/5
SPEC	Absolute position	1380	68.4	63.0	+3.67	3/4
	Relative position	232	77.3	69.8	+3.73	4/4
ARO	VG-Attribution	1281	92.3	59.3	+0.20	4/5
	VG-Relation	1250	69.9	53.1	+0.73	4/5

The same two conditions account for the benchmarks where generation does not help. Beyond the two COCO-based benchmarks, we screened four non-COCO candidates on a few hundred items each: ColorSwap and SPEC entered the panel, while What's Up (Kamath et al., 2023) and VALSE actions (Parcalabescu et al., 2022), run in full under the earlier single-readout gate, show no reliable gain, as a preregistered prediction had stated (Appendix B.6.5). Both nulls fit the two conditions of Figure 2: generation alone is near chance on What's Up (at most 0.553) and far below understanding on VALSE actions (0.59–0.64 versus 0.71–0.82). On Winoground (Thrush et al., 2022), a separate descriptive test, only one of five backbones improves on average (Appendix B.6.4).

U+G on a unified backbone exceeds dedicated image–text systems on four of five benchmarks. Against CLIP-L, NegCLIP, and BLIP-ITM, which use no target-task labels (NegCLIP and BLIP-ITM are fine-tuned on COCO), U+G on the best unified backbone of each benchmark, excluding the Qwen3.8-27B pairing, reaches 96.9%, 92.5%, 86.9%, and 86.5% on SugarCrepe, SugarCrepe++, SPEC, and ARO, versus at most 92.3%, 87.2%, 54.6%, and 71.6%, and trails only BLIP-ITM on ColorSwap (96.7% versus 97.4%). The margin comes mostly from the understanding side and from calibration labels, so these are reference points, not evidence for G (Appendix E).

Further diagnostics are in the appendix: interfaces (Appendix C.1), concept probes (Appendix C.3), and a cross-model test with Qwen3.8-27B (Appendix B.7).

4.3 Rewritten captions and caption length

Generation still helps when caption length is controlled. In our SugarCrepe++ pairing (Dumpala et al., 2024) the second positive captions were generated by an LLM and then validated and, where necessary, corrected by human annotators. Images and most hard negatives remain shared (82%

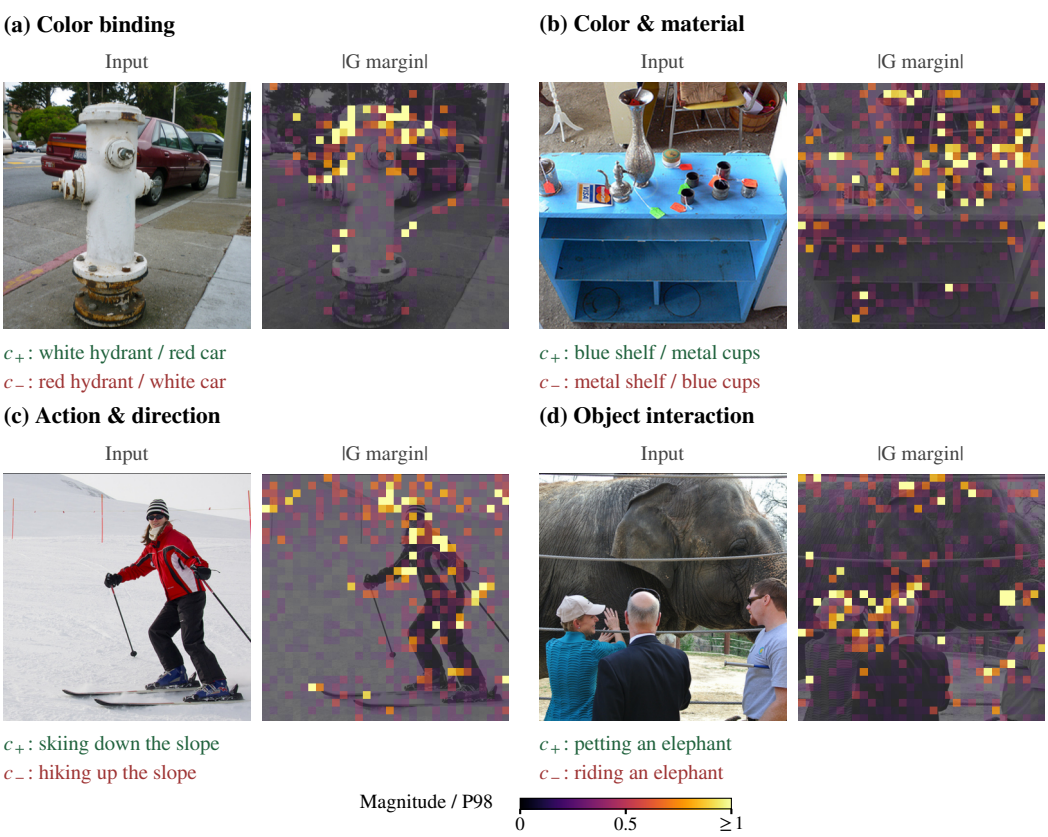


Figure 3: Chameleon cases with original-caption U wrong and G correct. Each pair shows the input (left) and $|\log p(t_i \mid c_+) - \log p(t_i \mid c_-)|$ overlay (right); c_+/c_-: correct/foil. Color and opacity encode magnitude; colors saturate at each map's 98th percentile. Captions are abbreviated; selected cases do not quantify gains.

verbatim), so this pairing tests caption robustness rather than a new image domain. The learned U-only combination adds 2–20 points over the selected ensemble here, and U+G improves three backbones over it by $+1.2$ to $+4.5$ points. Part of the understanding-side gain comes from caption length: rewritten positives are usually longer than the negatives, so choosing the longer caption alone reaches 81.3%, and summed likelihoods encode length. With the length difference added as an explicit feature to both sides (Table 4), generation still improves three models on all five splits, by $+1.8$ (Show-o), $+2.0$ (Chameleon) and $+3.2$ (Emu3.5) points.

Table 4: Caption-length control on SugarCrepe++ (five-split mean accuracy, %). +L adds caption-length difference to both combiners; gain compares the +L systems. Sig.: significant splits (§3.4); bold: at least four of five.

Model	U-only	U+G	Sig.	U-only+L	U+G+L	Gain	Sig.
Show-o	86.7	89.2	5/5	88.1	90.0	**+1.8**	5/5
Janus-Pro	89.3	90.2	3/5	90.5	91.1	+0.6	2/5
Emu3	87.2	88.4	4/5	88.5	89.4	+0.9	3/5
Chameleon	81.9	83.1	3/5	85.1	87.0	**+2.0**	5/5
BAGEL	92.0	92.5	1/5	92.9	93.2	+0.3	1/5
Emu3.5	80.9	85.4	5/5	84.3	87.4	**+3.2**	5/5

5 What visual evidence do generation scores contain?

Generation-score differences concentrate on the regions that distinguish the captions. On three autoregressive backbones, per-token score differences yield spatial maps. On SugarCrepe's swap-att items, the top twentieth of tokens holds 0.227–0.247 of absolute margin mass (uniform: one twentieth), with similar concentration on replace-rel. Figure 3 shows selected cases.

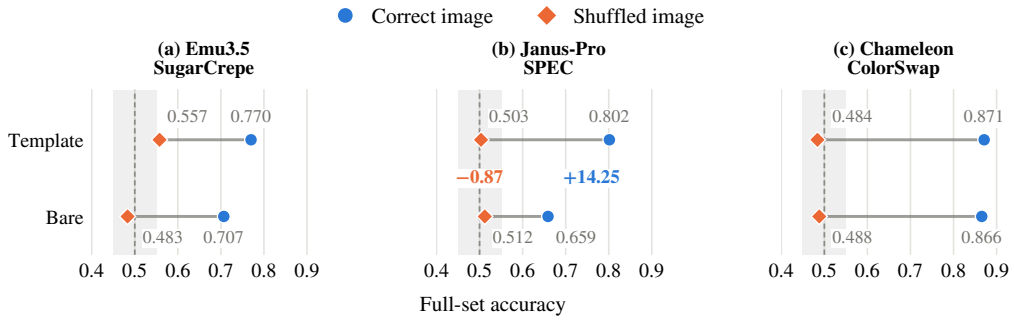


Figure 4: Full-set G accuracy with correct images (blue circles) and shuffled images from the same set (orange diamonds), with the caption template (Template) and without it (Bare). In (b), colored labels between rows give template-minus-bare gains (points). Shading marks the 0.45–0.55 chance band.

Occluding the named objects lowers generation accuracy more than occluding control regions. Two preregistered tests link these scores to named objects. Absolute margin mass in their boxes exceeds the area-normalized uniform expectation, and named objects have additional enrichment over unnamed ones, on every tested backbone. On originally correct Chameleon items, graying the named boxes reduces accuracy to 0.659/0.623 on swap-att/replace-rel, versus 0.770/0.744 for area-matched mirrored masks. The control-minus-target differences are 11.11/12.11 points, with whole-image 95% intervals $[2.90, 20.30]$/$[4.74, 19.30]$; boxes include background pixels (Appendix C.5).

Generation scores depend on the image. Shuffling images within each set drives three representative template-conditioned G readouts toward chance (Figure 4). The same readouts without the template all fall inside the chance band. Janus-Pro's template gain is image-contingent (+14.25 points with correct images versus -0.87 with shuffled ones), whereas Emu3.5 also has a substantial text-side contribution (+6.33 versus +7.42). Together with localization, these controls establish image dependence of the tested scores; they do not isolate the visual share of the residual gains.

6 Calibration and Deployment Requirements

U+G is label-efficient. Refit on random subsets (Table 5), U+G already exceeds U-only on all 10 improved pairs with 25 labeled pairs; with 100 pairs the median gain is 2.6 points versus 2.7 at full calibration.

Table 5: Median gain of U+G over learned U-only (points), refitting both on each label budget. All 10 pairs improved at full calibration retain a positive mean gain at every budget. Five-split means; no per-budget significance tests.

Calibration labels	25	50	100	200	400	Full
All 27 pairs	+0.8	+0.8	+1.3	+0.7	+0.9	+1.1
10 improved pairs	+2.0	+2.4	+2.6	+2.3	+2.5	+2.7

Generation adds four readouts per caption. Cost varies by architecture; a single U/G readout pair takes 0.26 s per item on Chameleon (one GPU; Appendix D).

7 Conclusion

A unified model's generation pathway is complementary evidence for understanding. Against equally informed understanding-only baselines, it significantly improves 10 of 27 model–benchmark pairs, most where understanding still errs, and its evidence depends on the regions that distinguish competing captions. Needing no new images or weight updates, it is a drop-in verifier for caption re-ranking and annotation screening. The evidence covers two-caption discrimination (POPE results: Appendix C.1); apart from the CREPE test, the benchmark comparisons are retrospective.

REPRODUCIBILITY STATEMENT

Section 3.3 and Appendix A specify models, scoring, exact templates, the combiner, and the statistical criteria. The numerical supplement provides the cached understanding and generation margins, image groups, calibration/test IDs, fitted configurations, and predictions for all 27 model–benchmark pairs, with CPU code that replays every comparison, including the four-level baseline ladder of Table 1 and its paired tests, without model inference. It also provides the localization and occlusion measurements with their recomputation code, and the frozen protocol, cached margins, and primary result of the prospective CREPE test (Appendix F).

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing, to draft sections of the paper, for retrieval and discovery of related work, and as coding assistants to help write and debug parts of the analysis code. The research question, method, experimental design, and interpretation of results are the authors' own; we have not used generative AI tools for the other tasks requiring disclosure under the ICLR 2027 AI policy. We have reviewed all AI-assisted work: the authors verified the manuscript content word by word, checked every cited related work against its original source, and all reported comparisons can be reproduced from the released code and cached scores. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vatsal Agarwal, Matthew Gwilliam, Gefen Kohavi, Eshan Verma, Daniel Ulbricht, and Abhinav Shrivastava. Towards multimodal understanding via stable diffusion as a task-aware feature extractor. *arXiv preprint arXiv:2507.07106*, 2025. URL https://arxiv.org/abs/2507.07106.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. ColorSwap: A color and word order dataset for multimodal evaluation. In *Findings of ACL*, 2024.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

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

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

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.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *Advances in Neural Information Processing Systems*, 2023.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025. URL https://arxiv.org/abs/2510.26583.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, 2000.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Advances in Neural Information Processing Systems*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. SUDER: Self-improving unified large multimodal models for understanding and generation with dual self-rewards. *arXiv preprint arXiv:2506.07963*, 2025. URL https://arxiv.org/abs/2506.07963.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, 2023.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, 2024.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *International Conference on Learning Representations*, 2024.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of EMNLP*, 2023.

Dohwan Ko, Ji Soo Lee, Minhyuk Choi, Zihang Meng, and Hyunwoo J. Kim. Bidirectional likelihood estimation with multi-modal large language models for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2507.23284.

Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? In *Advances in Neural Information Processing Systems*, 2023.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*, 2023b.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *International Conference on Machine Learning*, 2024a.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024b.

Jinyu Liu, Xincheng Shuai, Henghui Ding, and Yu-Gang Jiang. Unison: Benchmarking unified multimodal models via synergistic understanding and generation. In *International Conference on Machine Learning*, 2026.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5316–5330, 2022.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, 2002.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of ACL*, 2022.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. In *Proceedings of CVPR*, 2024.

Qwen Team. Qwen3.8-27B. `https://huggingface.co/Qwen/Qwen3.8-27B`, 2026. Model release.

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*, pp. 8748–8763. PMLR, 2021.

Sai Saketh Rambhatla and Ishan Misra. SelfEval: Leveraging discriminative nature of generative models for evaluation. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=0mGho8wrv5`.

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

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *International Conference on Computer Vision*, 2025.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. In *Advances in Neural Information Processing Systems*, 2023.

Chenlong Wang, Yuhang Chen, Zhihan Hu, Dongping Chen, Wenhu Chen, Sarah Wiegreffe, and Tianyi Zhou. Quantifying the gap between understanding and generation within unified multimodal models. *arXiv preprint arXiv:2602.02140*, 2026a. URL `https://arxiv.org/abs/2602.02140`.

Weixing Wang, Liudvikas Zekas, Anton Hackl, Constantin Alexander Auga, Parisa Shahabinejad, Jona Otholt, Antonio Rueda-Toicen, and Gerard de Melo. Beyond accuracy: Benchmarking cross-task consistency in unified multimodal models. *arXiv preprint arXiv:2604.25072*, 2026b. URL `https://arxiv.org/abs/2604.25072`.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Zimo Wen, Boxiu Li, Wanbo Zhang, Junxiang Lei, Xiaoyu Chen, Yijia Fan, Qi Zhang, Yujiang Wang, Lili Qiu, Bo Li, Ziwei Liu, Caihua Shan, Yifan Yang, and Yifei Shen. UniG2U-Bench: Do unified models advance multimodal understanding? *arXiv preprint arXiv:2603.03241*, 2026. URL https://arxiv.org/abs/2603.03241.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: A unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025.

GUIDE TO THE SUPPLEMENTARY MATERIAL

The supplement documents the paper's central result: against the strongest equally informed understanding-only baseline, the learned U-only combination, adding generation scores significantly improves 10 of 27 model–benchmark pairs by $+1.2$ to $+8.1$ points (Appendix B.1), 8 of them in both implementations (Appendix B.3.3). Zero-information replacements of the G columns pass on no pair (0/27); corrections and newly introduced errors account for 4.64% and 2.75% of all test items, respectively, on average; and the gains follow the product of understanding headroom and above-chance generation accuracy ($\rho = 0.86$, descriptive; Appendix B.4). Localization and occlusion show that G scores depend on the image regions that distinguish the captions (Appendix C.5). The sections establish:

- **Setup** (Appendix A): checkpoints, datasets, readouts, calibration, statistical criteria, and probe implementation.
- **Generation's incremental value** (Appendix B): the four-level baseline ladder, the learned combinations and their replication, where the gains arise, and generation gains that persist under stronger understanding controls and earlier single-readout protocols.
- **Visual evidence and interfaces** (Appendix C): localization, occlusion, error overlap, representation probes, and rendering; generation margins concentrate on named objects and lose more accuracy when those objects are occluded than under an area-matched control.
- **Practical requirements** (Appendix D): scoring costs, calibration budgets, configuration transfer, and deployment.
- **External references** (Appendix E): CLIP, NegCLIP, and BLIP-ITM on the same benchmark items, placing the U+G systems among standard image–text scorers.
- **Prospective evaluation** (Appendix F): the frozen CREPE protocol and its null result at a 96.65% baseline, the accuracy regime where the retrospective panel is also null.

Earlier gates and diagnostic tasks retain their own baselines, splits, and decision criteria. Their results are identified by protocol and should not be read as additional tests of the main learned U+G combination.

A EXPERIMENTAL SETUP AND IMPLEMENTATION

This section defines the models, data, readouts, and evaluation procedures used throughout the supplement. Caption-choice experiments evaluate likelihood-based systems; concept and verdict batteries use fitted representation probes, and rendering studies evaluate generated scenes under separate protocols. The main benchmark task is two-candidate caption discrimination. Open-ended VQA and captioning are outside the tested scope.

A.1 MODELS AND CHECKPOINT IDENTITIES

We evaluate six open unified backbones with weights frozen during likelihood scoring: Show-o (1.4B) (Xie et al., 2025), Janus-Pro (7B) (Wu et al., 2024; Chen et al., 2025), Emu3 (8B) (Wang et al., 2024), BAGEL (7B active) (Deng et al., 2025), Chameleon (7B) (Chameleon Team, 2024) through Anole (Chern et al., 2024), and Emu3.5-Image (34B) (Cui et al., 2025). Emu3.5 is generation-specialised and is evaluated on SugarCrepe and SugarCrepe++.

Checkpoint exceptions. Show-o, Janus-Pro, Chameleon/Anole, BAGEL, and Emu3.5 use one set of weights for both likelihood readouts. Emu3 instead uses the official Emu3-Chat and Emu3-Gen post-trainings of one pretrained model; its rows measure evidence added by a generation-specialised sibling checkpoint. Chameleon's concept and verdict probes use two separately trained, subsequently frozen LoRA adapters (Appendix A.5); its likelihood results use the same Anole weights for both paths. BAGEL's G score is a flow-matching proxy, not an exact likelihood. Show-o has the full initial order-bias, visual-dependence, and oracle controls.

Qwen3.8-27B (Qwen Team, 2026) is an external understanding-only model with no generation pathway. Its teacher-forced chat-template scores use thinking disabled; it receives separate-model G scores only in the cross-model experiments (Appendix B.7).

A.2 BENCHMARKS AND DIAGNOSTIC TASKS

Main panel. SugarCrepe (Hsieh et al., 2023) uses five length-matched slices (4,757 pairs); Appendix B.6.3 reports the full seven-slice set and the caption-length confound in the excluded slices. SugarCrepe++ (Dumpala et al., 2024) pairs an LLM-generated, human-validated second positive with the inherited SugarCrepe hard negative (82% verbatim), again on 4,757 pairs. All three captions were rescored. ARO (Yuksekgonul et al., 2023) uses 2,500 VG-Relation and 2,500 VG-Attribution items drawn with a fixed preregistered seed. ColorSwap (Burapacheep et al., 2024) uses AI-generated images and a separately preregistered validation pool excluding earlier probe groups. SPEC (Peng et al., 2024) tests spatial relations in synthetic images; its screening and full evaluations are in Appendix B.6.5.

Additional tasks. Winoground (Thrush et al., 2022) expands 400 groups into 800 caption-choice items: each image is paired with its correct caption and the group's other caption. It was designated descriptive before scoring because its 400-item test halves have limited power. POPE (Li et al., 2023b) tests object existence on 8,982 questions. Representation probes use 38 synthetic visual concepts (colour, position, counting, binding, and aggregate statistics) and nine verdict concepts, the latter covering 45 model–concept pairs, with held-out splits and null controls. The prospective CREPE test has its own protocol in Appendix F. Following a preregistered rule, each paired comparison keeps one fixed item set; results from different datasets or protocols are reported separately.

A.3 LIKELIHOOD READOUTS

Understanding scores the caption given the image, $\mathrm{U} = \log p(\text{caption} \mid \text{image})$; generation scores the observed image given the caption, $\mathrm{G} = \log p(\text{image} \mid \text{caption})$. The checkpoint exception for Emu3 is specified above. G uses each architecture's native estimator: masked-denoising pseudo-ELBO on Show-o, autoregressive likelihood on Janus-Pro, Chameleon, Emu3, and Emu3.5, and a flow-matching proxy on BAGEL. Scores are combined only within a model pairing; cross-backbone comparisons use accuracy, not raw score magnitude.

Both pathways score captions independently. On Show-o this yields order sensitivity 0.0042, versus 0.237 for prompted forced choice, which remains near chance after dual-order debiasing (0.471/0.508 on two templates). Per-caption yes/no prompting avoids that order bias but performs differently across backbones; Appendix B.5.7 reports the audit.

A.3.1 UNDERSTANDING SCORES

All are teacher-forced log-likelihoods of the caption given the image through each model's own understanding path; the primary normalisation is the per-token mean, the alternate is the raw sum. Show-o scores the caption through its multimodal-understanding pathway over MAGVIT-v2 image codes. Janus-Pro applies its chat template with the conditioning text "Describe the image." and scores the appended caption tokens. Emu3 (on its Emu3-Chat checkpoint) and Emu3.5 use the assistant template ("You are a helpful assistant. USER: [image] ASSISTANT:") with the caption scored after the assistant tag; Emu3.5 additionally uses its official separator tokens and resolution schedule, with flash attention (its vendored default attention path silently skips QK-Norm and was excluded after a sanity check; the supplement documents the incident). Chameleon (Anole) scores the caption in the single stream [image][caption]. Qwen3.8-27B uses its chat template with thinking disabled. The single-template repair in Appendix B.5 changes the caption string only: each caption is wrapped in ``` ``A photo of {c}.'' ``` before entering the same scorer, identically for the positive and the negative caption of every item, so the comparison stays a same-cost substitution. Its no-image control reuses the text-only scorer of the PMI baseline unchanged, with the same wrapping applied.

A.3.2 GENERATION SCORES

Each backbone is scored with the image-likelihood estimator its own architecture provides, so the family differs across backbones and we state each one. For the next-token backbones the readout is the mirrored teacher-forced likelihood of the true image's discrete tokens given the caption: 576 tokens for Janus-Pro (a 384^2 image through its generation head), 1,024 for Chameleon, and resolution-dependent counts for Emu3 and Emu3.5 under the official text-to-image template. Show-

o is a masked-denoising model with no autoregressive image factorisation, so its readout is a masked-denoising pseudo-ELBO instead: the real image's 256 MAGVIT-v2 codes are masked at eight frozen ratios (one tenth through four fifths, in steps of one tenth), with the masked position set drawn once from a fixed seed and reused for every item and both captions. The eight conditions are batched into one text-to-image forward call per caption, and the per-ratio masked-position log-probabilities averaged over the eight ratios. For these discrete-token readouts, the likelihood experiments use summed log-probability over the scored positions. On the autoregressive backbones the two captions score the same image over the same positions, so per-token normalisation would give identical decisions; on Show-o the masked-position counts differ across ratios and the two normalisations are therefore not interchangeable. We fix the summed form in the likelihood experiments; an earlier audit of fixed rules for choosing between prompted and likelihood readouts (the prompted-readout audit) uses its historical mean-normalised scores. BAGEL has no exact image likelihood; its proxy conditions the flow-matching head on the caption (same text-conditioning path as generation), VAE-encodes the real image (posterior mean), forms noised latents at five frozen timesteps (100–900 of 1,000, shared noise, seed 0 across the two captions), and scores minus the mean velocity-prediction error. The margin sign chooses the caption; its absolute magnitude ranks confidence for within-model calibration and gating. It is neither an exact log-likelihood nor a score calibrated across backbones.

A.4 Combination rules, calibration, and statistical criteria

The ensemble analyses and earlier single-readout experiments use different comparison protocols. We give the shared-score ensemble protocol first, then specify the earlier calibration procedures and reliability criteria so that their results can be interpreted under the appropriate comparison.

A.4.1 Ensemble controls and the fixed understanding system

The equal-label comparison, in which every method uses the same calibration labels, and the selected-ensemble comparison in Appendix B.3 cover all 27 model–benchmark pairs and their shared image-grouped splits, reuse its exact calibration/test IDs and five seeds, and make no new model calls. No examples are removed, and all methods in a pair receive identical calibration labels. SugarCrepe and SugarCrepe++ retain their shared-image assignments. The understanding pool is the twelve variants $\{t_0, t_1, t_2, t_3\} \times \{\text{mean}, \text{sum}\}$ and $\{t_0\text{PMI}, t_1\text{PMI}\} \times \{\text{mean}, \text{sum}\}$; the G pool has four caption templates. Their exact strings are

Template	String ($\{\texttt{c}\}$ is the verbatim candidate caption)
t_0	`{c}`
t_1	`A photo of {c}.`
t_2	`An image showing {c}.`
t_3	`A picture of {c}.`

The same string substitution is applied to both candidates and to U and G; original caption punctuation and the added template punctuation are retained. The U pool order is t_0, t_1, t_2, t_3, t_0PMI,t_1PMI, with mean before sum for each form; the G order is t_0, t_1, t_2, t_3. These orders resolve calibration ties. Prompted formats remain the separate audit in Appendix B.5.4; they are not silently added to this pool. Checkpoint identities are unchanged: Emu3 pairs Chat/Gen post-trainings, and Qwen rows use a separate Show-o generation model. The Qwen U controls use the receiver's readout pool, not an additional Show-o understanding readout.

Understanding-only controls. Six understanding-only methods, calibrated on the same labels as every G system, form the controls. Select the best single U by calibration accuracy, with ties resolved by the frozen pool order. On that base we fit a U–U gate and a fixed-base U fusion; another control jointly selects two distinct U readouts from all 66 pairs and their fusion weight. Further controls average all scaled U margins or take a majority vote (a tied vote retains the selected single U). The selected ensemble is chosen among these six methods by calibration accuracy, including the single U, with ties preferring the earlier method in that order. Selection uses no test labels. Table 9 reports every component, so calibration selection cannot hide a stronger tested control.

Gate and fusion rules. Gates initialize at zero delegation and traverse the frozen readout order and increasing coverage grid $\{0, .05, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1\}$. Only strictly higher calibration

accuracy replaces the incumbent. A nonzero coverage becomes a calibration quantile of the auxiliary readout's absolute raw margin; that threshold is fixed on test. For signed margins a, b, fusion uses

$$f_w = w\frac{a}{\sigma_{a,\mathrm{cal}} + 10^{-12}} + (1-w)\frac{b}{\sigma_{b,\mathrm{cal}} + 10^{-12}}, \qquad w \in \{1, .95, \ldots, 0\}. \tag{2}$$

Scores are scaled without subtracting their means. The default is the base alone; ties retain the earlier readout and larger base weight. Both normalizations and all choices use calibration data only. Cached margins use positive-caption minus negative-caption coordinates for evaluation; the zero threshold is the original caption-choice rule. Exact zero retains the original strict-positive correctness convention. No intercept or test normalization is fitted.

The fixed selected ensemble. After freezing the selected ensemble, we calibrate ensemble+U-gate, ensemble+U-fusion, ensemble+G-gate, and ensemble+G-fusion using the same families above. A further U-only system selects from the selected ensemble and its two U augmentations on calibration accuracy; a G system analogously selects the selected ensemble, G gating, or G fusion. Both are reported, and gate and fusion are reported as separate rules; neither verdict overrides the other. For fusion, the selected ensemble supplies its selected method's continuous margin: the scaled single margin, the original weighted/mean margin, the scaled branch margin from a U gate, or the signed vote difference with the single-U decision breaking ties. Its sign is checked against the original selected-ensemble predictions on calibration and test. It is then scaled by its own calibration standard deviation. The same labels fit both selection stages; increased selection complexity can still overfit them.

Statistics. Differences are paired on the same test items, and a reliable gain must replicate across splits. Exact two-sided McNemar tests are accompanied by 2,000 whole-image/scene bootstrap resamples, using the original split seed and retaining pair-weighted accuracy. A positive split requires positive difference, $p < .05$, and a 95% interval wholly above zero; negative splits are defined symmetrically. Reliability requires at least four of five. For the selection-based comparison, every specified U control and independently calibration-selected G-alone must be beaten on the *same* four or more splits. Failed significance is not equivalence. The overlapping splits provide stability checks, with no experiment-wide error-rate claim.

A.4.2 Earlier single-readout comparisons

The initial protocol compares the gate with U alone, G alone, and an alternate-normalisation second opinion: the same gate applied to the alternate length normalisation of U (the summed caption log-likelihood in place of the per-token mean; the renormalisation reuses the same forward pass, so it adds no model inference—Appendix A.3), using that reader's own margin. A preregistered follow-up adds a true second inference and a PMI-debiased understanding baseline (pointwise mutual information: the no-image caption likelihood is subtracted, the standard correction for surface-form and language priors (Holtzman et al., 2021; Lin et al., 2024a)). The stronger of original-caption U and PMI is selected on calibration data only, and the generation gate is recalibrated against that base. Five fixed splits and paired McNemar tests assess each comparison. Image-grouped splits additionally keep every caption pair for an image on the same side of calibration and test and use group-bootstrap confidence intervals. The initial splits provide a stability check; they are not five independent datasets. Visual-dependence ablations and the prompted-readout audit test whether the scores reflect image evidence.

The original-caption protocol fixes item order and permutes items once per split, using the first half for calibration and the second for testing. The best-single-readout gate protocol instead splits whole image/scene groups, with assignments shared across models as stated in §3.4; its item counts can differ between halves. The coverage grid is 0%, 5%, 10%, and every further 10% step to 100%; the calibrated coverage is converted to an absolute margin threshold (the matching quantile of the generation margins on the calibration half) and frozen before the test half is scored. Accuracy comparisons on the grid use a strict inequality, so ties resolve toward the smaller coverage and zero coverage (never delegate) wins unless delegation strictly helps. Significance is the exact two-sided binomial McNemar test on discordant pairs. The second-opinion control reuses the same forward pass under the alternate normalisation and therefore adds no model inference. It is a score-reuse control, not a compute-matched inference budget; the input-variant second read (Appendix C.1) is a true second inference.

The main benchmark task is a two-candidate caption choice; the concept, verdict, and rendering diagnostics use separate tasks. Open-ended VQA and captioning are out of scope. Paired comparisons share fixed items and calibration/test splits. The original-caption criterion requires significant positive exact two-sided McNemar tests against both the base and an alternate-normalisation second-opinion gate on at least four of five splits. The subsequent best-single-readout protocol selects one of twelve understanding variants and compares its G gate with that base. It does not by itself test superiority over an understanding ensemble. On its 27 pairs, the retrospective joint reliability rule also requires a whole-group bootstrap 95% interval above zero on the same four or more splits: 12 pairs pass the McNemar-only rule and 11 pass the joint rule, with SPEC/Chameleon the sole disagreement.

A.5 Representation-probe implementation

These probes measure information in hidden representations under fitted diagnostic readouts. Their outcomes appear in Appendices C.3 and C.1.

A.5.1 Concept set

U and G here denote linear class-mean difference probes of hidden image-token representations, pooled across tokens and a frozen late-layer band, not the benchmark likelihood readouts. Their axes and midpoint thresholds are fitted on calibration labels. Chameleon uses the independently trained LoRA adapters for understanding and generation objectives; each is frozen when extracting the diagnostic representations. The main Chameleon likelihood experiments instead use the same Anole weights for both paths. Stimuli are procedurally rendered 512×512 scenes of coloured shapes on a plain background, organised as minimal pairs: 48 positive and 48 negative images per concept, with a fixed contrast polarity declared per concept, so "positive" means the same thing in every pair. Shapes, colours, sides, and positions vary between pairs for diversity but are identical within a pair except for the concept itself; painted area and mean luminance are matched within pair by construction (computed, not eyeballed, and audited on every rendered pair before any model run) except where the concept is itself a quantity; the prompt is identical within pair. The 38 evidence concepts span colour, position, counting, attribute binding, and aggregate statistics, with rule-based ground truth throughout. Each concept is evaluated over 200 random half-splits (fixed seed schedule); axes are class-mean difference directions; arbitration replaces the lowest-confidence 30% of understanding-probe decisions on each test half, ranked by absolute distance to its fitted midpoint; a concept counts as helped only if the gated accuracy exceeds both the understanding baseline and a bootstrap self-arbitration null by at least five points, and the preregistered multiplicity audit of Appendix C.3 estimates the upper reference quantile of null pass counts per architecture under that bootstrap control.

A.5.2 Verdict set

These are class-mean difference probes of suffix-last hidden states, not decoded judgments. The same low-U-confidence 30% rule arbitrates between the probes, with a five-point improvement required over both U and bootstrap self-arbitration. Chameleon again uses the LoRA pair. Each of the nine verdict concepts corresponds to one error class of a claim about a scene. Stimuli reuse the same rendering family: 64 base scenes, each paired with claim variants under two prompt templates, give 128 positive items per concept (claims exhibiting that error) against 1,024 hard negatives (the correct claim and the other error classes over the same scenes). Because the variants of one scene share its content, the 200 half-splits are grouped by base scene (32/32; a scene's variants never straddle the calibration–test boundary), and all verdict accuracies are balanced accuracies to correct the class imbalance.

A.6 Emu3.5 prompted-readout implementation

The Emu3.5 extension in Appendix B.5.5 uses the following native prompt formats and inference checks.

Each A/B query uses the native sequence `BOS + system/USER + image tokens + question + ASSISTANT + BSS`, with the same helpful-assistant system text as the yes/no

audit. The primary query is "Which caption matches the image? A: {a} B: {b} Answer with A or B."; the two alternatives are "Look at the image. Which description is correct? A: {a} B: {b} Reply A or B." and "Choose the caption that best describes the picture. Option A: {a} Option B: {b} Answer." The positive caption's position alternates by original item index; the swap format reverses it. At the first answer position we compare the maximum logits over `A/␣A` versus `B/␣B`; ties count as incorrect. Yes/no instead scores each caption separately with "Does this caption correctly describe the image? Caption: {c} Answer yes or no.", comparing its maximum `Yes/yes` logit minus maximum `No/no` logit. The caption with the larger yes-minus-no margin is selected. The same checkpoint, image preprocessing, bf16 precision, and FlashAttention-2 are used throughout. After prefix caching failed a frozen numerical check, all new scores were recomputed with full-sequence forward passes; partial cached scores were excluded. All 20 repeated yes/no logit checks matched exactly. The new four-format inference used 1.867 GPU-hours on one NVIDIA L20X, including the discarded startup attempt (planned: 0.5–2.5 GPU-hours). An independent CPU implementation verified 47,308 held-out predictions and 70 paired statistics and group intervals. The original twelve-readout and twelve-plus-yes/no results were reproduced exactly.

B Incremental value of generation evidence

Generation evidence adds reliable accuracy beyond the strongest understanding-only baseline: against the learned U-only combination, U+G improves 10 of 27 pairs by $+1.2$ to $+8.1$ points. These results are documented by the baseline ladder (Appendix B.1), the fitting and replication details (Appendix B.3.3), and the exploratory gain analyses (Appendix B.4). The other subsections report alternative combination rules and earlier controls. In particular, a gain from a single-readout gate is relative to its stated U base, not the learned U-only baseline used in the main text.

Terms used in this appendix. The four baselines of §3.3 appear as the *original-caption* readout (level i), the *best single readout* (level ii), the *selected ensemble* (level iii), and the *learned U-only combination* (level iv); U+G is the main method. *Ensemble+U* and *ensemble+G* add one understanding or generation readout to the selected ensemble by margin-based gating or one-parameter fusion, and *single-readout gates* add a generation readout to the original-caption or best-single-readout base. U and G denote the understanding and generation pathways.

B.1 Baseline ladder

Generation adds significant gains at every level of the ladder, including level (iv), the strongest understanding-only baseline. The ladder applies one combiner at every level, so only the strength of the understanding baseline changes. For levels (i)–(iii) the understanding side is a single margin column: (i) the original-caption likelihood, with mean or summed aggregation chosen by calibration accuracy; (ii) the most accurate of the twelve understanding readouts on the calibration set; and (iii) the margin of the selected ensemble as constructed in Appendix A.4.1, which is the continuous margin for single-readout and fusion systems and the signed vote difference for majority vote. The understanding-only prediction at these levels is the sign of that margin. U+G appends the four generation margins and fits the combiner of §3.3. Level (iv) is the learned combination of Appendix B.3.3. All levels use the same five calibration/test splits and the significance criterion of §3.4, with 2,000 whole-group bootstrap draws per split.

Level (iv) reproduces the learned-combination analysis. Its set of 10 improved pairs matches Table 13, and its accuracies agree to within 0.001 (rounding). Level (i) accuracies can differ from the original-caption results of Appendix B.5, which use their own gating rule and protocol. The ladder is a retrospective reanalysis of the packaged margins, computed after the other comparisons were known; it adds no model inference and is reported for all 27 pairs. Table 6 gives the per-level accuracies.

Zero-information generation features pass on no pair. To test whether adding four uninformative features produces comparable gains, we rerun level (iv) with the four G columns replaced by features of the same shape that carry no information about the item: (a) G margins permuted across items, independently in the calibration and test sets, each with a random sign (the correct caption is always c_1 in the packaged margins, so an unsigned permuted column would act as an intercept); and (b) i.i.d. standard Gaussian columns. Combiner, grid, splits, and criterion are unchanged; each control

uses five independent draws. No pair passes in any of the ten draws (0/27 each). Across draws, 1–2 of the 135 pair-splits reach individual significance under (a) and 0–3 under (b), and the median gain is -0.04 points under both (range -0.74 to $+0.45$). The pass rule for these controls was written before they were run; specification and code are in the numerical supplement.

Table 6: Generation adds significant splits at every level of the ladder, including the strongest level (iv). Baseline ladder in full: U-only / U+G accuracy (five-split means) at each level, with the number of significant splits in parentheses (§3.4). Levels as in §3.3; main-text Table 1 reports level (iv) per pair and the per-level summaries.

Benchmark	Model	(i) Original caption	(ii) Best single readout	(iii) Selected ensemble	(iv) Learned combination
SugarCrepe	Show-o	0.740 / 0.833 (5)	0.852 / 0.883 (5)	0.870 / 0.893 (5)	0.882 / 0.902 (5)
	Janus-Pro	0.811 / 0.857 (5)	0.909 / 0.922 (4)	0.927 / 0.930 (0)	0.939 / 0.943 (0)
	Emu3	0.762 / 0.805 (5)	0.900 / 0.914 (5)	0.907 / 0.913 (2)	0.912 / 0.918 (2)
	Chameleon	0.658 / 0.717 (5)	0.782 / 0.812 (5)	0.837 / 0.854 (4)	0.871 / 0.883 (3)
	BAGEL	0.839 / 0.861 (5)	0.935 / 0.938 (0)	0.957 / 0.958 (0)	0.965 / 0.969 (1)
	Emu3.5	0.607 / 0.786 (5)	0.706 / 0.806 (5)	0.711 / 0.807 (5)	0.747 / 0.828 (5)
	Qwen × Show-o	0.797 / 0.887 (5)	0.942 / 0.949 (1)	0.962 / 0.964 (0)	0.972 / 0.973 (0)
SugarCrepe++	Show-o	0.734 / 0.805 (5)	0.760 / 0.810 (5)	0.777 / 0.825 (5)	0.867 / 0.892 (5)
	Janus-Pro	0.787 / 0.823 (5)	0.804 / 0.829 (5)	0.841 / 0.858 (5)	0.893 / 0.902 (3)
	Emu3	0.773 / 0.821 (5)	0.773 / 0.821 (5)	0.800 / 0.833 (5)	0.872 / 0.884 (4)
	Chameleon	0.653 / 0.706 (5)	0.695 / 0.734 (5)	0.706 / 0.740 (5)	0.819 / 0.831 (3)
	BAGEL	0.781 / 0.816 (5)	0.886 / 0.889 (0)	0.900 / 0.907 (2)	0.920 / 0.925 (1)
	Emu3.5	0.556 / 0.714 (5)	0.614 / 0.738 (5)	0.609 / 0.736 (5)	0.809 / 0.854 (5)
ColorSwap	Show-o	0.863 / 0.921 (5)	0.923 / 0.936 (1)	0.925 / 0.938 (0)	0.922 / 0.939 (5)
	Janus-Pro	0.936 / 0.958 (5)	0.945 / 0.964 (3)	0.960 / 0.969 (1)	0.956 / 0.967 (2)
	Emu3	0.866 / 0.917 (5)	0.913 / 0.949 (5)	0.915 / 0.942 (4)	0.919 / 0.948 (5)
	Chameleon	0.831 / 0.903 (5)	0.843 / 0.900 (5)	0.859 / 0.906 (5)	0.862 / 0.911 (5)
	Qwen × Show-o	0.935 / 0.951 (3)	0.971 / 0.973 (0)	0.977 / 0.978 (0)	0.978 / 0.979 (0)
SPEC	Show-o	0.578 / 0.621 (3)	0.728 / 0.731 (0)	0.749 / 0.748 (0)	0.762 / 0.757 (0)
	Janus-Pro	0.686 / 0.832 (5)	0.762 / 0.862 (5)	0.781 / 0.862 (5)	0.796 / 0.869 (5)
	Emu3	0.658 / 0.729 (5)	0.653 / 0.724 (5)	0.702 / 0.760 (5)	0.715 / 0.761 (3)
	Chameleon	0.498 / 0.554 (2)	0.497 / 0.551 (3)	0.492 / 0.549 (4)	0.514 / 0.547 (2)
ARO	Show-o	0.808 / 0.810 (0)	0.808 / 0.810 (0)	0.809 / 0.809 (0)	0.812 / 0.814 (0)
	Janus-Pro	0.802 / 0.819 (5)	0.802 / 0.819 (5)	0.814 / 0.824 (3)	0.816 / 0.829 (5)
	Emu3	0.797 / 0.798 (0)	0.797 / 0.798 (0)	0.810 / 0.809 (0)	0.815 / 0.814 (0)
	Chameleon	0.739 / 0.743 (1)	0.737 / 0.743 (3)	0.739 / 0.742 (0)	0.753 / 0.760 (0)
	BAGEL	0.856 / 0.856 (0)	0.856 / 0.856 (0)	0.865 / 0.864 (0)	0.864 / 0.865 (0)

Generation gains on SugarCrepe++ remain significant on three backbones under a caption-length control. In our SugarCrepe++ pairing the human-validated second positive is usually longer than the hard negative: choosing the longer caption, without the image or any model score, reaches 0.813 on the same test splits. Summed likelihood readouts encode length, so the learned understanding combination can exploit it. Main-text Table 4 adds the character-length difference (correct minus foil) as an explicit margin column to both the understanding-only combination and U+G, with the same combiner and criterion. Generation remains significant on Show-o, Chameleon and Emu3.5. This control is a retrospective reanalysis of the packaged margins.

B.2 Gating and fusion onto the selected ensemble

Adding a single G readout to the frozen selected ensemble already yields reliable gains: fusion passes on 10 of 27 pairs and gating on 6. The selected-ensemble comparison adds one auxiliary readout to the frozen selected ensemble. The selected ensemble (§3.2) is frozen once chosen. The U augmentation searches the twelve understanding configurations, the G augmentation the four generation configurations, and an independently selected G-only readout serves as a control. Each G addition is tested against the ensemble alone, both U additions and their calibration-selected system (Ens.+U), all preceding U-only controls, and G alone, under the criterion of §3.4. Earlier single-readout gates add G to the original-caption or best single U readout and retain their original criteria (Appendix B.5).

Margin-based gating. Let $m_b(x) = |b(x, c_1) - b(x, c_2)|$ be the absolute margin of an auxiliary readout b, U or G. The gated prediction follows b when $m_b(x) \geq \tau$ and the frozen selected ensemble otherwise, a reject-style rule that delegates rather than abstains (Chow, 1970; Geifman & El-Yaniv, 2017). Calibration selects a target coverage p^* from a fixed grid including zero delegation and converts it into a threshold τ on the calibration margins, frozen for testing.

Table 7: U+G significantly improves 10 of 27 pairs over the learned U-only combination (bold), by $+1.2$ to $+8.1$ points; 8 of the 10 are also significant in the independent implementation. Detailed scores for main-text Table 1: G alone, learned U-only, U+G, gain in percentage points, and significant splits. G alone is selected on calibration data. Five-split mean accuracies; bold denotes at least four significant splits, and ‡ marks a gain not significant in the independent implementation.

Benchmark	Model	G alone	U-only	U+G	Gain	Sig.
SugarCrepe	Show-o	0.795	0.882	**0.902**	**+2.0**	5/5
	Janus-Pro	0.755	0.939	0.943	+0.4	0/5
	Emu3	0.685	0.912	0.918	+0.6	2/5
	Chameleon	0.708	0.871	0.883	+1.2	3/5
	BAGEL	0.724	0.965	0.969	+0.3	1/5
	Emu3.5	0.771	0.747	**0.828**	**+8.1**	5/5
	Qwen × Show-o	0.795	0.972	0.973	+0.0	0/5
SugarCrepe++	Show-o	0.747	0.867	**0.892**	**+2.5**	5/5
	Janus-Pro	0.699	0.893	0.902	+0.9	3/5
	Emu3	0.699	0.872	**0.884**	**+1.2**	4/5
	Chameleon	0.657	0.819	0.831	+1.1	3/5
	BAGEL	0.629	0.920	0.925	+0.5	1/5
	Emu3.5	0.705	0.809	**0.854**	**+4.5**	5/5
ColorSwap	Show-o	0.864	0.922	**0.939**	**+1.7**‡	5/5
	Janus-Pro	0.904	0.956	0.967	+1.1	2/5
	Emu3	0.843	0.919	**0.948**	**+3.0**	5/5
	Chameleon	0.855	0.862	**0.911**	**+4.8**	5/5
	Qwen × Show-o	0.864	0.978	0.979	+0.1	0/5
SPEC	Show-o	0.577	0.762	0.757	−0.5	0/5
	Janus-Pro	0.800	0.796	**0.869**	**+7.3**	5/5
	Emu3	0.637	0.715	0.761	+4.6	3/5
	Chameleon	0.545	0.514	0.547	+3.3	2/5
ARO	Show-o	0.574	0.812	0.814	+0.2	0/5
	Janus-Pro	0.595	0.816	**0.829**	**+1.3**‡	5/5
	Emu3	0.533	0.815	0.814	−0.1	0/5
	Chameleon	0.522	0.753	0.760	+0.8	0/5
	BAGEL	0.588	0.864	0.865	+0.1	0/5

Score fusion. Fusion predicts with the sign of $f_w = w\,a/\sigma_a + (1-w)\,b/\sigma_b$, where a and b are the signed margins of the selected ensemble and of the auxiliary readout, σ their calibration standard deviations, and $w \in \{1, 0.95, \ldots, 0\}$ is chosen on calibration labels. Appendix A.4.1 gives the full formulas, the margin of each ensemble type, search grids, and tie-breaking rules.

The selected ensemble is frozen before adding either U or G. This controls for gains recoverable by combining understanding readouts alone; Tables 10 and 11 give all accuracies and paired comparisons.

Selected-ensemble results. On SugarCrepe, G fusion passes the joint criterion on Show-o, Chameleon, and Emu3.5; G gating passes only on Emu3.5. Fusion exceeds Ens.+U by 1.79 and 1.28 points on Show-o and Chameleon. Across all 27 pairs, fusion passes 10 and gating 6 (Table 11); Appendix B.3 gives the full comparisons and intervals.

These are alternative combination rules; the main method's learned combiner is stronger on both sides: its U-only baseline beats Ens.+U on 9 of the 27 pairs by up to 20 points, and U+G exceeds one-parameter fusion on 13 of 27 pairs and is never reliably worse (Appendix B.3.3). Against the learned U-only combination, U+G improves 10 pairs by $+1.2$ to $+8.1$ points (Table 1), 8 of them in both implementations. These counts are not pooled accuracies or independent replications.

Dataset differences. Fusion passes on four of six SugarCrepe++ backbones (2.75–12.05 points over Ens.+U), whereas gating fails on Emu3.5. On ColorSwap only Chameleon passes; on SPEC fusion passes for Janus-Pro and Emu3, and gating only for Janus-Pro. SPEC/Chameleon's $+5.00$-point fusion gain fails the G-alone control on every split. No ARO pair passes.

Scoring costs and single-readout calibration budgets are reported together in Appendix D; the learned U+G budgets are in main-text Table 5.

B.3 Main comparisons against understanding-only systems

Generation retains reliable gains on a subset of pairs in each of three comparisons against understanding-only systems: the best-single-readout G gate against stronger U systems, U/G additions to a frozen selected ensemble, and the learned U-only and U+G combinations. The first two use Appendix A.4.1; the third implements the main method. All reuse previously examined scores and splits. Rules were fixed before each new comparison was computed, but these are retrospective analyses rather than independent confirmations.

B.3.1 The original gate against stronger understanding-only systems

Table 8: Even the single-readout gate passes every understanding-only control and G-alone on SugarCrepe/Emu3.5 and SPEC/Janus-Pro (bold). Equal-label comparison of 27 model–benchmark pairs under the best-single-readout protocol; accuracies average five grouped test splits. Δ: single-readout gate minus the selected ensemble, in percentage points. $+/-$: jointly significant gains/losses versus the selected ensemble; Beats G: wins over G-alone. Pass: splits passing all five controls (single U, U–U gate, fixed-base U–U fusion, the selected ensemble, G-alone). Counts are out of five; bold requires at least four passing splits.

Model	Single U	Sel. ens.	G-alone	Single-readout gate	Δ	$+/-$	Beats G	Pass
SugarCrepe								
Show-o	0.852	0.870	0.795	0.877	+0.68	1/0	5	0
Janus-Pro	0.909	0.927	0.755	0.913	−1.34	0/5	5	0
Emu3	0.900	0.907	0.685	0.900	−0.68	0/1	5	0
Chameleon	0.782	0.837	0.708	0.794	−4.26	0/5	5	0
BAGEL	0.935	0.957	0.724	0.935	−2.13	0/5	5	0
Emu3.5	0.706	0.711	0.771	**0.802**	**+9.14**	5/0	5	**5**
Qwen × Show-o	0.942	0.962	0.795	0.945	−1.68	0/5	5	0
SugarCrepe++								
Show-o	0.760	0.777	0.747	0.798	+2.15	3/0	5	3
Janus-Pro	0.804	0.841	0.699	0.816	−2.46	0/5	5	0
Emu3	0.773	0.800	0.699	0.801	+0.02	0/0	5	0
Chameleon	0.695	0.706	0.657	0.725	+1.90	4/0	5	3
BAGEL	0.886	0.900	0.629	0.885	−1.51	0/5	5	0
Emu3.5	0.614	0.609	0.705	0.724	+11.49	5/0	2	2
ColorSwap								
Show-o	0.923	0.925	0.864	0.926	+0.11	0/0	5	0
Janus-Pro	0.945	0.960	0.904	0.957	−0.28	0/0	5	0
Emu3	0.913	0.915	0.843	0.923	+0.78	0/0	5	0
Chameleon	0.843	0.859	0.855	0.893	+3.37	4/0	5	3
Qwen × Show-o	0.971	0.977	0.864	0.971	−0.64	0/0	5	0
SPEC								
Show-o	0.728	0.749	0.577	0.725	−2.40	0/5	5	0
Janus-Pro	0.762	0.781	0.800	**0.829**	**+4.78**	5/0	4	**4**
Emu3	0.653	0.702	0.637	0.702	−0.00	0/0	5	0
Chameleon	0.497	0.492	0.545	0.543	+5.07	1/0	0	0
ARO								
Show-o	0.808	0.809	0.574	0.808	−0.10	0/0	5	0
Janus-Pro	0.802	0.814	0.595	0.802	−1.25	0/3	5	0
Emu3	0.797	0.810	0.533	0.797	−1.26	0/3	5	0
Chameleon	0.737	0.739	0.522	0.737	−0.19	0/0	5	0
BAGEL	0.856	0.865	0.588	0.856	−0.96	0/1	5	0

Even this single-readout gate passes every understanding-only control and G-alone on SugarCrepe/Emu3.5 and SPEC/Janus-Pro (Table 8). On SugarCrepe, the original gate is not reliably superior to U fusion on Show-o and is reliably worse on Chameleon. Emu3.5 remains a clear original-gate positive. Across the panel, SPEC/Janus-Pro also jointly exceeds the primary U controls and G-alone on four splits; SugarCrepe++/Emu3.5 exceeds U controls but G-alone on only 2/5. These are separate claims: a stronger G-only readout may explain an apparent gain over U without establishing a combination benefit. The complete negative comparisons in Table 8 are retained.

Table 9: The selected ensemble matches or exceeds the best single U readout on all but two pairs, making it a strong understanding-only reference. Understanding-only accuracy in the equal-label comparison (five-split means). U–U fix fuses with a fixed single-U base; U–U pair jointly selects both readouts and the weight. Mean/vote use twelve readouts. Calibration selects the selected ensemble from the six methods; all six serve as extension controls.

Model	Single U	U–U gate	U–U fix	U–U pair	Mean	Vote	Sel. ens.
SugarCrepe							
Show-o	0.852	0.863	0.871	0.870	0.865	0.864	0.870
Janus-Pro	0.909	0.914	0.926	0.928	0.928	0.927	0.927
Emu3	0.900	0.898	0.907	0.907	0.882	0.894	0.907
Chameleon	0.782	0.783	0.837	0.837	0.793	0.775	0.837
BAGEL	0.935	0.945	0.954	0.957	0.956	0.952	0.957
Emu3.5	0.706	0.706	0.711	0.711	0.686	0.697	0.711
Qwen × Show-o	0.942	0.958	0.962	0.962	0.962	0.961	0.962
SugarCrepe++							
Show-o	0.760	0.770	0.777	0.777	0.633	0.648	0.777
Janus-Pro	0.804	0.825	0.841	0.842	0.771	0.767	0.841
Emu3	0.773	0.784	0.800	0.800	0.670	0.684	0.800
Chameleon	0.695	0.702	0.706	0.706	0.528	0.583	0.706
BAGEL	0.886	0.887	0.900	0.900	0.809	0.801	0.900
Emu3.5	0.614	0.612	0.609	0.609	0.490	0.493	0.609
ColorSwap							
Show-o	0.923	0.925	0.924	0.925	0.926	0.917	0.925
Janus-Pro	0.945	0.955	0.961	0.960	0.956	0.948	0.960
Emu3	0.913	0.912	0.913	0.915	0.917	0.916	0.915
Chameleon	0.843	0.852	0.861	0.860	0.862	0.848	0.859
Qwen × Show-o	0.971	0.977	0.977	0.975	0.975	0.972	0.977
SPEC							
Show-o	0.728	0.744	0.753	0.752	0.702	0.696	0.749
Janus-Pro	0.762	0.764	0.779	0.782	0.781	0.753	0.781
Emu3	0.653	0.672	0.698	0.702	0.695	0.689	0.702
Chameleon	0.497	0.498	0.500	0.495	0.486	0.479	0.492
ARO							
Show-o	0.808	0.805	0.809	0.809	0.801	0.805	0.809
Janus-Pro	0.802	0.803	0.814	0.814	0.789	0.796	0.814
Emu3	0.797	0.802	0.805	0.810	0.803	0.799	0.810
Chameleon	0.737	0.738	0.742	0.741	0.719	0.736	0.739
BAGEL	0.856	0.858	0.865	0.865	0.855	0.862	0.865

B.3.2 Generation added after fixing the understanding system

With the understanding system frozen first, G fusion clears every U control and G-alone on 10 of 27 pairs (Table 11). On SugarCrepe/Show-o, the G gate beats the selected ensemble on 5/5 splits but all U controls plus G-alone on only 2/5; fusion clears the full criterion on 5/5. Chameleon has 0/5 for the gate versus 4/5 for fusion under this stronger criterion, while Emu3.5 supports both. Gate and fusion are therefore reported separately; calibration-selected G is a third system, not the better test result. The remaining passing pairs and all failures are in Table 11.

Table 12 reports two distinct paired differences: the original single-readout gate versus the selected ensemble, and G fusion versus U select. Machine-readable results include all configurations, correction/harm counts, delegation rates, tests, and intervals.

Negative outcomes remain. ARO's extended G gates choose zero delegation, but fusion has no such universal observed behavior. For example, the G fusion on ARO/Emu3 loses 1.08 points to the U fusion added to the selected ensemble on split 66, with a group interval $[-1.79, -0.38]$ and $p = 0.004$. SugarCrepe/Janus-Pro, SugarCrepe/Emu3, BAGEL's measured pairs, and both Qwen cross-model pairs fail the stronger reliability criterion. SPEC/Chameleon's large mean increment still does not reliably survive group uncertainty. Neither calibration nor a positive panel average guarantees improvement on every split or model–task pair.

Table 10: Adding G to the frozen selected ensemble outperforms adding U on every pair that meets the joint criterion (bold). Selected-ensemble accuracy (five-split means). All additions use the same frozen selected ensemble. U/G select chooses among the selected ensemble, gate, and fusion using calibration accuracy. Shading marks G additions; bold meets the joint criterion in Table 11. G-alone results are in Table 8.

Model	Sel. ens.	+U gate	+U fusion	U select	+G gate	+G fusion	G select
SugarCrepe							
Show-o	0.870	0.868	0.873	0.873	0.882	**0.891**	**0.891**
Janus-Pro	0.927	0.926	0.929	0.928	0.927	0.930	0.930
Emu3	0.907	0.907	0.907	0.907	0.907	0.912	0.912
Chameleon	0.837	0.842	0.836	0.842	0.844	**0.854**	**0.854**
BAGEL	0.957	0.957	0.959	0.959	0.957	0.957	0.957
Emu3.5	0.711	0.709	0.709	0.709	**0.804**	**0.801**	**0.802**
Qwen × Show-o	0.962	0.962	0.966	0.966	0.962	0.963	0.963
SugarCrepe++							
Show-o	0.777	0.777	0.778	0.778	**0.808**	**0.824**	**0.824**
Janus-Pro	0.841	0.841	0.840	0.841	0.845	0.856	0.856
Emu3	0.800	0.799	0.801	0.799	**0.816**	**0.831**	**0.831**
Chameleon	0.706	0.706	0.707	0.707	**0.728**	**0.734**	**0.734**
BAGEL	0.900	0.899	0.901	0.900	0.900	0.901	0.901
Emu3.5	0.609	0.610	0.609	0.608	0.723	**0.729**	**0.729**
ColorSwap							
Show-o	0.925	0.925	0.925	0.925	0.927	0.933	0.931
Janus-Pro	0.960	0.959	0.960	0.959	0.960	0.964	0.962
Emu3	0.915	0.914	0.916	0.915	0.922	0.944	0.944
Chameleon	0.859	0.860	0.859	0.860	**0.897**	**0.905**	**0.905**
Qwen × Show-o	0.977	0.977	0.977	0.977	0.977	0.977	0.977
SPEC							
Show-o	0.749	0.748	0.747	0.747	0.745	0.747	0.744
Janus-Pro	0.781	0.785	0.789	0.791	**0.830**	**0.858**	**0.858**
Emu3	0.702	0.699	0.707	0.707	0.739	**0.766**	**0.766**
Chameleon	0.492	0.491	0.493	0.491	0.539	0.541	0.541
ARO							
Show-o	0.809	0.808	0.809	0.808	0.809	0.807	0.807
Janus-Pro	0.814	0.814	0.813	0.812	0.814	0.822	0.822
Emu3	0.810	0.809	0.811	0.812	0.810	0.808	0.808
Chameleon	0.739	0.740	0.740	0.740	0.739	0.740	0.740
BAGEL	0.865	0.865	0.864	0.865	0.865	0.864	0.864

These results concern the tested scoring pools. Earlier delegated-item diagnostics and label-budget curves evaluate single-readout gates; U+G's label budgets are reported separately in Table 5.

B.3.3 The learned U-only combination as the strongest baseline

The learned U-only combination is the strongest understanding-only baseline, and generation still improves it. It beats ensemble+U on 9 of the 27 pairs by up to 20 points, and U+G improves it on 10 of 27.

Motivation and protocol. The controls above select readouts, average or vote over them, or fit a single fusion weight. To test whether unused U-side combination capacity explains the G increment, we add the learned U-only combination, a supervised linear combination of all twelve U margins, and U+G, the same combination over the twelve U and four G margins. Both are fit on exactly the calibration IDs of the 135 frozen splits, reuse the packaged margins and test items, and require no model inference. Because margins are stored in correct-minus-foil coordinates, the label is constant; we fit the equivalent symmetrised pairwise logistic model through the origin, stacking $[X; -X]$ against $[1; 0]$ with no intercept. Columns are divided by their calibration standard deviation and are not centered, matching the fusion convention of Appendix A.4.1; centering would break the odd symmetry of a margin. The L_2 strength is chosen from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$ by five-fold calibration-internal group cross-validation, ties preferring stronger regularization, and the model is refit on all calibration rows. Predictions use the strict-positive sign convention. No test label,

Table 11: G fusion passes every U control and G-alone on 10 of 27 pairs, and G gating on 6. Selected-ensemble gains and passing splits. Δ_{gate}: G gate minus the selected ensemble; Δ_{fusion}: G fusion minus U select, in mean percentage points. U: splits passing every U control; U, G: also passing G-alone. Passing requires both McNemar and group-interval tests. Sel. evaluates the calibration-selected G system. Counts are out of five; bold requires at least four.

Model	Δ_{gate}	Gate U	Gate U, G	Δ_{fusion}	Fusion U	Fusion U, G	Sel. U, G
SugarCrepe							
Show-o	+1.15	2	2	**+1.79**	5	**5**	**5**
Janus-Pro	+0.04	0	0	+0.20	0	0	0
Emu3	−0.04	0	0	+0.47	2	2	2
Chameleon	+0.74	0	0	**+1.28**	4	**4**	**4**
BAGEL	+0.00	0	0	−0.19	0	0	0
Emu3.5	**+9.32**	5	**5**	**+9.24**	5	**5**	**5**
Qwen × Show-o	−0.03	0	0	−0.28	0	0	0
SugarCrepe++							
Show-o	**+3.15**	5	**5**	**+4.60**	5	**5**	**5**
Janus-Pro	+0.41	0	0	+1.43	2	2	2
Emu3	**+1.58**	4	**4**	**+3.22**	5	**5**	**5**
Chameleon	**+2.18**	5	**5**	**+2.75**	5	**5**	**5**
BAGEL	+0.00	0	0	+0.14	0	0	0
Emu3.5	+11.41	5	3	**+12.05**	5	**4**	**4**
ColorSwap							
Show-o	+0.25	0	0	+0.75	0	0	0
Janus-Pro	+0.06	0	0	+0.56	0	0	0
Emu3	+0.75	0	0	+2.87	3	3	3
Chameleon	**+3.79**	4	**4**	**+4.51**	5	**5**	**5**
Qwen × Show-o	+0.00	0	0	+0.03	0	0	0
SPEC							
Show-o	−0.39	0	0	+0.02	0	0	0
Janus-Pro	**+4.92**	4	**4**	**+6.73**	5	**5**	**5**
Emu3	+3.69	3	3	**+5.96**	5	**5**	**5**
Chameleon	+4.74	1	0	+5.00	1	0	0
ARO							
Show-o	+0.00	0	0	−0.12	0	0	0
Janus-Pro	+0.00	0	0	+0.97	3	3	3
Emu3	+0.00	0	0	−0.34	0	0	0
Chameleon	+0.00	0	0	+0.05	0	0	0
BAGEL	+0.00	0	0	−0.08	0	0	0

full-set accuracy, or outcome-dependent grid change enters the fit. The main-panel replay uses Python 3.10.12, NumPy 1.24.4, SciPy 1.14.0, and scikit-learn 1.5.2, pinned separately from the prospective CREPE environment in the numerical supplement; version changes can alter grouped CV assignments.

Criterion. The joint rule of §3.4 requires a positive paired difference, exact two-sided McNemar $p < 0.05$, and a 95% whole-group bootstrap interval above zero (2,000 draws), on at least four of five splits. The four comparisons were fixed before calculation on already inspected scores.

Replication of this control. A second implementation, differing in the regularization grid and in selecting it on one calibration half rather than by five-fold grouped cross-validation, reproduces every accuracy in Table 13 to within 0.005 and the same count of 10 of 27 passing pairs, and 11 of 27 under an alternative calibration-scaling variant. It reassigns four borderline pairs, however: SugarCrepe/Chameleon and ColorSwap/Janus-Pro pass there and not here, ColorSwap/Show-o and ARO/Janus-Pro pass here and not there. Only the counts and the pairs passing in both implementations, chiefly SugarCrepe/Show-o and Emu3.5, should be read as stable; individual borderline pairs depend on how the combiner is regularized. Under the stricter requirement of also clearing every original U control and G-alone, the second implementation reports 9 of 27 passing pairs under both scaling variants, and finds that the one-parameter fusion clears the learned U-only combination on 4 of 27 rather than 5.

Table 12: Ensemble+G fusion gains with intervals above zero on 5/5 splits for Show-o and Emu3.5 and 4/5 for Chameleon, where the original gate loses on every split. SugarCrepe per-split gains and 95% intervals, in percentage points. Old: original single-readout gate minus the selected ensemble. New: ensemble+G fusion minus U select. Intervals bootstrap whole image groups; p: exact two-sided McNemar test.

Seed	Old Δ	Old 95% CI	Old p	New Δ	New 95% CI	New p
Show-o						
20261005	$+1.21$	$[+0.08, +2.35]$	0.042	$+2.08$	$[+1.19, +3.01]$	1.2×10^{-5}
75	$+1.03$	$[-0.18, +2.24]$	0.100	$+2.74$	$[+1.76, +3.76]$	6.0×10^{-8}
76	$+0.50$	$[-0.71, +1.67]$	0.441	$+1.47$	$[+0.22, +2.69]$	0.018
77	$+0.59$	$[-0.63, +1.74]$	0.365	$+1.14$	$[+0.08, +2.15]$	0.029
78	$+0.09$	$[-1.24, +1.34]$	0.945	$+1.50$	$[+0.38, +2.73]$	0.014
Chameleon						
20261005	-3.67	$[-5.22, -2.09]$	4.6×10^{-6}	$+2.12$	$[+1.02, +3.31]$	4.2×10^{-4}
75	-4.03	$[-5.66, -2.35]$	1.6×10^{-6}	$+1.20$	$[+0.08, +2.44]$	0.044
76	-4.40	$[-5.95, -2.84]$	1.9×10^{-8}	$+1.21$	$[+0.04, +2.43]$	0.037
77	-4.71	$[-6.47, -2.92]$	5.9×10^{-8}	$+0.08$	$[-1.10, +1.36]$	0.946
78	-4.51	$[-5.98, -3.10]$	8.2×10^{-10}	$+1.76$	$[+0.64, +2.92]$	0.005
Emu3.5						
20261005	$+10.00$	$[+7.92, +12.07]$	1.1×10^{-22}	$+9.71$	$[+7.90, +11.47]$	8.7×10^{-28}
75	$+9.22$	$[+7.06, +11.33]$	2.6×10^{-19}	$+9.31$	$[+7.32, +11.28]$	1.9×10^{-22}
76	$+9.13$	$[+7.03, +11.16]$	2.8×10^{-21}	$+9.42$	$[+7.27, +11.63]$	2.0×10^{-20}
77	$+9.00$	$[+6.93, +11.30]$	2.3×10^{-19}	$+8.79$	$[+6.70, +11.09]$	3.8×10^{-16}
78	$+8.37$	$[+6.17, +10.61]$	1.1×10^{-16}	$+8.97$	$[+6.69, +11.38]$	7.9×10^{-18}

Results. Across all 27 pairs (Table 13), learned U-only beats ensemble+U on 9 pairs, including all six SugarCrepe++ pairs and a 20.09-point gain on SugarCrepe++/Emu3.5. U+G improves this stronger baseline on 10 pairs spanning five datasets. Eight pass in both implementations, spanning four datasets; the ARO/Janus-Pro gain is not among them, so the five-dataset span itself should not be read as stable. One-parameter ensemble+G fusion beats learned U-only on only 5 pairs and falls below it on every SugarCrepe++ pair. U+G beats that fusion on 13 pairs and is never reliably worse. These comparisons motivate using the learned combiner on both sides; its label-budget curve is in §6.

Baseline accuracy and gains. Significant gains occur across a wide range of learned U-only accuracy, from 0.747 to 0.922. The 27 learned-combination outcomes do not separate monotonically by U accuracy: SugarCrepe pairs fail at 0.871 and pass at 0.882, SugarCrepe++ pairs fail at 0.819 and pass at 0.872, and ColorSwap gains occur at 0.922 and 0.919. The table below summarizes this relationship descriptively. The highest learned U-only accuracy with a significant increment is 0.922; 0.938 isolates the five highest-baseline pairs (SugarCrepe/Janus-Pro is 0.9389, rounded to 0.939). These are summaries of observed outcomes, not thresholds selected by a new test.

Learned U-only band	Pairs passing	Rate
learned U-only < 0.800	2/6	33%
$0.800 \le$ learned U-only < 0.900	6/12	50%
$0.900 \le$ learned U-only < 0.938	2/4	50%
all learned U-only < 0.938	10/22	**45%**
learned U-only accuracy ≥ 0.938	0/5	**0%**
all 27 pairs (base rate)	10/27	37%

None of the five pairs at or above 0.938 passes: SugarCrepe/Janus-Pro, /BAGEL, /Qwen×Show-o, and ColorSwap/Janus-Pro, /Qwen×Show-o. Below this band, 10/22 pass, with no monotone trend across the lower bands. The prospective CREPE null at 96.65% is consistent with this pattern (Appendix F). Given its observed 22 discordant pairs, the exact McNemar criterion required at least a $+0.60$-point increment; this conditional calculation is not a design-level detection limit. Five panel pairs and one prospective test cannot establish a predictive ceiling rule. Understanding headroom and informative G together describe the measured gains (§4.2).

Table 13: U+G improves the learned U-only combination on 10 of 27 pairs and beats one-parameter ensemble+G fusion on 13, while learned U-only beats ensemble+U on 9. Learned U-only combination and U+G across all 27 pairs, five-split means. Learned U-only: combination of the twelve U margins; U+G adds the four G margins. The four comparison columns give the mean difference in percentage points with the number of jointly positive splits in parentheses; bold marks four or more. All readouts use the same calibration labels, splits, and test items as Table 10, and no new model inference. SugarCrepe/BAGEL and SugarCrepe++/Show-o learned U-only differ by 0.001 from main-text Table 1 (rounding across the two implementations).

Model	Learned U-only	U+G	Learned U-only vs. ensemble+U	U+G vs. learned U-only	Ensemble+G fusion vs. learned U-only	U+G vs. ensemble+G fusion
SugarCrepe						
Show-o	0.882	**0.902**	+0.87 (2)	**+2.00 (5)**	+0.91 (2)	**+1.09 (4)**
Janus-Pro	0.939	0.943	**+1.04 (4)**	+0.40 (0)	−0.84 (0)	**+1.24 (5)**
Emu3	0.912	0.918	+0.45 (0)	+0.57 (2)	+0.02 (0)	+0.55 (3)
Chameleon	0.871	0.883	**+2.91 (5)**	+1.25 (3)	−1.63 (0)	**+2.88 (5)**
BAGEL	0.966	0.969	+0.67 (2)	+0.34 (1)	−0.85 (0)	**+1.19 (5)**
Emu3.5	0.747	**0.828**	**+3.84 (5)**	**+8.11 (5)**	**+5.40 (5)**	**+2.71 (5)**
Qwen × Show-o	0.972	0.973	+0.65 (3)	+0.03 (0)	−0.94 (0)	**+0.96 (4)**
SugarCrepe++						
Show-o	0.868	**0.892**	**+8.98 (5)**	**+2.46 (5)**	−4.38 (0)	**+6.84 (5)**
Janus-Pro	0.893	0.902	**+5.14 (5)**	+0.90 (3)	−3.71 (0)	**+4.60 (5)**
Emu3	0.872	**0.884**	**+7.33 (5)**	**+1.18 (4)**	−4.10 (0)	**+5.28 (5)**
Chameleon	0.819	0.831	**+11.25 (5)**	+1.14 (3)	−8.49 (0)	**+9.63 (5)**
BAGEL	0.920	0.925	**+2.05 (5)**	+0.45 (1)	−1.90 (0)	**+2.36 (5)**
Emu3.5	0.809	**0.854**	**+20.09 (5)**	**+4.46 (5)**	−8.04 (0)	**+12.50 (5)**
ColorSwap						
Show-o	0.922	**0.939**	−0.28 (0)	**+1.70 (5)**	+1.03 (1)	+0.67 (0)
Janus-Pro	0.956	0.967	−0.28 (0)	+1.14 (2)	+0.84 (2)	+0.31 (0)
Emu3	0.919	**0.948**	+0.33 (0)	**+2.95 (5)**	**+2.54 (4)**	+0.42 (1)
Chameleon	0.862	**0.911**	+0.28 (0)	**+4.85 (5)**	**+4.23 (5)**	+0.61 (1)
Qwen × Show-o	0.978	0.979	+0.11 (0)	+0.08 (0)	−0.08 (0)	+0.17 (0)
SPEC						
Show-o	0.762	0.757	+1.54 (1)	−0.54 (0)	−1.51 (0)	+0.97 (2)
Janus-Pro	0.796	**0.869**	+0.55 (0)	**+7.26 (5)**	**+6.18 (5)**	+1.08 (2)
Emu3	0.715	0.761	+0.78 (1)	+4.63 (3)	**+5.18 (5)**	−0.55 (0)
Chameleon	0.514	0.547	+2.31 (0)	+3.30 (2)	+2.69 (1)	+0.61 (0)
ARO						
Show-o	0.812	0.814	+0.35 (0)	+0.21 (0)	−0.47 (0)	+0.68 (1)
Janus-Pro	0.816	**0.829**	+0.42 (0)	**+1.30 (5)**	+0.55 (1)	+0.75 (1)
Emu3	0.815	0.814	+0.32 (1)	−0.06 (0)	−0.66 (0)	+0.61 (2)
Chameleon	0.753	0.760	+1.28 (3)	+0.77 (0)	−1.24 (0)	**+2.01 (4)**
BAGEL	0.864	0.865	−0.06 (0)	+0.06 (0)	−0.02 (0)	+0.08 (0)
Pairs passing	—	—	**9/27**	**10/27**	5/27	**13/27**

B.4 Where generation adds evidence: exploratory analyses

Gains concentrate where understanding leaves headroom and generation alone is informative, and corrections outnumber new errors on 25 of 27 pairs. **These analyses are post hoc and descriptive.** They reuse the per-item predictions of the learned U-only combination and U+G at level (iv), computed from the packaged margins on the same five splits, and the accuracy of the generation readout selected on calibration data alone. Accuracies aggregated from these predictions reproduce Table 1 exactly. No test was preregistered; the product index of Figure 2 was defined after the two conditions had been identified, and all accuracies are test-split means.

Two conditions jointly describe the gains. Across the 27 pairs, the gain of U+G correlates with understanding headroom (Spearman $\rho = 0.39$, $p = 0.04$) and with the accuracy of generation alone ($\rho = 0.35$, $p = 0.08$), and strongly with the product of headroom and above-chance generation accuracy ($\rho = 0.86$, $p = 7.6 \times 10^{-9}$). Because headroom and gain share the test accuracy of the

learned U-only combination, we recompute both conditions on calibration data alone: headroom as one minus its five-fold cross-validated calibration accuracy, and generation accuracy as that of the best G readout on the calibration set. The gains, still measured on test, give $\rho = 0.32$ ($p = 0.10$), 0.35 ($p = 0.07$), and 0.86 ($p = 6.8 \times 10^{-9}$), respectively.

Corrections outnumber new errors. Averaged across the 27 pairs and five splits, corrections and newly introduced errors account for 4.64% and 2.75% of all test items, respectively; corrections exceed new errors on 25 of 27 pairs (the exceptions are SPEC/Show-o and ARO/Emu3). On the 10 improved pairs the rates are 6.56% and 2.93%.

Error types. Gains are positive on at least half of the models for every foil type (main-text Table 3, which splits the gain by foil type). They are largest for replaced objects and attributes on both SugarCrepe variants and for spatial positions on SPEC; on both SugarCrepe variants, relation replacements gain less than object and attribute replacements, whereas on ARO the relation half gains more than the attribute half, whose understanding accuracy is near ceiling. Types differ in understanding accuracy and sample size, and no per-type significance test was run, so these differences are descriptive; in particular, swaps of attributes between objects do not gain more than replacements.

B.4.1 Earlier comparisons between gating and score fusion

The following comparisons test combination rules under their original single-readout protocols. They provide context for the ensemble experiment above and retain their original selection procedures and significance thresholds.

Original protocol. Scale-normalised weighted fusion was calibrated on the same labels as the gate. Under this comparison's preregistered three-of-five McNemar criterion, fusion never significantly beats gating (four ties, one gate win). On Chameleon, gating wins all five splits (0.751 versus 0.722, $p \leq 0.001$). Janus-Pro trends toward fusion on two significant splits, below the criterion. Fusion applies its fitted weight to every item; gating uses G only above its calibrated margin threshold.

Expanded search on eight earlier gaining pairs. This comparison searches over all bases and generation readouts on the eight pairs that gained under the original-caption protocol, the same pairs as the label-free analysis, using the final group-disjoint splits. Scale-normalised fusion fits on the same labels and readout pool as the gate: it jointly selects the base, generation readout, and mixing weight, whereas the gate first selects its base and then its readout and coverage. The comparison finds no reliable difference on 6 pairs; fusion exceeds the gate on 2 ($+1.4$ and $+1.7$ points, Show-o and Emu3 on SugarCrepe++); no pair favours the gate under the four-of-five McNemar rule. The original three-of-five rule selects the same pairs. The mean fitted generation weight is 0.55. These results describe the eight earlier gaining pairs. The subsequent selected-ensemble study covers all 27 pairs with a different fixed-base protocol and additional U controls (Appendix B.3); it does not relabel these historical comparisons.

B.5 Understanding-readout controls and robustness

Generation gains persist when the understanding readout is strengthened: after calibration selects both readouts, Show-o, Chameleon, and Emu3.5 retain reliable SugarCrepe gains. These experiments test how the generation increment changes when the understanding readout is strengthened. The best-single-readout summary is followed by its component controls, prompted-format audits, and prediction record. All gains in this subsection are relative to the stated single readout; Appendix B.3 supplies the comparisons against understanding-only combinations.

B.5.1 Original-caption and best-single-readout results

Table 14 and its discussion preserve the earlier single-readout protocol. Its original-caption columns use their original splits; only the best-single-readout columns share the image-grouped splits of the ensemble analyses in Appendix B.3.

Best-single-readout results. After calibration selects both readouts, Show-o, Chameleon, and Emu3.5 retain reliable gains; the other SugarCrepe backbones do not. No split loses significantly. The observed baseline accuracies separate these groups (≤ 0.852 versus ≥ 0.900), but this six-backbone pattern is descriptive. Zero delegation reproduces the base by construction, so the absence

Table 14: Generation gains on Show-o, Chameleon, and Emu3.5 survive the move from the original-caption to the best-single-readout base. Single-readout SugarCrepe results ($n = 4757$; five-split mean accuracy). Left: original-caption gate. Right: calibration selects from twelve U configurations and four G templates on image-grouped splits; Δ gives the splitwise gain range. The two protocols use different splits and are not paired across columns.

	Original-caption protocol		Best-single-readout protocol			
Pair	**base**	**+G**	**base**	**+G**	**Δ (pts)**	**sig.**
Show-o	0.718	**0.831**	0.852	**0.877**	$+1.88$ to $+2.96$	5/5
Chameleon	0.659	**0.751**	0.782	**0.794**	$+0.77$ to $+1.67$	4/5
Emu3.5	0.608	**0.727**	0.706	**0.802**	$+8.80$ to $+10.62$	5/5
Janus-Pro	0.811	**0.836**	0.909	0.913	$+0.25$ to $+0.56$	1/5
Emu3	0.762	**0.788**	0.900	0.900	-0.04 to $+0.25$	0/5
BAGEL	0.840	0.841	0.935	0.935	$+0.00$ to $+0.08$	0/5
Qwen3.8-27B × Show-o	—	—	0.942	0.945	$+0.08$ to $+0.50$	3/5

Table 15: Generation gating beats both U and the second opinion on all five splits for five of six backbones. Original-caption replication on six backbones ($n = 4757$). Accuracies are five-split means; gains include splitwise standard deviations. The splits column counts significant wins over both U and second opinion. The stronger-base column recalibrates the gate on the calibration-selected original-caption/PMI readout. †: dual criterion not met; dashes: untested. Chameleon uses Anole weights.

Backbone	**U base**	**+G gate**	**G alone**	**+second opinion**	**gain (pts)**	**splits**	**stronger base (pts)**
Show-o	0.718	**0.831**	0.798	0.734	$+11.2 \pm 0.6$	5/5	$+2.17$
Chameleon	0.659	**0.751**	0.714	0.672	$+9.1 \pm 0.9$	5/5	$+1.22$
Emu3	0.762	**0.788**	0.695	0.768	$+2.6 \pm 0.2$	5/5	$+2.59$
Janus-Pro	0.811	**0.836**	0.750	0.818	$+2.5 \pm 0.4$	5/5	$+0.88^{\dagger}$
BAGEL	0.840	0.841	0.685	0.847	$+0.05 \pm 0.2$	0/5	—
Emu3.5	0.608	**0.727**	0.706	0.614	$+11.8 \pm 0.7$	5/5	—

of significant loss is an observed loss count, not an equivalence test or a general safety guarantee; no reliable gain need not mean zero delegation.

The component controls below distinguish language-prior correction, caption templates, grouped evaluation, and prompted interfaces. They measure residual G value against each selected single U readout, without isolating its entire cause as additional visual information; the initial visual-dependence, order-bias, and oracle audits cover Show-o only. The stronger combination controls are in Appendix B.3.

B.5.2 Original-caption replication and understanding-side controls

Under the original-caption protocol, generation improves five of six backbones on all five splits. This replication uses the original-caption likelihood base and its original splits. Table 15 reports all six backbones; Table 16 gives Show-o's split-level results.

Original-caption gains. Five of six backbones beat both the U base and its alternate-normalisation gate on all five splits. BAGEL gains only 0.05 points with coverage 0.00–0.10. Emu3 improves despite weaker standalone G, consistent with useful evidence on the delegated subset. Emu3.5 has the largest G–U gap (0.706 versus 0.608). At coverage 0.70, its gate also beats G alone on all five splits ($p \leq 0.037$): of the 11.8-point gain over U, 9.7 points come from stronger standalone G and 2.1 from combining the pathways. Fair-coin and inverted-G replacements on the same delegated items change U accuracy by -7.5 and -26.9 points. The preregistered predictions and misses are retained in Appendix B.5.6.

Additional understanding controls. Alternate normalisation reuses the same forward pass. Its gated gains are 1.5, 0.7, 0.5, 0.7, 1.3, and 0.5 points on Show-o, Janus-Pro, Emu3, BAGEL,

Chameleon, and Emu3.5; G beats it significantly on all 25 replicating splits (half of the preregistered dual criterion). A true second inference instead uses a frozen nine-tenths centre crop on Show-o/Chameleon or a template rephrase on Janus-Pro/Emu3. Its full-set accuracy changes by -0.25 to $+1.41$ points, and its Show-o gate stays at the 0.718 base. G beats this control on all five splits for Show-o, Emu3, and Chameleon; Janus-Pro has no significant difference.

Show-o split-level results.

Table 16: G gating beats all three controls on every split. Show-o original-caption accuracy on each SugarCrepe split ($n = 4757$). Second opinion uses alternate U normalization; coverage is the calibrated delegation fraction.

Split	U base	+G gate	G alone	+second opinion	coverage
1	0.727	**0.829**	0.790	0.740	0.70
2	0.709	**0.827**	0.792	0.734	0.70
3	0.719	**0.836**	0.801	0.728	0.70
4	0.714	**0.825**	0.805	0.726	0.60
5	0.723	**0.836**	0.803	0.740	0.60
mean	0.718	**0.831**	0.798	0.734	—

Show-o improves from 0.718 to 0.831 on average (0.836 on its best splits). All five splits beat U ($p < 10^{-24}$), alternate-normalisation gating ($p < 10^{-17}$), and G alone ($p \leq 0.015$). Coverage is 0.60–0.70 on a grid including both zero and full delegation. The controls below measure how much of this gain remains after strengthening U.

B.5.3 Language-prior, template, and grouped-split controls

Language-prior correction leaves reliable generation gains on three backbones. Show-o's text-only U accuracy is 0.561; correct versus shuffled images yield 0.718 versus 0.580. For G, the matched figures are 0.796 versus 0.568. Both scores therefore depend on image content. Subtracting each backbone's text-only caption score (PMI) changes U accuracy by $+8.79$, $+2.86$, $+12.09$, and -2.02 points on Show-o, Janus-Pro, Chameleon, and Emu3. Selecting the stronger of U and PMI on calibration data and recalibrating the gate leaves gains of $+2.17$, $+2.59$, and $+1.22$ points on Show-o, Emu3, and Chameleon, respectively, each beating both the selected base and strongest second-opinion gate on 5/5 splits. Show-o reaches 0.832; Janus-Pro is a statistical tie. On SugarCrepe++, PMI instead hurts Emu3 and Chameleon by 9.31 and 10.76 points while their G gains persist. This correction recovers part, but not all, of the original-caption gain.

Caption-template repair. Replacing each scored caption with ``A photo of {c}.'' improves U by 9.5–13.4 points across all six backbones at the same inference cost. Two preregistered controls indicate a substantial text-side contribution. Show-o's no-image accuracy rises from 0.561 to 0.752; across four backbones, the no-image lift is 0.87–1.66 times the with-image lift and exceeds it on three. After subtracting each backbone's text-only score, the template's advantage is only -2.23 to $+2.14$ points. This is consistent with prior re-anchoring (Holtzman et al., 2021), without making template scoring equivalent to PMI or excluding image–text interactions.

On this stronger single-template base, G adds 2.45, 3.14, and 5.13 points on Show-o, Chameleon, and Emu3.5 (5/5 splits each; Table 17). The other three backbones show no reliable gain or significant split-level loss; the worst mean change is -0.03 points. U exceeds G by at most 5.4 points in the gaining group and at least 15.7 in the other group, a post hoc description of these six outcomes.

With this single-template base, calibration selects zero delegation on none of five splits for Janus-Pro, three for Emu3, and all five for BAGEL; the remaining splits delegate without significant changes.

Selecting both readouts. The best-single-readout comparison selects from the twelve U configurations and four G templates of Appendix A.4.1, using common image-grouped splits. Table 14 reports the resulting bases and gains. Show-o, Chameleon, and Emu3.5 pass on 5/5, 4/5, and 5/5 splits. Chameleon selects PMI; its gains are 0.77–1.67 points, compared with 1.88–2.96 on Show-o and 8.80–10.62 on Emu3.5. Emu3.5's bootstrap lower bounds all exceed 6.81 points.

Table 17: On the stronger single-template base, generation still adds reliable gains on Show-o, Chameleon, and Emu3.5 (5/5 splits each), with no significant split-level loss on the other three. Gating after the single caption-template repair ($n = 4757$). Base accuracies use the full item set; gains are five-split means. "Closed" means zero delegation on all five splits; "no reliable gain" allows nonzero delegation. Table 14 instead selects across the full repair suite.

Backbone	original-caption base	single-template base	+G gate on it (pts)	splits
Show-o	0.718	0.852	$\mathbf{+2.45}$	5/5
Chameleon	0.659	0.757	$\mathbf{+3.14}$	5/5
Emu3.5	0.602	0.702	$\mathbf{+5.13}$	5/5
Janus-Pro	0.812	0.908	$+0.16$ (no reliable gain)	0/5
Emu3	0.764	0.884	-0.03 (no reliable gain)	0/5
BAGEL	0.839	0.936	$+0.00$ (closed)	0/5

Generation-template selection matters: conditioning improves Emu3.5's G accuracy from 0.707 to 0.770 but reduces Show-o's from 0.796 to 0.754. Calibration selects the conditioned and bare scores, respectively, on every split. Emu3.5's U readout changes little with PMI (0.602 to 0.609), and its strongest single-template U score is 0.702. Figure 4 checks the image dependence of representative conditioned G readouts.

No SugarCrepe split loses significantly under this protocol. Emu3 and BAGEL select zero delegation on two and three splits; Janus-Pro never selects zero and has one significant gain, below the reliability criterion. On Emu3.5's fixed delegated items, fair-coin and inverted-G controls change accuracy by -15.08 to -12.98 and -40.42 to -35.58 points. Table 29 gives all replacement controls.

Grouped replay of the earlier original-or-PMI protocol. The 4,757 pairs use 1,542 images, which could appear on both sides of the earlier item-level splits. Five fresh preregistered image-grouped splits retain gains of 2.22, 2.61, and 1.19 points on Show-o, Emu3, and Chameleon, passing both McNemar and whole-image bootstrap criteria on 5/5, 5/5, and 4/5 splits. Comparisons with the strongest second opinion pass on the same counts. Original-caption gains also remain (10.65, 2.61, and 8.30 points; all 5/5). Thus these gains persist without shared images between calibration and test.

B.5.4 Prompted understanding-readout audit

A preregistered audit rescored every item under alternative understanding readout formats on Show-o, Emu3, and Janus-Pro; a preregistered follow-up extended the same five-format family to Chameleon and BAGEL's understanding-side language model. A subsequent extension adds Emu3.5-Image (Appendix B.5.5). Table 18 restricts all six backbones to the five length-matched slices used throughout the paper ($n = 4757$). These full-set accuracies are descriptive; the Emu3.5 extension selects its understanding base only on calibration labels.

Table 18: Per-caption yes/no is descriptively the strongest prompted format on five of six backbones. Prompted understanding accuracy on 4,757 SugarCrepe pairs. Bold shading marks each model's highest descriptive accuracy, without a significance claim. Emu3.5 uses the main experiment's Emu3.5-Image checkpoint.

Readout	Show-o	Janus-Pro	Emu3	Chameleon	BAGEL	Emu3.5
A/B	0.481	0.791	0.640	0.586	**0.973**	0.500
A/B, options swapped	0.490	0.802	0.662	0.593	0.969	0.498
A/B, template variant 2	0.482	0.809	0.694	0.623	0.961	0.500
A/B, template variant 3	0.500	0.822	0.653	0.654	0.955	0.500
Per-caption yes/no	**0.521**	**0.885**	**0.882**	**0.706**	0.967	**0.582**

A separate preregistered audit used the initial full item set and the strongest then-audited readout per backbone (forced choice for BAGEL). None of its frozen rules beat the best single readout on any of four backbones after multiplicity correction. Oracle gaps were 0.014–0.125, but the upper endpoint concerns U adding to stronger G on Show-o (G 0.801 versus U 0.658), and BAGEL's lower endpoint matches pairing U with an independent fair coin, $U + (1 - U)/2$, on unrounded held-out

counts. Oracle coverage alone therefore establishes neither useful routing nor the direction of the contribution.

A preregistered follow-up instead selected the strongest prompted U format on each calibration half and recalibrated the gate. Chameleon improves from 0.706 to 0.750 ($+4.31$ points, 5/5 splits, coverage 0.56). Janus-Pro, Emu3, and BAGEL show no reliable gain (worst mean -0.09 points), selecting zero delegation on two, four, and five splits. Emu3's single delegated split loses significantly. Chameleon's prompted U is 0.3 points below G, while the other three exceed G by at least 13.5 points. Show-o was excluded by the preregistered near-chance rule. These outcomes show a gain beyond prompted U on one backbone; their accuracy gaps do not establish a necessary condition for gains.

B.5.5 EMU3.5 PROMPTED-FORMAT EXTENSION

Protocol and timing. We complete the same fixed five-format family on Emu3.5-Image using all 4757 pairs and 1542 images. The completed yes/no scores and likelihood-family results were already known; the four additional A/B formats and analysis rules were frozen before their scores were collected. This is a targeted extension on an already studied dataset, not independent confirmation, and does not alter the 27-pair panel. We reuse its five image-grouped calibration/test splits, twelve likelihood readouts, four G templates, and coverage grid including zero delegation. For each split, calibration selects the understanding base, then the G template and coverage; the resulting absolute margin threshold is fixed before testing. We compare the prompted-only pool and the expanded pool of seventeen understanding readouts. All formats are retained, including poorly performing ones. Full-set accuracies in Table 18 are not used for selection. Table 19 reports the held-out comparisons.

Table 19: Generation adds $+9.56$ points to Emu3.5 on 5/5 splits even when prompted readouts join the understanding pool. Emu3.5-Image prompted audit: five-split mean accuracy with calibration-selected U bases. Joint counts positive gains with exact two-sided McNemar $p < 0.05$ and whole-image bootstrap 95% intervals above zero (2,000 draws). Splits are image-grouped but overlap across repetitions.

Understanding candidate pool	U	+G	Δ (pts)	Joint
Twelve likelihood readouts	0.706	**0.802**	**+9.56**	**5/5**
Five prompted readouts	0.578	**0.774**	**+19.54**	**5/5**
Twelve likelihood + five prompted	0.706	**0.802**	**+9.56**	**5/5**

Results and boundary. All five splits in the expanded pool still select the likelihood base (the "A photo of {c}." template, mean aggregation). Its gate is unchanged, with gains of $+8.80$ to $+10.62$ points and positive whole-image intervals on every split (smallest lower bound $+6.81$ points). Separately selecting G alone's template on the same calibration labels yields mean accuracy 0.771; the expanded gate exceeds it by $+3.07$ points, jointly significant on 5/5 splits. However, the four A/B accuracies are only 49.80–50.01%, and standard and swapped A/B select option A on 99.35% and 99.29% of items. Among 4,752 pairs without a tie in either order, the two orders select the same semantic caption on only 0.95%. These position-biased readouts do not strengthen the likelihood baseline. The audit preserves the measured G increment within this fixed family; it does not establish superiority over better prompts or the model's strongest possible understanding interface. The prompted-only pool selects yes/no on every split.

Prompt templates, inference checks, and computational accounting for this extension are specified in Appendix A.6.

B.5.6 PREREGISTERED PREDICTIONS FOR THE BACKBONE REPLICATIONS

The original-caption gate used predictions fixed before each new gated evaluation. The records include the following outcomes:

- **Janus-Pro and Emu3 round.** Before either gated outcome existed, the prediction that Show-o's gain would be at least twice theirs held (11.2 versus 2.5/2.6 points). Janus-Pro passed as predicted (5/5, $p < 10^{-2}$). The predicted Emu3 null failed: it passed 5/5 ($p < 10^{-4}$), despite G trailing U by seven points, with coverage 0.20–0.30.

- **BAGEL and Chameleon round.** Before either was scored, BAGEL's null was predicted correctly (0/5, $+0.05$ points, coverage 0.00–0.10). Chameleon's predicted null failed: G scored 0.714 versus U 0.659, and coverage 0.60 yielded $+9.1$ points ($p < 10^{-16}$ on every split). The predicted inverse ranking of gain and U accuracy, with Show-o maximal, held in direction ($\rho = -0.90$, exact two-sided permutation $p = 0.083$, $n = 5$; with five points this is consistency with the preregistered prediction, not an independent significance claim).
- **Emu3.5 round.** With only a 75-item rolling pilot mean known, two predictions were frozen and both missed. Its U score was lowest (0.608), rather than above the older backbones; its $+11.8$-point gain slightly exceeded Show-o's $+11.2$, rather than being substantially smaller.
- **Best-single-readout follow-up.** After the earlier repair results were known, the predicted gaining backbones (Show-o, Chameleon, Emu3.5), null group (Janus-Pro, Emu3, BAGEL), and selection of a non-bare G template all held. Emu3.5 exceeded the predicted one-to-five-point gain range. No direction was predicted for the cross-model pair.

The original-caption evaluations cover six backbones on SugarCrepe and five on SugarCrepe++, ARO, and Winoground; Emu3.5 joins SugarCrepe++ through the best-single-readout comparisons. ColorSwap has a separate four-backbone validation. Initial order-bias, visual-dependence, and oracle controls cover Show-o only. Of the twelve initial backbone and external-set predictions, three missed (Emu3, Chameleon, and ARO replication), and one ARO rank prediction was unevaluable because all five gains were zero. The two later Emu3.5 misses are additional. Prediction timing and outcomes for each external set are reported below. Replays on shared image-grouped splits are retrospective, not new independent tests of the original ordering.

B.5.7 Scope of the readout comparisons

The audit covers a fixed five-format family, not every possible prompted interface. Prompting exceeds the original-caption gated system on Emu3, Janus-Pro, BAGEL, and Qwen3.8-27B; all audited prompts remain below it on Show-o and Chameleon. Recalibration on prompted U yields a reliable gain only on Chameleon among the eligible backbones, with one significant Emu3 split-level loss. Emu3.5's additional A/B formats remain near chance and position-biased, leaving its expanded-pool result unchanged.

The original-or-PMI Janus-Pro comparisons against the stronger base and second inference have only 2/5 and 1/5 significant splits. Subsequent single-readout repairs change which backbones benefit; their observed accuracy cutoffs are descriptive. BAGEL's null also combines strong U with a weak flow-matching G proxy, which this design cannot disentangle. The original-caption ordering does not predict best-single-readout gains in general, and the realized gains establish complementarity relative to the measured bases without attributing the full increment to additional visual information. All these conclusions are relative to the tested readouts; stronger combination controls are in Appendix B.3.

B.6 Cross-dataset generalization and benchmark extensions

Generation gains extend beyond SugarCrepe to rewritten captions (SugarCrepe++) and synthetic imagery (ColorSwap and SPEC). This subsection collects rewritten-caption tests, changes of image domain, the full SugarCrepe benchmark, and the descriptive and screened boundary sets. The summary first reports best-single-readout outcomes; the corresponding selected-ensemble results are in Appendix B.3. Earlier protocols and their preregistered predictions remain attached to each evaluation.

B.6.1 Single-readout outcome summary and interpretation

The summary below concerns the best-single-readout gate; the selected-ensemble results use the separate criterion in Appendix B.3. Gain ranges in Table 20 are the minimum and maximum test-split accuracy changes, in percentage points. Sig. counts positive McNemar tests; the dagger identifies the disagreement with the joint rule.

Best-single-readout summary. Table 20 separates the datasets and protocols. SugarCrepe++ tests rewritten captions on shared images; Emu3.5 has the largest gain and BAGEL no reliable gain.

Table 20: Generation gains hold on five of six SugarCrepe++ backbones and extend to ColorSwap (Chameleon) and SPEC (Janus-Pro, Emu3); ARO and Winoground show no reliable gain. Best-single-readout transfer results on shared group-disjoint splits. The first four blocks use the protocol of Table 14; backbone coverage differs by dataset. Winoground reports the range of backbone mean gains under its separate descriptive protocol.

Backbone	Gain / outcome	Sig.
SugarCrepe++ ($n = 4757$)		
Show-o	$+3.35$ to $+4.33$	5/5
Emu3	$+1.59$ to $+3.78$	5/5
Chameleon	$+2.87$ to $+3.29$	5/5
Emu3.5	$+9.92$ to $+12.01$	5/5
Janus-Pro	$+1.04$ to $+1.47$	5/5
BAGEL	No reliable gain; zero delegation on 4/5 splits	0/5
ColorSwap ($n = 1436$; 718 groups not used in earlier probes)		
Chameleon	$+3.90$ to $+5.99$	5/5
Show-o, Emu3, Janus-Pro	No reliable gain or significant loss	—
SPEC, synthesised ($n = 3220$)		
Janus-Pro	$+4.64$ to $+8.18$	5/5
Emu3	$+2.01$ to $+6.79$	4/5
Chameleon[†]	$+2.66$ to $+5.77$; joint rule not met	4/5
Show-o	No reliable gain; one split loses -1.09 significantly	0/5
ARO ($n = 5000$)		
All five backbones	No reliable gain; zero delegation on 25/25 splits	—
Winoground (descriptive; original protocol)		
All five backbones	-1.65 to $+1.80$; inconsistent	—

[†]SPEC/Chameleon selects $p^* = 1$ and delegates every test item on 4/5 splits, yielding G alone on those splits. Only 1/5 whole-group intervals exclude zero; the noise verdict is mixed (Table 29). It is excluded from the reliable-gain counts; the observed gain and McNemar count are retained.

ColorSwap tests AI-generated imagery and retains a reliable gain only on Chameleon. ARO's G readouts remain weak (at best 0.601), and every calibration selects zero delegation. On SPEC, Janus-Pro and Emu3 gain reliably; Chameleon fails the whole-group uncertainty criterion despite its positive McNemar count, and Show-o has one significant loss. Winoground remains descriptive, with only one backbone improving on average.

On the four backbones common to all five panel datasets, reliable gains number 2, 4, 1, 2, and 0 for SugarCrepe, SugarCrepe++, ColorSwap, SPEC, and ARO. The dataset dependence thus persists with matched model coverage. The following sections give the protocols, prediction outcomes, and limits.

B.6.2 Earlier protocols and screening record

This section preserves the original and best-single-readout evaluations; the selected-ensemble results in Appendix B.3 use different baselines and criteria.

Caption robustness: SugarCrepe++. Its second positive captions are LLM-generated and human-validated; images and most hard negatives are shared with SugarCrepe (82% verbatim). All three original-protocol predictions, registered before scoring, held: the four previously gaining backbones replicated (Show-o $+7.10$, Emu3 $+2.85$, Chameleon $+2.70$, Janus-Pro $+1.12$ points; 5/5 splits except Janus-Pro at 4/5), BAGEL was null (-0.13, coverage 0.02), and gain remained inversely ranked with U accuracy ($\rho = -0.70$). Rescoring the original captions checked for pipeline drift.

The best-single-readout round shares SugarCrepe's image assignments and selects both readouts on calibration data. Gains are 3.35–4.33, 1.59–3.78, 2.87–3.29, and 1.04–1.47 points on Show-o, Emu3, Chameleon, and Janus-Pro, all significant on 5/5 splits. BAGEL has no reliable gain over its 0.886 base and selects zero delegation on 4/5 splits. Emu3.5's selected U stays at 0.607–0.627,

Table 21: Under the earlier protocol, generation gains replicate on four of five SugarCrepe++ backbones. Earlier single-readout transfer results on two benchmarks (five-split mean accuracy). Spl. counts significant wins over both U and second opinion; bold requires at least 4/5. ARO selects zero delegation on all 25 backbone–split combinations. SugarCrepe results are in Table 15.

	SugarCrepe++ $n = 4757$			ARO $n = 5000$		
Backbone	**base**	**+gate**	**spl.**	**base**	**+gate**	**cov.**
Show-o	0.728	**0.799**	5/5	0.810	0.810	0.00
Chameleon	0.651	**0.678**	5/5	0.740	0.740	0.00
Emu3	0.776	**0.804**	5/5	0.790	0.790	0.00
Janus-Pro	0.780	**0.791**	4/5	0.804	0.804	0.00
BAGEL	0.781	0.780	0/5	0.849	0.849	0.00

below all its selected G scores, and gains 9.92–12.01 points on 5/5 splits. Both Chameleon and Emu3.5 select conditioned G on every split.

Image-domain transfer: ColorSwap. A preregistered screen compared four non-COCO candidates: What's Up (Kamath et al., 2023), SPEC (Peng et al., 2024), VALSE actions (Parcalabescu et al., 2022), and ColorSwap (Burapacheep et al., 2024). ColorSwap had promising G scores but exceeded the screen's U-accuracy cap. The screen was left unamended; a separate preregistered validation used 718 previously untouched groups (1,436 pairs), excluding all 282 groups from earlier probes. The other candidates were subsequently evaluated in full (Appendix B.6.5).

Under the original-or-PMI protocol with group-held-out splits, Chameleon improves from 0.837 to 0.888 ($+5.13$ points) and Show-o from 0.896 to 0.920 ($+2.42$; PMI had raised its base from 0.863). Both beat their bases on every split with positive group intervals. Against the second opinion, Chameleon passes 5/5 splits and Show-o only 3/5, missing the composite prediction that both would pass at least four. A preregistered extension adds Emu3 (0.877 to 0.911, $+3.34$ points; 5/5 versus both controls and positive intervals) and Janus-Pro (base 0.942, mean gain $+0.08$, no negative split).

The best-single-readout replay corrects the earlier item-level splitting: each ColorSwap group, including both related images and caption choices, stays wholly on one side. Each half has 359 groups and 718 pairs; all five model combinations share assignments. This retrospective replay keeps the five seeds, scores, candidate pools, and selection rules. U repair lifts Show-o from 0.863 to 0.927. Pairs with selected U at least 0.898 show no reliable gain or significant loss; Chameleon, whose base remains at most 0.851, gains 3.90–5.99 points with positive intervals on all five splits. It selects conditioned G throughout (full-set G 0.866 to 0.871).

ARO and Winoground. The preregistered prediction that the four replicating backbones would also gain on ARO failed. Original G accuracy is 0.519–0.575 versus U 0.740–0.849; all 25 backbone–split calibrations select zero coverage. The BAGEL null holds, while the gain-rank prediction is unevaluable because every gain is zero. Readout selection does not change this outcome: G reaches at most 0.601, and all 25 gates retain their selected U base. Original-caption U is the best full-set variant on four backbones, and a caption template on Chameleon. ARO's abstention therefore does not guarantee useful or harmless delegation on a new dataset.

Winoground does not show the same abstention: coverage is 0.02–0.46 and mean gains range from -1.65 to $+1.80$ points. Only one backbone improves on average, missing the predicted three of four previously gaining backbones. Its descriptive results are in Appendix B.6.4; neither useful nor harmless delegation is established on this dataset.

B.6.3 The full seven-slice benchmark

The main panel excludes SugarCrepe's add-object and add-attribute slices, whose negatives append a phrase and are systematically longer. A preregistered follow-up evaluated all six backbones on both the full seven slices and the excluded pair (Table 22), using the original-caption gate. Five backbones improve on the full set on 5/5 splits, from 3.4 points on Janus-Pro to 20.5 on Show-o; BAGEL shows a -0.1-point mean change with near-zero coverage. This does not extend the best-single-readout or learned-combination results to all seven slices.

On the excluded slices, mean caption likelihood falls to or below chance (Show-o 0.402, Chameleon 0.463), while summed likelihood approaches an artificial ceiling also attainable by caption length alone. G scores the same image-token positions for both captions, avoiding this direct scored-caption-length confound, though conditioning can still introduce length bias. Five gates select coverage 0.68–1.00 and reach 0.656–0.809; BAGEL selects zero. These accuracies are reported separately because the caption-length artifact dominates the U comparisons.

Table 22: Generation improves five of six backbones on the full seven-slice set, each on 5/5 splits. Gating on the full seven-slice SugarCrepe set ($n = 7511$) and excluded slices ($n = 2754$). These differ from the 4,757-pair panel. Spl.: significant wins over U; cov.: mean calibrated coverage.

	Full seven slices				Excluded pair alone			
Model	**base**	**gate**	**Δ**	**spl.**	**base**	**gate**	**Δ**	**cov.**
Show-o	0.600	**0.805**	+20.5	5/5	0.402	0.796	+39.4	1.00
Janus-Pro	0.735	**0.770**	+3.4	5/5	0.608	0.679	+7.1	0.68
Emu3	0.702	**0.745**	+4.3	5/5	0.609	0.681	+7.2	0.70
BAGEL	0.793	0.792	−0.1	0/5	0.721	0.721	+0.0	0.00
Chameleon	0.584	**0.704**	+11.9	5/5	0.463	0.656	+19.3	0.96
Emu3.5	0.594	**0.745**	+15.1	5/5	0.598	0.809	+21.2	0.96

B.6.4 WINOGROUND: DESCRIPTIVE EVALUATION

Winoground's 400 groups yield 800 caption-choice items (the text-score direction: given the image, choose the correct caption); a 400-item test half has limited power for small paired effects, so this set was declared *descriptive* before any score existed—the numbers below carry no significance tests and make no claims, by design.

Table 23: Descriptive Winoground results, averaged over five splits; no significance tests. Gains are percentage points over U. Dir+ counts positive-gain splits; coverage is the mean delegation fraction. These results establish neither reliable gains nor harmlessness.

Backbone	**U**	**G alone**	**gate gain (pts)**	**dir+**	**coverage**
Show-o	0.583	0.566	−1.65	1/5	0.46
Chameleon	0.557	0.562	+1.80	5/5	0.42
Emu3	0.601	0.566	−0.50	0/5	0.26
Janus-Pro	0.665	0.599	−0.05	1/5	0.16
BAGEL	0.719	0.601	−0.20	0/5	0.02

Chameleon is the only backbone with both G at least as accurate as U and a positive mean gate gain (positive on 5/5 splits). This post hoc pattern on 400-item test halves does not validate a rule for predicting gains.

B.6.5 SCREENED NON-COCO CANDIDATES AND THEIR PREDICTIONS

Before pursuing ColorSwap we screened four non-COCO candidates on 300 pairs each and pursued only the one whose generation readout looked usable. A later preregistration committed a prediction for each of the remaining three and then ran the full gated protocol on all of them: null on What's Up (both pathways near chance) and null on VALSE actions (understanding far above generation), with no prediction recorded for SPEC, whose two pathways sit at parity but both weakly. Splits are group-disjoint (every pair of an image on one side), four fresh seeds, the gate and its calibration unchanged from the original-caption protocol in Appendix A.4.2. Each dataset retains its own fixed item set and separate comparison block in Table 24.

Results and limits. Both predicted nulls hold. SPEC has a signed U-minus-G gap of -4.9 to $+1.4$ points and gains, whereas VALSE actions has a gap of $+9.7$ to $+22.8$ and no reliable gain. What's Up shows why parity alone is insufficient: two backbones are near parity, the four-model gap range is -2.8 to $+9.1$, and G is near chance throughout (at most 0.553). SPEC was explicitly left without a prediction. Both positive non-COCO domains, SPEC and ColorSwap, use synthetic imagery; the two photographic candidates are null.

Table 24: Generation gains on three of four SPEC backbones (4/4 splits each), and both predicted nulls hold on What's Up and VALSE actions. Earlier non-COCO screening results, averaged over four splits. Δ: gate minus U, in percentage points. Sig. splits: significant wins over U out of four. The five-split reliability criterion does not apply.

Model	U	G	Δ (pts)	Sig. splits
What's Up ($n = 1236$)				
Show-o	0.527	0.553	$+1.78$	1/4
Janus-Pro	0.580	0.538	$+0.57$	0/4
Emu3	0.639	0.549	$+0.00$	0/4
Chameleon	0.498	0.526	$+1.46$	1/4
VALSE actions ($n = 1821$)				
Show-o	0.778	0.633	$+0.99$	1/4
Janus-Pro	0.801	0.640	$+0.44$	0/4
Emu3	0.705	0.608	-0.19	0/4
Chameleon	0.822	0.594	$+0.00$	0/4
SPEC, synthesised ($n = 3220$)				
Show-o	0.588	0.589	$+2.40$	3/4
Janus-Pro	0.642	0.659	$\mathbf{+5.98}$	4/4
Emu3	0.649	0.635	$\mathbf{+3.66}$	4/4
Chameleon	0.511	0.560	$\mathbf{+5.50}$	4/4

The best-single-readout comparison on the frozen SPEC sample selects from twelve U variants and four G templates on grouped splits. Janus-Pro gains 4.64–8.18 points on 5/5 splits, with templates raising G from 0.659 to 0.802 (0.503 after image shuffling). Emu3 gains 2.01–6.79 on 4/5 splits. Chameleon gains 2.66–5.77 with 4/5 positive McNemar tests but only 1/5 positive group intervals. It delegates every item on 4/5 splits, the only full-delegation endpoint in the 27-pair panel. Its U variants reach at most 0.514, and its coin-replacement control is mixed, although G exceeds the coin expectation on every split. This case shows G carrying signal against near-chance U rather than complementarity between two effective readouts. Show-o's G is also weak (at best 0.589); one delegated slice loses 1.09 points significantly ($p = 0.005$), the only significant loss in the best-single-readout panel, a reminder that a calibrated threshold is an empirical control, not a guarantee.

B.6.6 Transfer scope and prediction limits

SugarCrepe++ establishes rewritten-caption robustness on shared images and mostly shared negatives. ColorSwap and SPEC add evidence on synthetic image domains; the photographic non-COCO candidates are null. SPEC had no registered prediction, and ColorSwap exceeded the original screening cap and missed its composite validation criterion on Show-o's second-opinion comparison. Its reliable gains narrow from three original-or-PMI backbones to one after best-single-readout repair. ARO's zero delegation is specific to that evaluation; Winoground shows that calibration can still delegate without a reliable gain. Positive transfer covers rewritten captions, a robustness transfer rather than an independent domain, and synthetic imagery, not broad image-domain generalization. These results support domain-limited transfer, not a general rule for predicting gains on new datasets.

B.7 Cross-model generation evidence

B.7.1 Summary of the cross-model evidence

Generation scores from separate models improve Qwen3.8-27B's weaker likelihood readouts on SugarCrepe, but no reliable gain remains over its selected or prompted readouts. Table 25 separates the original and aligned-control protocols; the strongest combination controls are in Appendix B.3. These experiments test cross-model transfer, not complementarity within Qwen's weights.

Receiver readout. Qwen3.8-27B scores the same 4,757 SugarCrepe pairs by teacher-forced mean caption likelihood, without sampling. Its original-caption accuracy is 0.797, versus 0.946 for per-caption yes/no. This readout gap bounds the interpretation of the following gains.

Table 25: Show-o generation margins raise Qwen3.8-27B's likelihood base from 0.797 to 0.880 and beat every aligned understanding control. Qwen3.8-27B cross-model accuracy on 4,757 SugarCrepe pairs (five-split means). Top: original generation models. Middle: independently calibrated generation models on aligned items; coin/sign-flip controls retain Show-o's delegated items. Bottom: best-single-readout base. Blocks use separate protocols and are not paired comparisons.

Evidence source / configuration	**Accuracy**
Original likelihood-base protocol	
None (Qwen3.8-27B likelihood base)	0.797
Show-o generation margins	**0.880**
Janus-Pro generation margins	0.833
Chameleon generation margins	0.826
Emu3 generation margins	0.814
BAGEL generation margins	0.801
Aligned control comparison	
Show-o generation margins	**0.879**
Qwen3.8-27B's second likelihood readout	0.807
Show-o understanding margins (pathway swap)	0.796
Janus-Pro understanding margins (accuracy-matched)	0.822
Coin on the same delegated slice	0.691
Sign-flipped generation answers	0.510
Best-single-readout protocol	
Strongest selected scoring base	0.942
Same base + Show-o generation gate	0.945
Residual: $+0.08$ to $+0.50$ points; 3/5 significant splits.	

Original-caption gains and controls. All five generation models improve this likelihood base. The analysis was preregistered in two stages: an audit on the original seeds and a control run on fresh seeds. Show-o (1.4B) raises accuracy to 0.880 ($p < 10^{-33}$ on every seed), with coverage 0.4 on all five calibrations from a grid including zero and full delegation. In the aligned control run, the G gate reaches 0.879. Replacing its delegated answers with a fair coin or inverted G gives 0.691 or 0.510; on those items U is correct on 0.740–0.765 and G on 0.924–0.957.

Three understanding controls use the same aligned items and seeds, with both margin normalisations available to each auxiliary model. A second Qwen likelihood readout reaches 0.807. Show-o U selects zero delegation under both normalisations and retains the 0.796 base. An approximately accuracy-matched Janus-Pro U readout (0.811 versus Show-o G 0.796) reaches 0.822 when gated. Show-o G beats all three on 5/5 McNemar tests.

Error overlap and grouped evaluation. Oracle headroom over the Qwen base is 0.114–0.164 across the five G sources. With Show-o G, the union accuracy is 0.961, near the 0.959 error-independence reference; with Qwen's alternate U normalisation it is only 0.852, against the same 0.959 reference. Replaying the control protocol on five fresh image-grouped splits (1,542 images) yields 0.799 to 0.878, with positive McNemar tests and group-bootstrap intervals on every split.

Stronger receiver baselines. On five fresh seeds, calibration always selects PMI over original-caption U. The base rises to 0.852; the same Show-o margins add 2.15 points to reach 0.874 ($p < 10^{-4}$ on every seed, coverage 0.4). Of 781 original-base errors that G can correct, 607 are also correctable by PMI.

Further readout repair removes the reliable residual. The fixed caption template scores 0.945 on the full set, near yes/no's 0.946. Under the common grouped protocol, the calibration-selected base averages 0.942 and its G gate 0.945: gains are 0.08–0.50 points, significant on 3/5 splits and below the four-of-five criterion. A preregistered gate against the stronger yes/no readout also finds no reliable change on any seed. Neither result establishes a capability ceiling. The positive evidence concerns weaker receiver readouts, even though the G source is nineteen times smaller. Model size alone gives no uniform ordering: BAGEL exceeds Qwen on some audited readouts.

C INTERFACES AND VISUAL EVIDENCE

Spatial tests show that G scores depend on the image regions that distinguish the captions on the tested backbones. The other diagnostics use separate rendering, representation-probe, or gate protocols. They do not form a common performance ranking. Earlier delegated-item controls require explicit gate decisions and have not been rerun on the corrections made by U+G or selected-ensemble fusion.

C.1 INTERFACE COMPARISONS AND DIAGNOSTIC CONTROLS

C.1.1 SUMMARY OF THE INTERFACE DIAGNOSTICS

Original-caption likelihood results are in Table 15. Rendering, probes, and replacement controls use distinct tasks and protocols, detailed in Appendices C.2, C.3, and C.4.3. Additional understanding-inference controls are in Appendix B.5.2. These diagnostics do not directly explain the learned U+G gains.

Coverage. Original-caption likelihood comparisons cover six backbones. True second-inference controls cover Show-o, Janus-Pro, Emu3, and Chameleon: G wins on 5/5 splits for all except Janus-Pro, which has no reliable difference. Rendering covers BAGEL and Emu3. Verdict probes cover nine concepts on Show-o, Janus-Pro, Emu3, BAGEL, and Chameleon; none of the 45 model–concept pairs reaches the preregistered five-point balanced-accuracy gain over both U and self-arbitration. This permits smaller gains and does not test decoded verbal judgments. Emu3.5 was not evaluated with these probes.

POPE ensemble-control extension. The later extension evaluates 8,982 questions from 499 images on Show-o, Janus-Pro, Emu3, and BAGEL, using five image-grouped calibration/test splits. It selects and freezes the selected ensemble before adding U or G, with 16 U candidates (12 caption-scoring configurations and four direct yes/no readouts) and four G candidates. All model weights remain frozen. Neither fusion nor gating has a jointly positive split on any backbone under the criterion in §3.4. Mean fusion increments over the U-select system are -0.19, -0.24, $+0.16$, and -0.08 percentage points, respectively. These results do not establish an advantage on object existence. The splits overlap, and the previously studied benchmark makes this extension a follow-up rather than an independent confirmation; U/G search menus also differ in size. Whole-image sign-flip tests with within-split Holm correction across models likewise yield no jointly positive splits.

Earlier exploratory POPE delegation controls. The earlier prompted-readout audit uses one frozen calibration/test split with 4491 test questions, prompted forced-choice U and the original G readout (mean-normalised for Show-o), not the final twelve-variant protocol. Its G-margin rule selects zero coverage and has zero increment on all four backbones. The separately calibrated agreement rule yields a Show-o increment of $+1.4$ points, with a bootstrap 95% interval $[+0.3, +2.4]$ and uncorrected one-sided $p = 0.0064$. POPE's rule comparisons were not multiplicity-corrected, so this is exploratory evidence under a different baseline and protocol; it does not establish gains over the stronger understanding controls in the later extension above.

Earlier prompted-gate attempt. A gate built on the order-biased prompted readout of Appendix A.3 failed all eight preregistered model–benchmark comparisons, including controls that re-asked the question. Reanalysis of its frozen logs with G-margin confidence recovered a Janus-Pro gain from 0.782 to 0.800 on one frozen split, beating equal-budget random routing; the other three backbones selected zero delegation. The subsequent likelihood-based gate improves five backbones under its own protocol (Appendix B.5). This sequence motivates checking the readout interface rather than treating prompted and likelihood scores as interchangeable.

C.2 DRAWING, READING BACK, AND CANDIDATE SELECTION

Paired rendering protocol. Motivated by rendered intermediates (Hu et al., 2024), we compare understanding answers about a scene with drawings of the same specification, scored by a rule-based detector. The preregistered test suite covers orientation, left–right position, counting, colour binding, majority colour, and claim verification on BAGEL and Emu3. Specifications are shared item by item; only those with both outputs enter paired comparisons. A frozen, value-blind legibility precheck determines generation sample sizes for each task and model.

For BAGEL, conditional drawing accuracies are 0.958 for left–right position, 0.977 for colour binding, and 0.974 for majority colour, but legibility rates are only 0.60, 0.54, and 0.65. Orientation rarely passes legibility (0.05); counting is usually legible (0.95) but less accurate (0.649). These results separate failures to produce a scorable scene from errors in its task-relevant values. They compare pathway competence on this suite, not an end-to-end system that adds drawings to understanding.

Candidate selection. Best-of-eight follow-ups test whether resampling can recover correct drawings and whether label-blind selectors can identify them. We use self-verification, a value-blind legibility check, and an oracle upper bound. The protocols and results below separate BAGEL's 270 specifications from Emu3's 75-specification subset.

C.2.1 Best-of-eight protocols and detailed results

Both readouts were preregistered before any candidate was drawn. Every specification keeps its frozen first drawing as candidate $k = 0$ and adds seven redraws under fresh derived seeds; the selection rules are label-blind: self-verification keeps the first candidate the model endorses when asked the understanding question about it, legibility keeps the first that passes the value-blind gate, and the oracle-of-eight is the non-deployable ceiling. The two readouts run on different fixed specification sets—BAGEL on all 270, Emu3 on the 75 specifications of the three patterns where its understanding readout errs, chosen before the run so that the rescue denominator is non-empty—so they are tabulated separately and never pooled.

Table 26: BAGEL readout (270 specifications). Detector-scored drawing accuracy per pattern for the frozen single draw, self-verified best-of-eight, and the oracle over the same eight candidates; U is understanding accuracy on the same specifications, whose errors form the rescue denominator (den).

Pattern	n	U	$N{=}1$	self-verified	oracle-of-8	den
Orientation	20	0.700	0.200	0.350	0.750	6
Left–right position	40	1.000	0.575	0.600	0.975	0
Counting	60	1.000	0.617	0.833	1.000	0
Colour binding	80	1.000	0.700	0.700	1.000	0
Majority colour	60	1.000	0.983	0.983	1.000	0
Claim verification	10	1.000	0.200	0.300	1.000	0

Table 27: Emu3 drawing accuracy on a separate set of 75 specifications, not directly comparable to Table 26. Self-verification and oracle select among eight draws; U is understanding accuracy, and den counts U errors eligible for rescue.

Pattern	n	U	$N{=}1$	self-verified	oracle-of-8	den
Orientation	20	0.200	0.000	0.000	0.150	16
Left–right position	40	0.950	0.125	0.125	0.675	2
Counting	15	0.733	0.133	0.133	0.533	4

Selection results. On BAGEL, self-verification raises drawing accuracy from 0.670 to 0.737, while the same candidates permit an oracle accuracy of 0.978. The selector recovers 22% of oracle headroom and remains below U's 0.978 on these specifications. BAGEL endorses 0.863 of its drawings although only 0.670 are correct.

Emu3 has 22 U errors eligible for rescue, compared with BAGEL's six (a 3.7-fold larger denominator). Eight draws leave its rescue rate at 0.045. Only 0.507 of specifications have any correct candidate; 38 of 75 have no endorsed candidate. The seven selections that move from the first draw never change correctness. Emu3 also over-endorses: 0.398 of drawings are endorsed, while 0.093 are correct.

Neither selector meets the preregistered rescue criterion: a significant McNemar win over the first draw and an absolute rescue rate of at least one quarter. BAGEL's denominator of six limits power; the Emu3 follow-up adds a larger rescue pool with the same null conclusion. The measured bottlenecks differ: BAGEL often has a correct candidate that selection misses, while Emu3 often has

Table 28: Generation-probe arbitration helps every architecture on at least two concepts, led by BAGEL (9/38). Concepts gaining at least five points over both the U probe and self-arbitration. Coverage reports helped/tested concepts; Shared 20 restricts counts to concepts tested on every model. These synthetic representation-probe diagnostics use a separate gate; Chameleon also uses separately trained LoRA adapters.

Model	Helped concept families	Coverage	Shared 20
BAGEL	colour, vertical position, touching, near–far, size–colour & shape–position binding, aspect ratio, symmetry ($\times 2$)	9/38	6
Janus-Pro	majority colour/shape family, size–colour binding, symmetry	7/38	2
Show-o	majority colour ($\times 3$), majority size	4/38	1
Chameleon	vertical position, row–column, hollow–filled	3/20	3
Emu3	majority shape, majority colour	2/33	0

none. These task-specific comparisons neither establish a general failure of rendering nor a causal advantage for likelihood scoring, which is evaluated on different tasks.

C.3 Concept evidence in hidden representations

Generation-side probes help every tested architecture on some visual concepts, exceeding the self-arbitration null. We use the class-mean probes and synthetic stimuli of Appendix A.5: G replaces the lowest-confidence 30% of U decisions on 38 visual concepts and nine verdict concepts. These are representation probes, distinct from caption likelihoods. Chameleon's trained LoRA pair is a separate diagnostic instrument; the main likelihood method updates no model weights. Table 28 reports both model-specific coverage and the 20 concepts shared by all five models.

Each architecture has helped concepts within its tested set, exceeding the preregistered self-arbitration null (95th-percentile pass count: zero per architecture). Four architectures retain helped concepts among the shared 20; Emu3's two lie outside that set. Help spans colour, spatial and shape properties, binding, and global statistics. None of the 45 verdict pairs passes its five-point criterion, though smaller improvements are possible. SugarCrepe swap-att also contains U-wrong/G-right items, motivating localization; such examples alone do not establish a net gain on that slice.

C.4 Error complementarity and generation preferences

We distinguish potential error coverage from improvements achieved by an implementable gate. The oracle analysis measures the former; fixed-delegation replacement controls test whether the generation answers carry useful directional information on the selected items.

C.4.1 Oracle error coverage on Show-o

Generation errors are close to independent of understanding errors on Show-o, whereas two understanding variants largely repeat the same errors. An oracle that takes either readout's answer when correct reaches 0.936 for U$\times$G—within 0.007 of the 0.943 reference point that two error-*independent* readouts of the same accuracies would show (exceeding it would require anti-correlated errors). U$\times$U (two variants of the understanding readout) reaches only 0.795, falling 0.130 below its own independence reference of 0.925: the alternate understanding normalisation largely repeats the same errors, while the generation pathway's errors are close to independent of them. The prompted readout paired with G does no better than pairing G with a fair coin (-0.0004) on this control. The gate converts +3.3 points of this headroom into realised accuracy beyond the stronger single readout (0.798 $\rightarrow$ 0.831).

C.4.2 Interpreting the overlap of errors and evidence

Aggregate patterns and item-level gains. Under the original-caption SugarCrepe protocol, adding Emu3.5 gives a six-backbone rank correlation of $\rho = -0.94$ between gain and the signed U-minus-G gap (exact two-sided permutation $p = 0.017$). U accuracy gives the identical ranking; these

are two summaries of the same outcomes. The five-backbone result and prediction record are in Appendix B.5.6. Nominal significance appears only after adding the sixth backbone, six points remain few, and the ordering does not persist under stronger U baselines; this analysis does not validate either statistic prospectively.

Standalone accuracy does not determine which errors G corrects: Emu3 gains despite weaker G, while BAGEL has concept-level gains despite its SugarCrepe null. Calibration finds useful delegation on ColorSwap, zero delegation on ARO, and unreliable delegation on Winoground. The relevant condition is the balance of corrections and new errors on the delegated items.

For delegation event D and correctness indicators U, G, the gain is exactly

$$\Delta = \Pr(D, U = 0, G = 1) - \Pr(D, U = 1, G = 0).$$

Thus the relevant condition is that corrections exceed corruptions within the delegated set. Neither global U accuracy nor its gap to G alone determines that condition; observed aggregate separations are descriptive.

C.4.3 Replacement controls on fixed delegated items

Fixed-delegation controls. Replacing G's delegated answers with a fair coin or inverted G loses on every nonempty SugarCrepe and SugarCrepe++ split. For each best-single-readout split, Table 29 holds the delegated items fixed and replaces G's answers with the exact fair-coin expectation or inverted G. All ColorSwap pairs have clean controls; SPEC has three clean and one mixed pair (Chameleon); ARO is not powered because its gates never delegate.

On Show-o/SugarCrepe, U and G accuracies on the delegated slice average 0.890 and 0.953. The gate gains 2.53 points overall, versus -15.75 for the coin and -34.03 for inverted G. These comparisons test useful answer direction on a selected slice, not reliable improvement over U: Emu3, Janus-Pro, and BAGEL lack reliable SugarCrepe gains despite costly random replacement. Nor do these controls evaluate U+G or fusion corrections.

The best-single-readout BAGEL gate delegates on 2/5 SugarCrepe splits and 1/5 SugarCrepe++ splits. Both pairs have fewer than three nonempty delegated splits and are marked not powered by the original noise-control rule. Their observed changes remain reported in Table 29; Table 30 preserves the earlier diagnostic. The final controls on the other backbones address the same random-replacement explanation under the revised protocol, without assuming that the original and final gates delegate the same items.

Table 29: G's delegated answers are directional: coin and inverted-G replacements lower accuracy on every SugarCrepe, SugarCrepe++, and ColorSwap pair. Noise controls for 27 model–benchmark pairs under the best-single-readout protocol. Changes are percentage points over selected U, averaged across five splits including empty-delegation cases. Coin reports its expectation; inverted-G reverses G on the same delegated items. Live counts nonempty splits; not powered does not indicate control success.

Pair	Live	G	Coin	Inverted-G	Noise verdict
SugarCrepe					
Show-o	5/5	+2.53	−15.75	−34.03	clean
Janus-Pro	5/5	+0.42	−8.74	−17.90	clean
Emu3	3/5	+0.08	−2.48	−5.04	clean
Chameleon	5/5	+1.21	−9.74	−20.69	clean
BAGEL	2/5	+0.03	−1.51	−3.05	not powered
Emu3.5	5/5	+9.56	−14.53	−38.62	clean
Qwen × Show-o	5/5	+0.30	−8.77	−17.84	clean
SugarCrepe++					
Show-o	5/5	+3.82	−15.33	−34.48	clean
Janus-Pro	5/5	+1.22	−7.49	−16.20	clean
Emu3	5/5	+2.72	−9.67	−22.07	clean
Chameleon	5/5	+3.03	−7.84	−18.71	clean
BAGEL	1/5	−0.04	−0.41	−0.77	not powered
Emu3.5	5/5	+11.00	−7.57	−26.14	clean
ColorSwap					
Show-o	5/5	+0.25	−23.62	−47.49	clean
Janus-Pro	5/5	+1.17	−27.83	−56.82	clean
Emu3	5/5	+1.00	−21.49	−43.98	clean
Chameleon	5/5	+4.99	−25.91	−56.80	clean
Qwen × Show-o	4/5	+0.00	−4.32	−8.64	clean
SPEC					
Show-o	4/5	−0.30	−1.24	−2.19	clean
Janus-Pro	5/5	+6.68	−17.07	−40.83	clean
Emu3	5/5	+4.90	−4.00	−12.90	clean
Chameleon	5/5	+4.59	+0.04	−4.50	mixed
ARO					
Show-o	0/5	+0.00	+0.00	+0.00	not powered
Janus-Pro	0/5	+0.00	+0.00	+0.00	not powered
Emu3	0/5	+0.00	+0.00	+0.00	not powered
Chameleon	0/5	+0.00	+0.00	+0.00	not powered
BAGEL	0/5	+0.00	+0.00	+0.00	not powered

Table 30: BAGEL's original-caption noise diagnostic, retained separately from the best-single-readout controls in Table 29. Entries are accuracy changes in points averaged over the original five splits, including zero changes on empty-delegation splits.

Benchmark	+G	coin flip	inverted-G
SugarCrepe	+0.05	−1.85	−3.75
SugarCrepe++	−0.13	−0.51	−0.88

C.5 Spatial localization and causal intervention

C.5.1 Concentration and detailed localization results

Generation evidence is spatially concentrated. On the three tested autoregressive backbones—Chameleon, Janus-Pro, and Emu3—the generation readout sums per-token log-probabilities over the image's vector-quantised (VQ) grid, so its margin decomposes into the spatial map used here; Show-o's masked pseudo-ELBO and BAGEL's flow-matching proxy are not evaluated by this exact decomposition. It is highly concentrated: on the swap-att slice the top twentieth of tokens carries 0.239 of the total margin mass on Chameleon, 0.227 on Janus-Pro, and 0.247 on Emu3 (a uniform map would give one twentieth), with nearly identical figures on the replace-rel slice (0.222/0.217/0.236). The three backbones span a fixed 32×32 grid, a fixed 24×24 grid, and Emu3's variable native grids

of some eight thousand tokens, showing that concentration recurs across tokenizers and resolutions. In the rescued pool—items the understanding readout gets wrong and the generation readout gets right (134 of 666 swap-att items on Chameleon, 58 of 666 on Emu3)—the mass sits on the swapped objects themselves (Figure 3): these examples connect the score difference to the visual regions that distinguish the captions.

Generation margins concentrate in named-object regions and are sensitive to their occlusion. Two preregistered follow-ups make that reading quantitative. First, a pointing test against COCO bounding-box masks: the absolute margin mass inside boxes of named categories, divided by their grid area, exceeds the uniform-map value of one in each model–slice test; every interval still excludes one under a retrospective whole-image bootstrap reanalysis. Each backbone uses its own eligible item set (Table 31). Unnamed objects also attract margin mass; on items with both measurements, named objects receive additional enrichment, with every paired difference interval excluding zero. Second, an intervention on originally correct Chameleon items: graying the named-object boxes and re-encoding drops the generation readout's accuracy to 0.659 on swap-att and 0.623 on replace-rel, against 0.770 and 0.744 when an area-matched mirrored region is grayed instead on 135 and 223 items from 121 and 173 images. The whole-image 95% intervals for control minus target accuracy are $[2.90, 20.30]$ and $[4.74, 19.30]$ percentage points. Original item-level McNemar values were $p = 0.024$ and $p = 0.003$; these do not account for repeated images. The two masks cover an identical mean grid area of 0.20. The median surviving margin is 0.43 of the original when their boxes are occluded versus 0.63 for the control. The localization is therefore not only a concentration statistic: occluding the target regions causes a larger loss than the matched control on this selected pool. Boxes include background pixels, so this is region-level evidence, not segmentation-level object attribution. Localization covers three backbones and occlusion one; neither establishes the necessity of every region or token. Table 31 tabulates the pointing and intervention tests.

C.5.2 Pointing and intervention protocols

Both follow-ups were preregistered before any number was computed. The *pointing test* reuses the frozen per-token margins of §5 and adds no inference. For each item we take the union of COCO bounding boxes of the object categories the positive caption names (whole-word match against the 80 category names, their plurals and four frozen aliases), map them onto the token grid under each backbone's *actual* preprocessing—shortest-edge-512 with a centre crop for Chameleon, a square resize for Janus-Pro, aspect-preserving scaling for Emu3—and report named-object mass, the share of $|\mathrm{margin}|$ mass inside those grid cells divided by the share of grid cells they occupy. Items are eligible when at least one named category has a mask covering between 2% and 50% of the grid. Unnamed-object mass repeats the computation for boxes of categories present in the image that the caption does *not* name. Because named-object mass uses absolute margins, it quantifies discriminative magnitude rather than net positive evidence for the correct caption.

The *intervention* fills the named-object bounding boxes with gray, re-encodes, and re-scores the frozen generation readout; the control grays the same masks mirrored to the other side of the image, so the two conditions remove an identical pixel area (mean intersection-over-union with the target, at most 0.08). Items enter when the generation readout is originally correct and the mirrored mask overlaps the target by no more than a quarter. Its effect is therefore conditional on originally correct, geometrically eligible items. The original intervals and McNemar tests used caption-choice items as units. The retrospective reanalysis resamples whole images, retaining all eligible pairs from each sampled image and the pair-weighted estimand. It uses 2,000 bootstrap draws with seed 20260929 (20260930 for named-minus-unnamed enrichment), without changing scores, masks, or eligibility. The paired enrichment statistic uses only items with both measurements. Table 31 reports the unchanged point estimates and new grouped intervals. These per-comparison intervals are not adjusted across tests; the reanalysis adds no new images or independent confirmation.

Scope. Localization covers three backbones; occlusion covers eligible, originally G-correct items on Chameleon. The results support region-level visual dependence, without explaining all U+G or fusion corrections.

Table 31: Generation margins concentrate on named objects in every model–slice test, and occluding those objects costs Chameleon 11.11 and 12.11 points more than an area-matched control. Spatial grounding tests with retrospective whole-image bootstrap 95% intervals (2,000 draws). (a) Named-object margin mass (uniform expectation 1), unnamed-object mass, and their paired difference. Named-object mass uses all eligible pairs; unnamed-object mass and D use pairs with both measurements, so D need not equal the difference of the displayed means. (b) Chameleon occlusion accuracy and control-minus-target difference in points. (c) Median retained signed margins. Eligibility and point estimates are unchanged; masks include background pixels.

(a) Pointing: named-object enrichment and paired contrast

Model / slice	Pairs / images	Named mass [95% CI]	Unnamed mass	Paired pairs / images	Difference [95% CI]
Chameleon / swap	295/263	1.43 [1.37, 1.48]	1.26	185/166	0.12 [0.03, 0.20]
Chameleon / rel.	581/374	1.41 [1.35, 1.47]	1.26	383/257	0.12 [0.07, 0.18]
Janus-Pro / swap	333/298	1.47 [1.42, 1.53]	1.13	210/190	0.28 [0.21, 0.35]
Janus-Pro / rel.	656/421	1.41 [1.36, 1.46]	1.23	444/296	0.11 [0.04, 0.16]
Emu3 / swap	345/309	1.52 [1.47, 1.58]	1.29	218/197	0.18 [0.11, 0.25]
Emu3 / rel.	692/437	1.49 [1.44, 1.54]	1.35	472/306	0.10 [0.04, 0.17]

(b) Chameleon: accuracy after area-matched occlusion

Slice	Pairs / images	Gray named	Gray control	Δ (pts)	95% CI (pts)
swap	135/121	0.659	0.770	11.11	[2.90, 20.30]
rel.	223/173	0.623	0.744	12.11	[4.74, 19.30]

(c) Chameleon: median fraction of margin retained

Slice	Gray named	Gray control
swap	0.43	0.63
rel.	0.26	0.72

D Calibration requirements and computational cost

For the main U+G method, Table 5 reports label budgets and Appendix D.5 gives the deployment procedure. The detailed budget, label-free, and transfer analyses below concern earlier single-readout gates. U+G and selected-ensemble deployment latency has not been measured; we report their scoring requirements separately from the measured single-readout latency.

D.1 Inference and selection costs

Single-readout pipelines. Per candidate caption, U requires one teacher-forced sequence evaluation. Autoregressive G adds one image-sequence evaluation; Show-o batches eight masked conditions in one forward call, while BAGEL evaluates five flow timesteps sequentially with text-conditioning overhead. Both margins are computed for every item regardless of delegation. A best-single-readout gate runs the selected U and G configurations, plus text-only scoring if U uses PMI. It requires no weight training or rendering. The checkpoint exceptions are listed in Appendix A.1.

Chameleon's original two-readout, two-caption pipeline takes 0.26 s per item on one GPU, the only measured end-to-end latency. Image sequences have 256 codes on Show-o, 576 on Janus-Pro, 1,024 on Chameleon, and variable counts on Emu3 and Emu3.5. These architectural differences preclude treating two scoring directions as a universal doubling of compute. The drawing diagnostics use different tasks and offer no matched latency comparison.

U+G and ensemble deployment. U+G requires four image-conditioned U passes, two text-only PMI passes, and four G readouts per caption. Mean and sum aggregation reuse the same scores, giving twelve U and four G margins. A selected-ensemble gate or fusion instead runs only the readouts required by that ensemble and its auxiliary G configuration; the ensemble itself need not cost one U pass.

On Chameleon, Emu3, and Emu3.5, U and G each use a teacher-forced pass over an image–caption sequence and score different positions. Janus-Pro uses separate understanding and generation in-

terfaces, Show-o masked conditions, and BAGEL flow timesteps. Equal calibration labels or pass counts therefore do not establish matched FLOPs or latency.

Calibration search. After base selection, U and G gates search 132 and 44 nonzero readout/coverage pairs; fixed-base U fusion has 252 grid entries and free two-U fusion 1,386. The selected-ensemble extension reuses these auxiliary menus. All combination analyses run on cached margins and add no GPU inference. The budget curves below refit single-readout gates; they do not measure reduced-label performance of the larger selected-ensemble search.

D.2 Calibration label requirements

Original single-threshold protocol. Subsampling labels for the coverage threshold reaches within 0.5 points of full-calibration accuracy with 200 pairs on Show-o, Janus-Pro, and Emu3, 400 on Chameleon, and 25 on BAGEL. This protocol selects only coverage; the following experiment also selects U and G configurations.

All best-single-readout pairs. A retrospective replay covers all 27 pairs on the same five grouped splits, reproducing the earlier eight-pair analysis. Its plan was fixed before computing the additional curves. Within each dataset, models share nested labeled-pair subsets; the original subsets are retained and extensions use the same frozen permutation rule. Budgets are 25, 50, 100, 200, 400, 800, and 1600 when every calibration half supports them, plus full calibration. SPEC omits 1600; its full halves contain 1570–1640 pairs. Each subset selects the U base, G readout, coverage, and threshold before evaluation on the unchanged test half. Full-budget configurations and predictions match the original panel.

Table 32 reports the smallest tested budget m^* reaching within 0.5 points of that pair's full-calibration mean accuracy; Table 33 gives every curve. This is a descriptive approximation to full performance, including pairs without reliable G gains, not a minimum label requirement for obtaining a gain. Curves can be nonmonotone, so larger budgets need not meet the tolerance after it is first reached. Selection never uses test labels.

Subsets reach the tolerance on 23/27 pairs, including 9/11 pairs with reliable full-calibration gains; their budgets range from 25 to 1600. The other two gaining pairs require full calibration. At 25 labels, however, 8/27 have negative mean increments over U selected with the same labels. All five ARO pairs lose 0.73–3.38 points despite selecting zero delegation at full budget.

Against a different reference, the U base selected with full calibration, 14 pairs fall behind by more than 0.5 points at some budget of at least 50 labels (34 budget–pair instances). This combines changes in base selection and delegation, rather than isolating G. All budget comparisons are means over overlapping splits, without new significance claims.

D.3 What labeled calibration adds

Original protocol. A label-free gate can delegate the 10% of items with the largest G margins, using the unlabeled set's own quantile. On the four originally replicating backbones it gains 1.1–2.7 points on SugarCrepe and 1.1–1.8 on SugarCrepe++, but loses 1.2–2.6 on ARO, where calibrated gates select zero. Delegating whenever G's within-set margin rank exceeds U's is also label-free; it loses 9.7 points on BAGEL/SugarCrepe++, where the calibrated gate loses 0.13 points. Margins supply a ranking; calibration determines whether and how much to delegate.

Expanded search on earlier gaining pairs. The expanded search, over all bases and generation readouts, covers the eight pairs that gained under the original-caption protocol plus five ARO pairs, using the same group-disjoint splits as the final results. The fixed rule uses original-caption readouts and delegates the ten percent with the largest generation margins at the unlabelled set's own quantile. It gains $+1.2$ to $+3.5$ points on the eight selected pairs, recovering 18–47% of the full system's gain over original-caption understanding, and loses 1.1 to 2.5 points on the five ARO pairs, where calibration selects zero delegation. This selected-panel comparison has not been extended to all 27 pairs.

Table 32: Label subsets reach within 0.5 points of full calibration on 23/27 pairs. Calibration budgets for 27 model–benchmark pairs under the best-single-readout protocol. m^*: smallest tested budget reaching full-calibration accuracy minus 0.5 points or higher. $\Delta_{\text{full}}/\Delta_{25}$: mean G gains over U selected with the same full/25-label budget, in percentage points. Sig. counts positive McNemar tests; joint also requires a positive group interval. Bold requires 4/5 joint positives. Zero_{25} counts zero-delegation splits.

Model	m^*	Δ_{full}	Sig.; joint	Δ_{25}	zero_{25}
SugarCrepe					
Show-o	50	$\mathbf{+2.53}$	5/5; 5/5	$+1.77$	0/5
Janus-Pro	100	$+0.42$	1/5; 1/5	$+0.55$	2/5
Emu3	100	$+0.08$	0/5; 0/5	-1.35	3/5
Chameleon	25	$\mathbf{+1.21}$	4/5; 4/5	$+2.63$	0/5
BAGEL	200	$+0.03$	0/5; 0/5	$+0.00$	5/5
Emu3.5	400	$\mathbf{+9.56}$	5/5; 5/5	$+7.90$	0/5
Qwen × Show-o	50	$+0.30$	3/5; 3/5	-0.09	4/5
SugarCrepe++					
Show-o	800	$\mathbf{+3.82}$	5/5; 5/5	$+3.62$	1/5
Janus-Pro	1600	$\mathbf{+1.22}$	5/5; 5/5	$+0.76$	0/5
Emu3	200	$\mathbf{+2.72}$	5/5; 4/5	$+1.81$	1/5
Chameleon	1600	$\mathbf{+3.03}$	5/5; 5/5	$+1.31$	1/5
BAGEL	200	-0.04	0/5; 0/5	$+0.09$	4/5
Emu3.5	1600	$\mathbf{+11.00}$	5/5; 5/5	$+12.12$	0/5
ColorSwap					
Show-o	100	$+0.25$	0/5; 0/5	$+1.31$	2/5
Janus-Pro	100	$+1.17$	2/5; 2/5	$+0.33$	2/5
Emu3	full	$+1.00$	1/5; 1/5	$+1.11$	2/5
Chameleon	full	$\mathbf{+4.99}$	5/5; 5/5	$+2.79$	1/5
Qwen × Show-o	400	$+0.00$	0/5; 0/5	$+0.00$	5/5
SPEC					
Show-o	100	-0.30	0/5; 0/5	-0.26	2/5
Janus-Pro	100	$\mathbf{+6.68}$	5/5; 5/5	$+9.45$	0/5
Emu3	full	$\mathbf{+4.90}$	4/5; 4/5	$+1.70$	1/5
Chameleon	full	$+4.59$	4/5; 1/5	$+4.44$	0/5
ARO					
Show-o	800	$+0.00$	0/5; 0/5	-0.73	3/5
Janus-Pro	800	$+0.00$	0/5; 0/5	-2.09	1/5
Emu3	400	$+0.00$	0/5; 0/5	-3.38	3/5
Chameleon	100	$+0.00$	0/5; 0/5	-1.25	2/5
BAGEL	400	$+0.00$	0/5; 0/5	-0.74	4/5

Table 33: Best-single-readout gate accuracy at each calibration budget (five grouped test splits). Bold shading marks m^* from Table 32; gray marks full calibration, with sizes listed by dataset. n/a: budget unavailable on some splits. Selection uses unrounded values.

Model	25	50	100	200	400	800	1600	full
SugarCrepe; full = 2357–2428 labelled pairs								
Show-o	0.871	**0.873**	0.871	0.871	0.871	0.872	0.873	0.877
Janus-Pro	0.884	0.887	**0.913**	0.911	0.915	0.914	0.915	0.913
Emu3	0.876	0.883	**0.899**	0.897	0.899	0.900	0.899	0.900
Chameleon	**0.789**	0.782	0.783	0.792	0.788	0.795	0.788	0.794
BAGEL	0.924	0.923	0.930	**0.933**	0.935	0.934	0.936	0.935
Emu3.5	0.774	0.781	0.785	0.796	**0.799**	0.805	0.804	0.802
Qwen × Show-o	0.933	**0.943**	0.942	0.942	0.942	0.945	0.946	0.945
SugarCrepe++; full = 2357–2428 labelled pairs								
Show-o	0.763	0.777	0.777	0.791	0.791	**0.796**	0.799	0.798
Janus-Pro	0.801	0.786	0.804	0.801	0.798	0.810	**0.814**	0.816
Emu3	0.779	0.772	0.794	**0.803**	0.803	0.802	0.801	0.801
Chameleon	0.684	0.690	0.703	0.714	0.715	0.718	**0.726**	0.725
BAGEL	0.827	0.824	0.877	**0.886**	0.885	0.886	0.886	0.885
Emu3.5	0.686	0.708	0.699	0.713	0.716	0.717	**0.720**	0.724
ColorSwap; full = 718 labelled pairs								
Show-o	0.904	0.909	**0.921**	0.922	0.926	n/a	n/a	0.926
Janus-Pro	0.943	0.943	**0.952**	0.946	0.952	n/a	n/a	0.957
Emu3	0.893	0.900	0.911	0.915	0.914	n/a	n/a	**0.923**
Chameleon	0.863	0.886	0.878	0.885	0.884	n/a	n/a	**0.893**
Qwen × Show-o	0.951	0.965	0.962	0.964	**0.970**	n/a	n/a	0.971
SPEC; full = 1570–1640 labelled pairs								
Show-o	0.670	0.693	**0.721**	0.720	0.722	0.718	n/a	0.725
Janus-Pro	0.786	0.801	**0.825**	0.832	0.826	0.837	n/a	0.829
Emu3	0.673	0.653	0.684	0.687	0.693	0.696	n/a	**0.702**
Chameleon	0.532	0.530	0.531	0.526	0.521	0.537	n/a	**0.543**
ARO; full = 2417–2504 labelled pairs								
Show-o	0.792	0.775	0.787	0.789	0.797	**0.807**	0.808	0.808
Janus-Pro	0.762	0.772	0.767	0.779	0.783	**0.799**	0.802	0.802
Emu3	0.745	0.768	0.780	0.785	**0.794**	0.797	0.797	0.797
Chameleon	0.725	0.731	**0.735**	0.732	0.738	0.738	0.736	0.737
BAGEL	0.818	0.839	0.836	0.842	**0.852**	0.849	0.853	0.856

D.4 THRESHOLD AND CONFIGURATION TRANSFER

Freezing each split's absolute margin threshold on SugarCrepe and applying it unchanged to the full SugarCrepe++ set gives descriptive gain ratios of 87% on Show-o, 101% on Emu3, and 48% on Chameleon relative to native test-half gains. Target evaluation reuses source-calibration images; these are not paired, disjoint transfer estimates. Janus-Pro's transferred gain is -1.13 points, with transferred mean coverage 0.48 versus native SugarCrepe++ coverage 0.10–0.20.

The expanded search transfers the full SugarCrepe-selected configuration to SugarCrepe++ with aligned images and thresholds re-quantiled without target labels. Source-calibration images are excluded from target test. Transferred accuracy is 2.6–12.5 percentage points below native target-task calibration on all four shared backbones. These results favour recalibration to recover the available gains.

D.5 APPLYING THE METHOD TO A NEW MODEL OR BENCHMARK

U+G needs labeled target-task pairs, cached margins, and one fitted combiner. To apply the method to a new model or benchmark, draw a few hundred labeled caption pairs from the target distribution and group them by image. Score both captions of every pair with the twelve understanding readouts and the four generation readouts of §3.1, and store the signed margins. Scale each margin by its calibration standard deviation, fit the no-intercept L_2-regularized logistic combiner of §3.3 on the symmetrized pairs, choose the regularization strength by five-fold grouped cross-validation, and refit on all calibration pairs. Fit the same combiner on the twelve understanding margins alone and keep generation only if U+G is reliably better on held-out groups. Table 5 reports observed mean gains with 25–400 labels; it does not test significance at each budget.

Check the two conditions for a gain first. Gains appear where understanding leaves errors to correct and generation alone is informative (§4.2). If the understanding-only combination is already near ceiling, or generation alone is close to chance on the calibration pairs, little benefit should be expected. Prompted yes/no readouts can exceed likelihood readouts on some models (Appendix B.5.4), while A/B formats can show strong position bias (Appendix B.5.5). Evaluate candidate interfaces on the same labeled pairs before choosing the understanding pool.

A single-readout gate is a cheaper alternative. When scoring all sixteen readouts is too expensive, one understanding and one generation readout can be combined by the margin-based gate of Appendix A.4.2: select the base, generation readout, and coverage on the labeled pairs, including zero coverage, and freeze the margin threshold. Table 32 gives its label budgets. In either case, calibration is an empirical assessment of the model–task pair, not a guarantee under distribution shift, as the Winoground and configuration-transfer results show.

E EXTERNAL IMAGE–TEXT REFERENCE SYSTEMS

On SugarCrepe, U+G on Show-o, Chameleon and Emu3.5 exceeds CLIP-L by 6.5–13.9 points, and U+G on BAGEL and on Qwen3.8-27B with Show-o generation scores exceeds BLIP-ITM (descriptive comparisons). We score every frozen image–caption pair using CLIP ViT-L/14 (Radford et al., 2021), NegCLIP ViT-B/32 (Yuksekgonul et al., 2023), and BLIP-ITM base COCO (Li et al., 2022). We preserve the exact 4,757 SugarCrepe, 4,757 SugarCrepe++, 1,436 ColorSwap, 3,220 SPEC, and 5,000 ARO pair IDs and the five grouped calibration/test splits of the main comparison. The external systems use no target-task calibration labels or fitted combination. We report the arithmetic mean of their five test accuracies, not full-pool accuracy or numbers quoted from previous papers. The overlapping tests are not five independent replications.

Frozen implementation. CLIP uses the official OpenAI checkpoint and processor. NegCLIP uses the original authors' released checkpoint with the ViT-B/32 QuickGELU architecture. Both score normalized image–text cosine similarity using original captions and a 77-token limit. BLIP uses `Salesforce/blip-itm-base-coco` and its image–text matching head, taking the positive-class softmax probability with original captions and a 35-token limit. We follow the CLS input used by the official standalone BLIP-ITM forward. Reusing image features reproduces the public full forward exactly on the verification batch. Softmax is evaluated in float64 from FP32 logits to avoid

Table 34: BLIP-ITM is the strongest external reference on every dataset. External reference accuracy (%) on the original five grouped tests. These are standalone system comparisons with different pretraining and supervision, not controlled evidence-addition ablations.

Dataset	CLIP ViT-L/14	NegCLIP ViT-B/32	BLIP-ITM
SugarCrepe	76.36	83.82	92.30
SugarCrepe++	76.87	73.52	87.25
ColorSwap	62.65	68.69	97.35
SPEC	52.20	49.48	54.61
ARO	56.63	66.04	71.56

artificial saturation ties. The positive-minus-foil margin must be strictly positive; ties count as errors. CLIP and NegCLIP truncate no captions. BLIP truncates 7 SugarCrepe pairs and 10 SugarCrepe++ pairs; all are retained. No system produces a score tie in these pools. Checkpoint hashes, scores, truncation flags, and per-split statistics are retained with the experiment record.

Interpretation. On SugarCrepe, U+G on Show-o, Chameleon and Emu3.5 (0.902, 0.883, 0.828) exceeds CLIP-L (0.764) by 6.5–13.9 points; we did not run paired tests for these external comparisons. NegCLIP reaches 83.82% and BLIP-ITM 92.30%, above these three U+G systems, while U+G on BAGEL (0.969) and Qwen3.8-27B with Show-o generation scores (0.973) exceed BLIP-ITM. NegCLIP and this BLIP checkpoint were fine-tuned for COCO retrieval, and the unified systems use task-specific calibration labels. Thus the external rows provide empirical reference points across different training settings; they do not isolate G's contribution. The latter is addressed by the within-backbone comparisons and learned U control.

The reference also exposes task-dependent weaknesses. On SugarCrepe++, Chameleon and Emu3.5's original ensemble+G fusion accuracies (0.734 and 0.729) fall below CLIP-L (0.769), while their learned U+G combinations reach 0.831 and 0.854. On SugarCrepe, CLIP-L exceeds Emu3.5's selected understanding base (0.711) and the learned U-only combination (0.747), whereas U+G reaches 0.828. These are descriptive system comparisons; the contribution of G is assessed against the corresponding U-only system.

F Prospective test of generation evidence beyond learned U scores

Question and scope. After the retrospective panel, we fixed a single new test of whether the learned combination of twelve U and four G margins improves accuracy over the learned combination of the same twelve U margins. The backbone is the existing Show-o checkpoint (1.4B unified model, with a Phi-1.5 core); its earlier results were known when choosing this test. The dataset is CREPE Productivity (Ma et al., 2023), which had not been used in the accessible experiment history. We adapt its multi-candidate retrieval setting to two-caption discrimination. This tests the learned combinations of Appendix B.3.3, rather than retesting the original ensemble+G fusion or the five overlapping splits.

Freeze and data selection. Before any CREPE model scores were produced, a local, timestamped protocol fixed the data, readouts, fitting rule, sole primary endpoint, and decision criterion; code, data and checkpoint hashes were recorded. This was not an external-registry preregistration. We used the official complexity 4–12 atom, swap, and negate CSV files from CREPE repository revision `1fa81c425f442396fe304f170c8eb6dc0747c814`. We excluded every Visual Genome image linked to COCO in the official metadata and every image in the original ARO relation/attribution annotations. Eligibility also required an available valid image, a positive-area crop, and a nonempty foil different from its positive. Of 4,170 eligible images, we selected 3,000, with one supplied positive/foil pair per image and the supplied region crop. There was no manual caption rewriting or score-based exclusion.

Selection used SHA-256 with seed string `20260925-crepe-confirmation-v1`: choose the minimum pair hash within each image, select the first 3,000 image IDs under an independent sample hash, and order them under a separate split hash. The first 1,000 images form the calibration set; the remaining 2,000 form the sole held-out test. The exact identifiers, captions, source rows, crops and

split assignments are retained. Calibration and test have no original-image overlap. No alternative split, backbone or dataset was selected after viewing outcomes; model pretraining exposure to these images is unknown.

Readouts and fitting. We reused the existing Show-o scorer: FP32 weights, native 256-pixel image encoding, eight mask ratios from 0.1 to 0.8 with mask seed zero, and the four caption templates of Appendix A. U includes mean/sum caption likelihoods for all four templates and mean/sum PMI for the first two; G includes summed masked-image pseudo-likelihood margins for all four. Two previously scored SugarCrepe items reproduced all 32 packaged margins exactly before the new-data run. Both combinations use the same 1,000 calibration labels and the symmetric, no-intercept L2 logistic learner. Features are divided by their calibration standard deviations plus 10^{-12} without centering. As in the retrospective learned control, this calibration-only scale is shared across the internal folds. Unshuffled five-fold GroupKFold selects $C \in \{.001, .01, .1, 1, 10, 100\}$ by mean validation accuracy; ties within 10^{-12} choose smaller C. The learner is then refitted on all calibration items, with 5,000 maximum iterations and tolerance 10^{-8}. The analysis used NumPy 1.26.4, SciPy 1.17.1, and scikit-learn 1.8.0; the numerical supplement pins these versions for replay. Both readouts selected $C = 10$. They have twelve versus sixteen coefficients, so this is an equal-label comparison with a common learner, not equal capacity. A strictly positive combined margin is correct; zero counts as an error.

Sole primary endpoint and result. The prespecified endpoint is the test accuracy of U+G minus that of the learned U-only combination, in percentage points. Confirmation requires a positive difference, an exact two-sided McNemar $p < 0.05$, and a 95% percentile whole-image bootstrap interval entirely above zero (10,000 resamples, seed 2026092501). With one pair per image, resampling images is equivalent to resampling pairs. Both systems classify 1,933 of 2,000 test pairs correctly (96.65%). The primary difference is **0.00 percentage points**, with 95% interval $[-0.45, +0.45]$ and $p = 1.0$. Adding G recovers eleven errors and introduces eleven errors. The frozen criterion therefore does not confirm an accuracy increment on this task. No interim task accuracies were printed during inference, and no secondary subgroup hypothesis tests were run.

Interpretation and reproducibility. The null is consistent with the headroom condition. Its 96.65% baseline lies above every panel pair with a significant increment (learned U-only accuracy ≤ 0.922, Table 13) and in the same high-accuracy regime as the five panel pairs above 93.8%, none of which meets the improvement criterion. Conditional on the observed 22 discordant pairs, the frozen exact McNemar rule would require at least 17 in G's favour, a $+0.60$-point increment on 2,000 test items. This is a conditional calculation, not a prespecified power bound. The null provides no evidence of an increment on this task; it neither invalidates the retrospective positives nor provides a rule for predicting new-task gains. CREPE's automatically generated negatives can contain language artifacts, and this adapted task does not establish visual grounding. The numerical supplement provides the protocol, frozen sample manifest, cached margins, fitted coefficients, test predictions and CPU replay of the one endpoint. Independent checks reconstructed all 48,000 margins from raw caption scores, all 4,000 test predictions, and the primary statistics. Scoring took about 30 minutes across four GPU processes, whose scoring wall times sum to 1.91 hours; this is not exclusive GPU allocation time. Peak allocated memory per process was 7.08 GB.