ALIGNED IS NOT SUPPORTED: ATTRIBUTING IMAGE–TEXT ALIGNMENT TO CAPTION PHRASES FOR DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image–text filters for CLIP-style pretraining score each pair with one number. We show that this number saturates: on CC3M, once the least-aligned 30% of pairs are removed, stricter alignment thresholds stop improving compositional benchmarks and begin to reduce zero-shot transfer. The reason is structural. A global score measures whether a pair matches, not whether the caption's phrases carry that match; a caption can be well aligned yet *hollow*, sustained by one salient object, an OCR fragment, or a text prior. We introduce Counterfactual Phrase Intervention (CPI), which attributes the alignment score back to the caption: each nominal head is replaced by a nonce token matched in CLIP-BPE length and surface form, and the resulting similarity drop measures how much of the match that phrase carries. In other words, CPI asks not whether a caption matches its image, but which words in it do the matching. Mean phrase sensitivity correlates only weakly with alignment ($r = 0.084$), and rankings induced by noun-phrase and relational contexts are nearly independent ($\rho = 0.059$), so the signal adds information that no global score can recover. Ranking the aligned pool by this signal yields a 50%-data subset that improves VL-CheckList Relation by $+1.73$ over full data and $+0.98$ over alignment-only filtering at matched budget (five paired seeds, $p = 0.002$), is within seed variance of the 70% alignment pool on compositional metrics with 28.7% fewer samples, and preserves zero-shot and retrieval transfer. Used without any change to the objective, the same subset improves NegCLIP by $+3.84$ and CE-CLIP by $+4.92$ on VL-CheckList Relation, and the effect replicates on CC12M, on ViT-S/32 and ViT-L/14, and under a SigLIP scorer (80.6% selection overlap against 71.3% at chance). We present these results as controlled evidence at CC3M–CC12M scale; curation costs 1.70 GPU-hours with cached embeddings, and code is provided.

1 INTRODUCTION

A data filter for image–text pretraining has one number per pair. CLIPScore (Hessel et al., 2021), DataComp (Gadre et al., 2023), DFN (Fang et al., 2024), T-MARS (Maini et al., 2024), and Sieve (Mahmoud et al., 2024) produce that number differently, but all threshold a scalar that says how well an image and a caption match. This paper asks whether that scalar is the right quantity to threshold once the obvious mismatches are gone.

On CC3M it is not. We sweep the alignment retention ratio from 100% to 50% and pretrain CLIP ViT-B/32 on each subset (Section 3). Every subset beats the full corpus on SugarCrepe, so alignment does remove harmful pairs; yet SugarCrepe Overall peaks at the 70% pool (58.34) and falls to 57.92 at 50%, SugarCrepe++ Replace falls from 47.82 to 46.65, and CIFAR10 zero-shot drops from 46.85 to 42.49. We call this regime *alignment saturation*: past coarse mismatch removal, ranking plausible pairs by the same scalar no longer tracks the supervision they provide.

The cause is what a global score cannot see. Two captions with the same similarity can differ in *where* it comes from: in one, each object, attribute, and relation contributes; in the other, a single salient object, an OCR fragment, or a text prior sustains it and the remaining phrases are decorative. We call the second kind a *hollow caption*. Hollow captions pass any pair-level filter and are the pairs

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.

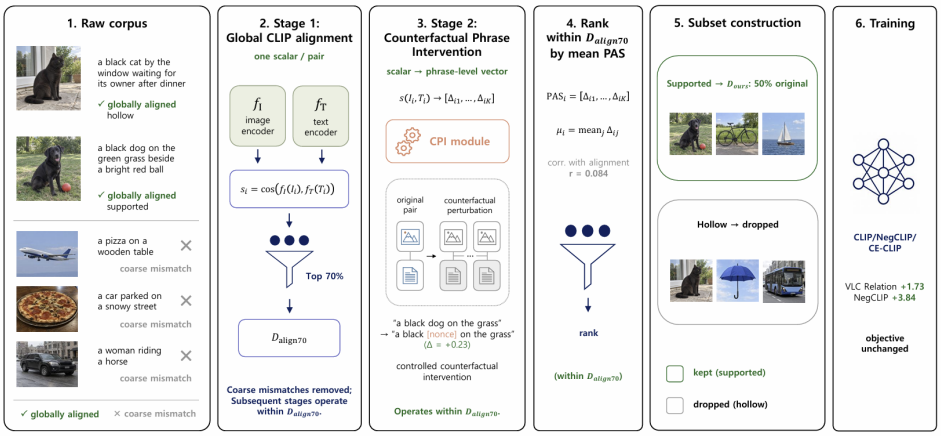


Figure 1: **Alignment tells how well a pair matches; CPI tells which phrases carry the match.** Stage 1 removes coarse mismatches by global CLIP alignment (top 70%). Stage 2 replaces each nominal head with a length- and form-matched nonce token, averages the similarity drops into a phrase-sensitivity score ($r = 0.084$ with alignment), and keeps the top 50% of the corpus. Training on this subset with an unchanged objective improves VL-CheckList Relation by $+1.73$ over full data and $+0.98$ over alignment-only filtering (CC3M, ViT-B/32, five paired seeds, $p = 0.002$), and by $+3.84$ under NegCLIP. Curation costs 1.70 GPU-hours.

compositional benchmarks expose (Yuksekgonul et al., 2023; Thrush et al., 2022; Hsieh et al., 2023). Separating them requires asking not *how well* a pair matches but *which phrases carry the match.*

We answer that question with **Counterfactual Phrase Intervention (CPI)**, a curation signal that attributes the alignment score to the phrases of the caption. For each phrase p_j of caption T, CPI replaces its nominal head w_j with a nonce token r_j and measures the drop in image-conditioned similarity, $\Delta_j = s(I, T) - s(I, \tilde{T}_j)$. A *Three-Invariance Replacement Protocol* makes r_j match w_j in exact CLIP-BPE subtoken count and in capitalization, number, and possessive form while carrying no lexical meaning. Under these constraints Δ_j is a first-order sensitivity of the fixed scorer to removing the content of w_j (Proposition 2), and the caption mean is the first-order term of a Shapley attribution of the match to its phrases (Proposition 3; Shapley, 1953). Concretely, a caption whose phrases each move the score when removed is supported; one whose score barely moves is hollow. The protocol covers 99.79% of aligned captions and scores CC3M in 1.70 GPU-hours with cached embeddings.

CPI is the second stage of a coarse-to-fine pipeline (Figure 1): Stage 1 keeps the 70% most aligned pairs, where saturation begins; Stage 2 ranks that pool by mean phrase sensitivity and keeps 50% of the corpus. The stages measure different things (mean sensitivity correlates with alignment at $r = 0.084$ within the pool; Section 6). In simpler terms, Stage 1 asks whether a pair matches; Stage 2 asks whether the caption earned the match.

We state the scope before the results. CPI intervenes on nominal heads in four syntactic contexts, not on adjectives, prepositions, or verbs, and it neither localizes phrases nor certifies grounding; it measures context-conditioned sensitivity of a fixed scorer to nominal substitution. Whether that signal, used only to select positives, improves compositional pretraining is the empirical question; the benchmarks are the test, not the construct.

The answer, at CC3M–CC12M scale, is yes. Training CLIP ViT-B/32 on the CPI subset raises VL-CheckList Relation from 37.90 to 39.63 ($+1.73$ over full data; $+0.98$ over alignment-only filtering at the same budget, higher in all five paired seeds, $p = 0.002$) and SugarCrepe Overall by $+0.89$ and $+0.40$ ($p = 0.010$), with zero-shot, retrieval, and linear-probe transfer within seed variance of the strongest baseline (Section 5). The subset is within seed variance of the 70% pool on both metrics with 28.7% fewer samples, whereas random or alignment-based pruning of the same pool falls below the pool, and it matches or exceeds T-MARS on every compositional metric given the same pool. Used unchanged, it improves NegCLIP by $+3.84$ and CE-CLIP by $+4.92$ on Relation, and the effect

replicates on CC12M, ViT-S/32, ViT-L/14, and under a SigLIP scorer. Gains on SugarCrepe++ and word-order probes are small and in some settings negative; we report and interpret them in full.

This work makes three claims: pair-level alignment filtering saturates, and pruning past the 70% pool by alignment lowers compositional and zero-shot metrics; controlled nonce substitution turns a pair-level score into a phrase-level attribution that is nearly orthogonal to alignment, stable across extractors and nonce protocols (88–97% selection overlap), and robust to a change of scorer; and selecting positives by this attribution improves relation-sensitive evaluation without touching the objective and compounds with hard-negative objectives. We present this as controlled evidence at CC3M–CC12M scale; from-scratch web-scale pretraining remains open (Section 7).

2 Related Work

Four lines of work meet here: filters that score pairs, selection rules that score examples by their value to a learner, objectives that make contrastive models composition-sensitive, and interventions that attribute an output to parts of its input. We state what each measures and where CPI departs.

Pair-level filtering scores how well a pair matches. Contrastive pretraining on web pairs (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023) depends on curation, and the curation literature has converged on one design: assign each pair a scalar and threshold it. CLIPScore thresholds pretrained similarity (Hessel et al., 2021); DataComp standardizes the recipe and shows it beats random and text-heuristic baselines (Gadre et al., 2023); DFN learns the scalar from high-quality pairs (Fang et al., 2024); T-MARS suppresses pairs whose similarity is explained by rendered text (Maini et al., 2024); Sieve rescores with a captioner (Mahmoud et al., 2024); MetaCLIP balances the pool over metadata (Xu et al., 2024); caption rewriting replaces the text before scoring (Nguyen et al., 2023); and filtering scaling laws show the value of a scalar filter depends on compute (Goyal et al., 2024). All answer the same question with better instruments: is this pair a good match? None asks how the match is distributed over the caption's phrases. CPI keeps the scalar for coarse mismatch removal and adds a per-phrase attribution that the scalar cannot express; the two are nearly independent inside the aligned pool ($r = 0.084$), and given the same pool and budget CPI matches or exceeds T-MARS on every compositional metric (Section 5).

Example-level selection scores value to a learner. Influence functions (Koh & Liang, 2017), early gradient norms (Paul et al., 2021), margin statistics (Sorscher et al., 2022), holdout-loss reduction (Mindermann et al., 2022), online batch utility (Evans et al., 2024; Udandarao et al., 2025b), and deduplication (Abbas et al., 2023; Tirumala et al., 2023) score an example's relation to a learner or to the corpus and need a training loop, a reference model, or corpus statistics. CPI scores a property of the pair itself, the internal support of its caption under a fixed scorer, with no training loop.

Objective-side methods change the negatives; CPI changes the positives. Contrastive models behave like bags of words on order- and relation-sensitive tests (Yuksekgonul et al., 2023; Thrush et al., 2022; Diwan et al., 2022; Kamath et al., 2023). The standard remedy modifies the objective: NegCLIP adds word-order hard negatives (Yuksekgonul et al., 2023), CE-CLIP ranks intra- and cross-modal negatives (Zhang et al., 2024), TripletCLIP and generative negative mining synthesize negatives in both modalities (Patel et al., 2024; Sahin et al., 2024), structured captions and scene graphs inject relational supervision (Doveh et al., 2023b;a; Herzig et al., 2023), and FSC-CLIP regularizes fine-tuning (Oh et al., 2024). Every one takes the positives as given and works on the loss; CPI takes the loss as given and works on the positives. The levers compose: the CPI subset, applied unchanged, improves NegCLIP by $+3.84$ and CE-CLIP by $+4.92$ on VL-CheckList Relation (Section 5.4).

Benchmarks and their biases set the evidentiary bar. Compositional benchmarks probe objects, attributes, relations, and order (Zhao et al., 2022; Parcalabescu et al., 2022; Ma et al., 2023). SugarCrepe removes the plausibility and fluency artifacts that let text-only models succeed (Hsieh et al., 2023), SugarCrepe++ adds lexical-paraphrase sensitivity (Dumpala et al., 2024), and a recent audit shows blind heuristics still match CLIP on several older benchmarks (Udandarao et al., 2025a). We therefore treat VL-CheckList Relation as a benchmark-specific measurement, report SugarCrepe as the de-biased primary test, give SugarCrepe++ under both aggregation conventions because the

benchmark defines no official overall, and report the word-order probes on which CPI does not improve; phrase-level evaluation of generated images (Cho et al., 2024) shares the premise at evaluation rather than curation time.

Controlled substitution moves from diagnosis to selection. Occlusion (Zeiler & Fergus, 2014), integrated gradients (Sundararajan et al., 2017), Shapley values (Shapley, 1953; Lundberg & Lee, 2017), concept vectors (Kim et al., 2018), and counterfactual explanations (Goyal et al., 2019) attribute a prediction to input components; minimal pairs (Marvin & Linzen, 2018; Ettinger, 2020; Ribeiro et al., 2020), counterfactual data and models (Kaushik et al., 2020; Feder et al., 2021; Pearl, 2009), and mediation or editing interventions (Vig et al., 2020; Meng et al., 2022; Oikarinen & Weng, 2023) do the same for language. All diagnose a trained model. CPI applies the same operation, a length-matched (Sennrich et al., 2016) lexical ablation, to a fixed scorer before training, and uses the attribution to decide which pairs a new model sees.

Baseline scope. The primary matched-budget comparison is alignment-only filtering at 50%, which differs from CPI only in whether the pool is further ranked by phrase sensitivity; we add the 70% pool itself, a random draw from it, and T-MARS as a standalone filter and as an alternative Stage-2 rule. DFN and Sieve need a learned filter or a captioner tuned to other corpora and operating points and are treated as complementary systems, not matched controls.

3 Alignment Saturation

Pair-level filtering is used as if it were monotone: a stricter threshold should keep better pairs and train a better model. We test the premise before building on it. We rank all CC3M pairs by CLIP similarity, keep the top $\rho \in \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5\}$, and pretrain CLIP ViT-B/32 on each subset under identical optimization (single seed; Table 4 in Appendix A).

Two facts follow. Alignment is a useful coarse signal: every subset, including top50, exceeds Full on SugarCrepe Overall, so the least-aligned pairs do hurt. Alignment is not a useful fine signal: SugarCrepe Overall peaks at top70 (58.34) and drops to 57.92 at top50, SugarCrepe++ Replace falls from 47.82 to 46.65, CIFAR10 zero-shot falls from 46.85 at top80 to 42.49 at top50, and Flickr T2I R@1 falls from 11.020 to 10.225. The decline appears on compositional, retrieval, and zero-shot axes at once. We call the regime past the peak *alignment saturation*.

Saturation is what one expects if the global score answers only one question. Below the peak it separates pairs that match from pairs that do not; past it, every surviving pair matches, and sorting by the same scalar sorts by whatever residual variation it carries: caption length, tokenization, OCR content, or a single dominant object. Section 4 makes this precise: the global score is the sum of a component that survives removal of the caption's content and a component that does not, and a threshold on the sum cannot separate them. We fix the top70 pool D_{align70} (1,557,650 pairs) as the Stage-1 candidate set for every later experiment.

4 Counterfactual Phrase Intervention

Proofs are in Appendix K.

4.1 Setup and the alignment decomposition

Let f_I and f_T be the image and text encoders of a fixed scorer, mapping to the unit sphere, and $s(I,T) = \langle f_I(I), f_T(T) \rangle$ its similarity. A phrase extractor returns K disjoint spans $p_1, \ldots, p_K$ of caption T, each with one nominal head token $w_j \in p_j$ (Appendix B). An *intervention* on phrase j substitutes its head, and the *phrase attribution score* (PAS) is the image-conditioned similarity drop it causes:

$$\mathcal{T}_j T := T[\, w_j \leftarrow r_j \,],\ r_j \in \mathcal{R}(w_j); \quad \Delta_j(I,T) := s(I,T) - s(I,\mathcal{T}_j T); \quad \boldsymbol{\Delta} := (\Delta_1, \ldots, \Delta_K). \tag{1}$$

The admissible set $\mathcal{R}(w_j)$ is defined in Section 4.2; $\boldsymbol{\Delta}$ is the *sensitivity profile*. Pair-level filters observe only $s(I,T)$. The following identity says what they miss.

Lemma 1 (Alignment decomposition). *For every pair with $K \geq 1$ admissible interventions,*

$$s(I,T) = h(I,T) + \mu(I,T), \qquad h := \frac{1}{K}\sum_{j=1}^{K} s(I, \mathcal{T}_j T), \qquad \mu := \frac{1}{K}\sum_{j=1}^{K} \Delta_j. \tag{2}$$

The identity is immediate; its content is what it denies a pair-level filter. The *hollow component* h is the similarity the pair keeps, on average, after one content word is removed; the mean PAS μ is what it loses. In simpler terms, every alignment score is the sum of a part that survives removal of the caption's content and a part that does not. A threshold on s is a threshold on $h + \mu$, so $(h, \mu) = (s, 0)$ and $(s - \mu, \mu)$ are indistinguishable to any pair-level filter. Empirically the split is nearly independent of the total: within D_{align70}, $\text{corr}(s, \mu) = 0.084$ (Section 6). Stage 1 thresholds s; CPI ranks by μ.

4.2 The intervention operator and the Three-Invariance Protocol

The identity holds for any replacement; its use depends on r_j removing lexical content and nothing else. With τ the CLIP-BPE tokenizer, σ the surface-form signature of a token (capitalization, number, possessive marking), $\mathcal{N}$ the set of phonotactically generated nonce strings with no intended lexical meaning, and $\sqsubseteq$ the substring relation,

$$\mathcal{R}(w) := \{\, r \in \mathcal{N} : \ |\tau(r)| = |\tau(w)|,\ \sigma(r) = \sigma(w),\ r \not\sqsubseteq w,\ w \not\sqsubseteq r \,\}. \tag{3}$$

A deterministic generator picks one $r_j \in \mathcal{R}(w_j)$ from the lemma of w_j (Appendix B.5). The constraints hold three confounders fixed: token count (positions and attention masks do not move), surface form (morphology does not change), and lexical meaning (no competing concept enters); we call their joint satisfaction the *Three-Invariance Replacement Protocol*. On D_{align70} it yields an admissible intervention for 99.79% of captions, 2.81 per caption on average (Appendix B.7).

4.3 What the score estimates

Two results state what Δ_j and μ measure; the first is local and analytic, the second combinatorial and exact. For the first, let $\mathbf{e}_w \in \mathbb{R}^m$ be the input embedding block of a token, assume it decomposes as $\mathbf{e}_w = \boldsymbol{\phi}(\sigma(w)) + \mathbf{c}_w$ into a surface-form and a lexical-content component, with a nonce carrying content $\mathbf{c}_r = \boldsymbol{\varepsilon}_r$ that has no target-specific meaning; let $\mathbf{u} = f_I(I)$, and let $g(\mathbf{e}) = \langle \mathbf{u}, f_T(T[\mathbf{e}_{w_j} \leftarrow \mathbf{e}]) \rangle$ be twice differentiable on the segment from $\mathbf{e}_{w_j}$ to $\mathbf{e}_{r_j}$ with Hessian norm at most M_j.

Proposition 2 (First-order sensitivity under Three-Invariance). *For $r_j \in \mathcal{R}(w_j)$, with J_j the Jacobian of f_T with respect to the block of w_j at T,*

$$\Delta_j = \underbrace{\langle \mathbf{u}, J_j \mathbf{c}_{w_j} \rangle}_{\textit{content sensitivity}} - \underbrace{\langle \mathbf{u}, J_j \boldsymbol{\varepsilon}_{r_j} \rangle}_{\textit{nonce artifact}} - R_j, \qquad |R_j| \leq \tfrac{1}{2} M_j \|\mathbf{e}_{r_j} - \mathbf{e}_{w_j}\|^2. \tag{4}$$

Relaxing token-count invariance adds a positional term, relaxing surface-form invariance adds $\langle \mathbf{u}, J_j(\boldsymbol{\phi}(\sigma(r_j)) - \boldsymbol{\phi}(\sigma(w_j))) \rangle$, and a real-word r_j replaces $\boldsymbol{\varepsilon}_{r_j}$ by its content vector; none of the three need be zero-mean, so none can be assumed to average out over phrases.

That is, Δ_j is the first-order response of the fixed scorer to removing the lexical content of one head, up to a nonce artifact and a second-order remainder. The artifact is bounded empirically, not assumed away: text-only displacement correlates negatively with Δ_j ($r = -0.215$), and nonce drops lie strictly between same- and cross-category real-word replacements (Section 6). Nothing in Equation 4 localizes w_j in the image; PAS is a sensitivity, not a grounding.

The second result says why the *mean* is the right aggregate. For $S \subseteq [K]$ let T_S be the caption with the heads of all phrases outside S replaced by their nonces, $v(S) := s(I, T_S)$, so $v([K]) = s(I,T)$ and $v(\emptyset)$ is the similarity of the fully hollowed caption; let $d(R)$ be the Harsanyi dividends of v and φ_j its Shapley value (Shapley, 1953).

Proposition 3 (PAS is the first-order Shapley term). *$\Delta_j = v([K]) - v([K] \setminus \{j\})$, and*

$$\Delta_j = \varphi_j + \sum_{R \ni j,\, |R| \geq 2} \Big(1 - \tfrac{1}{|R|}\Big) d(R), \qquad \mu = \frac{v([K]) - v(\emptyset)}{K} + \frac{1}{K}\sum_{|R| \geq 2} \big(|R| - 1\big) d(R). \tag{5}$$

If phrases do not interact ($d(R) = 0$ for $|R| \geq 2$), then $\Delta_j = \varphi_j$ and μ is the per-phrase share of similarity that lexical content carries over the fully hollowed caption.

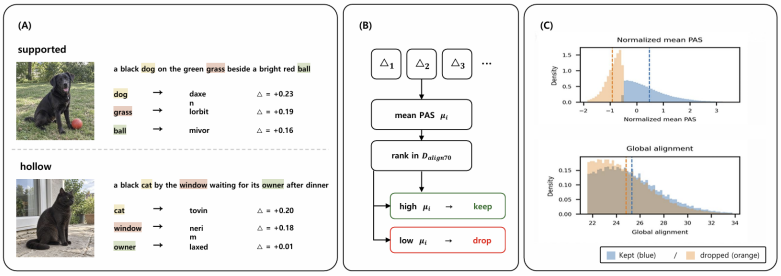


Figure 2: **Stage 2 scoring. (A)** Head-to-nonce interventions on a phrase-supported caption (each removal lowers similarity) and a hollow caption (removals barely move it). **(B)** Drops are averaged into μ_i and ranked within D_{align70}. **(C)** Kept (blue) and dropped (orange) pairs separate on normalized mean PAS ($+0.48$ vs. -0.92) but overlap on global alignment (0.253 vs. 0.248), so Stage 2 uses information that Stage 1 does not.

Concretely, a caption is hollow when $v(\emptyset) \approx v([K])$. Mean PAS estimates $(v([K]) - v(\emptyset))/K$ from K interventions rather than 2^K coalitions, at the price of the interaction term, which we treat as a residual and test indirectly (Section 6).

4.4 Selection rule and pipeline

The CPI score of pair i is its mean PAS. We rank the valid Stage-1 samples by this score and retain the top-scoring samples required to reach 50% retention relative to the original corpus:

$$\mu_i = \frac{1}{K_i} \sum_j \Delta_{ij}; \quad D \xrightarrow{s \geq s_{(0.7)}} D_{\text{align70}} \xrightarrow{\text{top-}N \text{ by } \mu} D_{\text{ours}} = \underset{S \subseteq D_{\text{align70}},\, |S|=N}{\arg\max} \sum_{i \in S} \mu_i. \tag{6}$$

The scoring pipeline additionally stores a normalized version of mean PAS for analysis and reporting; this normalization preserves the induced ranking. Thus, the selection program is solved by sorting the valid Stage-1 samples by mean PAS. On CC3M, $|D| = 2{,}222{,}261$, $|D_{\text{align70}}| = 1{,}557{,}650$, and $|D_{\text{ours}}| = 1{,}111{,}131$. Of the D_{align70} samples, 0.21% admit no valid intervention and are excluded. Stage 2 then removes approximately the bottom 28.5% of the remaining valid pool by mean PAS (Appendix C). Both stages use one OpenAI CLIP ViT-B/16 scorer, so no cross-encoder calibration enters; image embeddings are cached once, each intervention costs one text forward pass, and scoring CC3M takes 1.70 GPU-hours (Appendix L).

Spans come from a dependency parser in four syntactic contexts: noun phrase, compact noun phrase, noun phrase with a spatial preposition, and participial predicate with a nominal argument. In every context the replaced token is the nominal head; the contexts differ in what surrounds it. The restriction is deliberate: the invariances in Equation 3 are defined for nominal tokens, which keeps Proposition 2 well specified. Whether a nominal-only signal carries relational information is answered empirically: rankings from relational and from plain noun-phrase contexts are nearly independent ($\rho = 0.059$; Section 6). Figure 2 shows the operator on two captions and the kept-versus-dropped distributions. The subset is then used unchanged, with no change to objective, optimizer, or architecture, in every run that follows (extraction rules, nonce generation, and skip diagnostics in Appendix B; aggregation ablations in Section 6).

5 Experiments

Each subsection states the claim it tests, the setting, and the cells that decide it. All ViT-B/32 CLIP runs continue pretraining from one ViT-B/32 checkpoint (pretrained on Full for 13,575 steps) for 5,430 steps with identical optimizer, schedule, batch size, and seeds; only the training identities differ, so the comparison isolates the effect of continued training on selected data. We evaluate zero-shot classification (ImageNet-1K, CIFAR10/100), Flickr30K retrieval, linear probes, SugarCrepe, SugarCrepe++, and the Visual Genome split of VL-CheckList. Values are means $\pm$ standard deviations over matched seeds; *CPI50* names the Stage-2 subset (Appendix D).

5.1 Stage 2 adds information that alignment cannot supply

Claim. At a fixed 50% budget, ranking the aligned pool by mean PAS improves relation- and presence-sensitive evaluation relative to ranking it by alignment, without a cost in general transfer. *Setting.* Five paired seeds; four conditions: Full (100%), Full-Random50, Align-only top50, and CPI50 (Table 1).

Table 1: Main result on CC3M, ViT-B/32 (mean ± standard deviation over five seeds). Top block: general transfer; Flickr columns average T2I and I2T, CIFAR ZS averages CIFAR10 and CIFAR100. SC denotes SugarCrepe and SC++ SugarCrepe++. SC++ Pooled is the example-weighted accuracy over the five subsets and SC++ Macro the unweighted subset mean; the benchmark defines no official overall score. CPI50 is highest on SC Overall, SC Add, SC++ Replace, and VLC Object, Relation, and Overall; Align-only top50 is highest on SC Replace, both Swap rows, both SC++ aggregates, and VLC Attribute. CPI50 is highest on Flickr R@5 and CIFAR ZS and within one standard deviation of the best condition on ImageNet and Flickr R@1. Subset-level, direction-specific retrieval, linear-probe, and paired-seed results are in Appendices E and D.

Metric	Full	Full-Random50	Align-only top50	CPI50
IN-1K ZS	7.40±0.12	6.99±0.09	**7.43±0.12**	7.37±0.08
Flickr R@1 (avg)	**12.29±0.15**	11.80±0.25	12.03±0.38	12.13±0.32
Flickr R@5 (avg)	30.08±0.42	29.19±0.23	30.45±0.31	**30.65±0.31**
CIFAR ZS (avg)	29.73±0.28	29.73±0.40	28.22±0.39	**30.02±0.07**
SC Overall	57.72±0.22	57.82±0.33	58.21±0.26	**58.61±0.23**
SC Add Avg	59.61±0.30	59.76±0.27	59.35±0.57	**60.83±0.64**
SC Replace Avg	58.14±0.24	58.26±0.41	**59.13±0.55**	58.90±0.06
SC Swap Attribute	52.04±0.36	52.46±1.56	**53.96±1.45**	52.88±0.78
SC Swap Object	51.43±1.55	51.43±2.33	**53.31±1.57**	51.59±0.68
SC++ Replace Avg	47.07±0.19	46.06±0.38	47.46±0.23	**47.52±0.12**
SC++ Pooled	45.42±0.22	44.39±0.14	**45.92±0.23**	45.59±0.10
SC++ Macro	42.63±0.28	41.87±0.30	**43.09±0.29**	42.48±0.33
VLC Object	70.84±0.20	71.12±0.04	70.79±0.04	**71.18±0.11**
VLC Attribute	56.93±0.28	56.96±0.24	**57.47±0.03**	57.38±0.28
VLC Relation	37.90±0.29	39.42±0.45	38.65±0.25	**39.63±0.23**
VLC Overall	66.68±0.19	67.08±0.02	66.74±0.05	**67.17±0.10**

The deciding cells are the VL-CheckList Relation and SugarCrepe Overall rows of Table 1. Relation rises from 37.90±0.29 (Full) and 38.65±0.25 (Align-only top50) to 39.63±0.23: +1.73 and +0.98. SugarCrepe Overall rises from 57.72±0.22 and 58.21±0.26 to 58.61±0.23: +0.89 and +0.40. Seeds are matched across conditions, so the gains over Align-only top50 admit paired tests: CPI50 is higher in all five seeds on both metrics (one-sided sign test $p = 0.031$), $t(4) = 7.60$, $p = 0.002$ (Relation) and $t(4) = 4.58$, $p = 0.010$ (SugarCrepe), with 95% intervals $[+0.62, +1.34]$ and $[+0.16, +0.65]$ (Appendix E.4).

The table is not uniform, and we read it whole: CPI50 is highest on six of the twelve compositional rows and Align-only top50 on the other six (listed in the caption). The split follows the mechanism: CPI attributes similarity to the presence of content words, so it moves the axes that test whether a named object or relational argument is in the image (Add Object +1.99 over Align-only, Relation +0.98), whereas swap probes permute a fixed token multiset, which a presence-based attribution cannot see; Section 5.4 shows an order-sensitive objective supplies what selection cannot. General transfer is preserved (top block of Table 1; Appendices E and D).

5.2 Matched controls and a competitive filter

Claim. The Stage-2 rule, not the Stage-1 pool, produces the gain, and the rule is competitive with T-MARS given the same pool. *Setting.* Two seeds; every condition except Full and Align-only top50 starts from the same D_{align70} pool (Table 18, Appendix F).

Random-from-Align70 is the cleanest control in the paper: same pool, same budget, random Stage 2. Against it CPI adds +0.59 SugarCrepe Overall, +0.93 Relation, and +0.20 ImageNet, isolating the ranking rule. Against the pool itself, CPI is within seed variance of Align70 on SugarCrepe Overall (58.49±0.40 vs. 58.35±0.01) and Relation (39.62±0.12 vs. 39.31±0.41) with 28.7% fewer samples, whereas alignment or random pruning of the same pool to the same size loses 0.36–0.45 SugarCrepe

and 0.62–0.69 Relation; Align70 stays ahead on ImageNet and SugarCrepe++. Given the identical pool and budget, CPI matches or exceeds T-MARS on every compositional column (SugarCrepe 58.49 vs. 58.25, SugarCrepe++ 45.60 vs. 45.59, Relation 39.62 vs. 39.20) while T-MARS leads on ImageNet (7.63 vs. 7.43): the two filters remove different pairs (similarity carried by rendered text versus similarity not carried by content words) and trade general against compositional transfer in opposite directions. CC3M contains little rendered text, so this tests T-MARS as a Stage-2 rule rather than in its intended regime (Appendix F).

5.3 Scaling across corpus and backbone

Claim. The relation-side effect is not specific to CC3M or to ViT-B/32. *Setting.* CC12M with three seeds at the same step budget (Table 2); ViT-S/32 (three seeds) and a pretrained ViT-L/14 (two seeds) on CC3M at a matched 50% budget (Table 23, Appendix G).

Table 2: CC12M at a fixed 5,430-step schedule (mean ± standard deviation over three seeds). Align70 is the Stage-1 pool; Align50, Random50, and CPI50 share a 50% budget. CPI50 is highest on both Flickr directions, SC++ Both, and VLC Relation; Full is highest on SC Overall and VLC Overall and Align70 on ImageNet. Extended results are in Appendix G.1.

Metric	Full	Align70	Align50	Random50	CPI50 (Ours)
IN-1K	7.38±0.23	**7.63±0.04**	7.59±0.08	7.38±0.06	7.54±0.02
Flickr T2I R@1	9.34±0.30	9.41±0.08	9.29±0.12	9.26±0.13	**9.79±0.27**
Flickr I2T R@1	12.83±0.40	13.10±0.92	12.77±0.72	13.07±0.49	**13.90±0.46**
SugarCrepe Overall	**61.61±0.52**	61.03±0.40	61.06±0.16	61.13±0.25	61.43±0.08
SC++ Both	47.27±0.45	47.56±0.65	47.47±0.44	46.91±0.39	**48.13±0.19**
VLC Relation	39.47±0.38	39.30±0.75	39.65±0.58	39.62±1.06	**40.05±0.31**
VLC Overall	**68.04±0.11**	67.42±0.33	67.23±0.09	67.98±0.16	67.26±0.11

On CC12M the selection signal replicates before training (coverage 99.68% vs. 99.79%; $\mathrm{corr}(s, \mu) = 0.079$ vs. 0.084), and after training CPI50 is highest on Relation (40.05 vs. 39.47 Full, 39.65 Align50), on SugarCrepe++ Both (48.13 vs. 47.27, 47.47), and on both Flickr directions; Full keeps SugarCrepe Overall and VL-CheckList Overall, Align70 keeps ImageNet. ViT-L/14 with CPI50 is higher than Align50 on all six columns of Table 23, including SugarCrepe++ Both (70.59 vs. 69.96, all five subsets improved) and Relation (54.87 vs. 53.91); ViT-S/32 shows the same relation and retrieval pattern. The strict SugarCrepe++ metric, behind at CC3M, is thus ahead at CC12M and on ViT-L/14 (Appendix G).

5.4 The subset compounds with objective-side methods

Claim. Phrase-supported positives and hard negatives are different signals. *Setting.* NegCLIP and CE-CLIP trained with their own objectives on Full versus CPI50, nothing else changed (two seeds, Table 3).

Table 3: Plug-in with objective-side methods (mean ± standard deviation over two seeds); only the training subset changes, and bold compares Full with CPI50 within each objective. Every column improves for both objectives; VLC Relation moves by +3.84 (NegCLIP) and +4.92 (CE-CLIP). Controls with Align70, Align50, and Random50 are in Appendix H.

Objective	Setting	IN-1K	SC Ovr	SC++ Ovr	VLC Rel	VLC Ovr
NegCLIP	Full	7.05 ± 0.06	58.91 ± 0.10	43.46 ± 0.07	37.72 ± 1.13	67.46 ± 0.23
NegCLIP	**CPI50**	**7.15 ± 0.06**	**59.53 ± 0.62**	**43.76 ± 0.22**	**41.56 ± 0.23**	**68.49 ± 0.14**
CE-CLIP	Full	5.41 ± 0.10	60.20 ± 0.48	35.46 ± 0.95	32.72 ± 0.09	70.24 ± 0.41
CE-CLIP	**CPI50**	**5.56 ± 0.06**	**61.07 ± 0.43**	**36.60 ± 0.04**	**37.64 ± 3.38**	**70.72 ± 0.47**

Replacing the training data alone improves every column for both objectives; the largest movements are on Relation, NegCLIP 37.72 → 41.56 (+3.84) and CE-CLIP 32.72 → 37.64 (+4.92, seed variance ±3.38). Against each objective's matched Align50 control, NegCLIP with CPI data leads on

Relation (+1.21) and VL-CheckList Overall but not SugarCrepe++, and CE-CLIP leads on Relation (+3.63) but not the SugarCrepe aggregates (Appendix H). The word-order result closes the loop with Section 5.1: on the same subset, moving from CLIP to NegCLIP raises swap-object from 51.84 to 56.12. Selection supplies positives whose content words carry the match; the objective supplies the order sensitivity selection cannot (Appendix H.1).

6 What the Score Measures

Not alignment re-ranked. Within D_{align70}, normalized mean PAS correlates with alignment at Pearson $r = 0.084$ and Spearman $\rho = 0.088$; kept and dropped pairs separate on normalized mean PAS ($+0.479$ vs. -0.915) but barely on alignment (0.2531 vs. 0.2484; Figure 2C), as Lemma 1 predicts (Appendix J).

Not a text-side artifact. On 50K interventions, Δ_j correlates negatively with text-only embedding displacement ($r = -0.215$, $\rho = -0.204$). On 100K interventions, nonce drops (0.0227) lie strictly between same-category real-word replacements (0.0177) and cross-category distractors (0.0257), $p < 10^{-16}$ for both paired gaps; a generic out-of-vocabulary effect would put nonce at or above the distractor (Appendices J.2 and J.3).

Not object salience, and not one phrase. On the 309,455 captions that admit both a plain noun-phrase and a relational or predicate intervention, the two context-specific rankings are nearly independent (Spearman $\rho = 0.059$; top-50% overlap 52.9%, chance $\approx 50\%$), which rules out a pure object-replacement reading. In a matched comparison ($\approx$ 2.0M captions, 800 bins on intervention count, alignment, and length), CPI-high captions have higher second, third, and minimum PAS ($+0.052$, $+0.035$, $+0.043$) and a 9.2-point lower top-1 share: sensitivity is distributed across phrases, the condition under which the interaction term of Proposition 3 is least likely to dominate (Appendix J.7).

Not one scorer's bias. Re-scoring D_{align70} with SigLIP ViT-B/16 gives a ranking that differs from CLIP's in detail (Kendall $\tau = 0.385$) yet selects 80.6% of the same pairs (71.3% at chance, since either rule keeps 71.3% of the pool), and training on the SigLIP subset gives Relation 39.16 ± 0.06, above Align50 (38.62), retaining $+1.45$ of the $+1.91$ two-seed gain over Full. Two scorers with different data and losses thus agree on which captions are hollow (Appendix J.8).

Not the implementation, and not verbosity. A different parser keeps 94.76% of the selection and changes reported metrics by at most 0.10 points; other target, nonce, and BPE-matching variants change 3–12% of the selection, with the largest drop (-0.93 SugarCrepe++ Both) from removing length matching. The mean is the best functional of the profile, and caption length and phrase count trail CPI on ImageNet, Flickr R@1, SugarCrepe Overall, and Relation (Appendices I–N).

7 Conclusion and Limitations

A pair-level filter thresholds one number, and that number is the sum of a component that survives removal of the caption's content and one that does not (Lemma 1); on CC3M the threshold saturates at the 70% pool. CPI measures the second component with K matched nonce interventions per caption (Propositions 2–3). Ranking by it yields a 50% subset that improves VL-CheckList Relation by $+1.73$ over full data and $+0.98$ over alignment-only filtering in all five paired seeds ($p = 0.002$), stays within seed variance of the 70% pool on compositional metrics with 28.7% fewer samples, matches or exceeds T-MARS given the same pool, and preserves general transfer; the direction holds on CC12M, three backbones, and under SigLIP, and grows to $+3.84$ and $+4.92$ under NegCLIP and CE-CLIP, at a curation cost of 1.70 GPU-hours. Subset indices and code are released.

Limitations. CPI intervenes on nominal heads and therefore measures the presence of content, not word order, which it leaves to objectives such as NegCLIP; the evidence stops at CC12M with fixed budgets, and from-scratch web-scale pretraining is open (Appendix L).

More broadly, pair-level filtering misses the share of similarity carried by the caption's own words, which CPI measures with K text forward passes.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing of the manuscript, including improving the grammar, clarity, readability, and phrasing of author-written text, as well as to assist with the preparation of selected figure elements. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure that they accurately reflect the intended scientific meaning. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies data curation using existing image–text datasets and does not involve human-subject experiments or the collection of new personal data. As with other web-scale vision–language datasets, the source data may contain societal biases, inappropriate content, or material subject to copyright and licensing restrictions. CPI is designed as a data-selection method for improving compositional supervision and is not intended to serve as a content-safety, fairness, or copyright filter. Consequently, applying CPI does not by itself remove these risks, and downstream users should follow the usage conditions of the underlying datasets and consider appropriate dataset auditing and filtering for their intended applications.

REPRODUCIBILITY STATEMENT

The phrase-extraction and intervention protocol, target-token rule, nonce constraints, and skip diagnostics are specified in Appendix B; pool and subset statistics in Appendix C; training and evaluation settings, run counts, and reporting conventions in Appendix D; proofs in Appendix K; and curation cost in Appendix L. The supplementary material contains the scoring and subset-construction code, configuration examples, and reference subset indices for the primary CC3M data-selection conditions.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. In *Workshop on Understanding Foundation Models at ICLR 2023*, 2023.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is Winoground hard? Investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

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 (NeurIPS) Datasets and Benchmarks Track*, 2024.

Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 2020.

Talfan Evans, Shreya Pathak, Hamza Merzic, Jonathan Richard Schwarz, Ryutaro Tanno, and Olivier J. Hénaff. Bad students make great teachers: Active learning accelerates large-scale visual understanding. In *European Conference on Computer Vision (ECCV)*, 2024.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations (ICLR)*, 2024.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2), 2021.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. Scaling laws for data filtering—data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning (ICML)*, 2019.

Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

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 (NeurIPS) Datasets and Benchmarks Track*, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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 (CVPR)*, 2023.

Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, and Ari S. Morcos. Sieve: Multimodal dataset pruning using image captioning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Pratyush Maini, Sachin Goyal, Zachary C. Lipton, J. Zico Kolter, and Aditi Raghunathan. T-MARS: Improving visual representations by circumventing text feature learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sören Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning (ICML)*, 2022.

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

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained VLMs for improving vision-linguistic compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Tuomas Oikarinen and Tsui-Wei Weng. CLIP-Dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations (ICLR)*, 2023.

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 the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

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

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, and Volker Tresp. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*. Princeton University Press, 1953.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

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 (CVPR)*, 2022.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Vishaal Udandarao, Mehdi Cherti, Shyamgopal Karthik, Jenia Jitsev, Samuel Albanie, and Matthias Bethge. A good CREPE needs more than just sugar: Investigating biases in compositional vision-language benchmarks. *arXiv preprint arXiv:2506.08227*, 2025a.

Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J. Hénaff. Active data curation effectively distills large-scale multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations (ICLR)*, 2024.

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 (ICLR)*, 2023.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

APPENDIX ROADMAP

Each main-text claim is backed by one appendix section; the table lists the map.

Main-text claim	Appendix
Alignment sweep and saturation point (Sec. 3)	A
Intervention operator, extraction rules, skip diagnostics (Sec. 4.2)	B
Pool and subset statistics (Sec. 4.4)	C
Training, evaluation, reporting conventions, run counts (Sec. 5)	D
Five-seed subset-level results and paired tests (Sec. 5.1)	E
Matched controls and T-MARS (Sec. 5.2)	F
CC12M, ViT-S/32, ViT-L/14 (Sec. 5.3)	G
NegCLIP and CE-CLIP controls, routed diagnostic (Sec. 5.4)	H
Aggregation, drop-ratio, and top-3 ablations (Sec. 6)	I
Non-redundancy, calibration, construct validity, cross-scorer, protocol robustness (Sec. 6)	J
Proofs of Lemma 1 and Propositions 2–3	K
Curation cost (Sec. 4.4)	L
Extended discussion and limitations (Sec. 7)	M
Qualitative disagreement cases	N

A ALIGNMENT PRUNING SWEEP

Table 4 reports the single-seed alignment-only sweep used in Section 3 to locate the saturation point and fix the Stage-1 operating point at top70. The columns are the same benchmarks as in the main comparison; no repeated-seed claim is made from this table.

Table 4: Alignment-only filtering sweep on CC3M (single seed). All runs share architecture, initialization, optimizer, and step budget. Bold marks the best per column. Every subset beats Full on SugarCrepe Overall, but no column improves monotonically with a stricter threshold: SugarCrepe Overall, SugarCrepe++ Replace, and Flickr T2I peak at top70 and fall thereafter.

Method	Keep	IN-1K	F-T2I R@1	F-I2T R@1	C10 ZS	C100 ZS	SC R/A	SC Ovr	SC++ Rep	SC++ Ovr
Full	100%	7.413	10.260	14.05	45.685	14.100	58.63	57.69	46.81	45.32
top90	90%	7.492	10.700	13.500	44.880	**16.120**	58.81	57.86	47.46	45.64
top80	80%	7.458	10.700	**15.300**	**46.850**	14.950	58.99	57.82	47.06	45.36
top70	70%	7.558	**11.020**	14.200	45.730	16.070	**59.20**	**58.34**	**47.82**	**45.95**
top60	60%	**7.592**	10.560	13.500	43.740	14.760	59.43	58.10	46.92	45.72
top50	50%	7.477	10.225	13.100	42.490	14.100	58.70	57.92	46.65	45.69

B FULL PHRASE-PERTURBATION PROTOCOL

This appendix provides the implementation details behind the Stage 2 CPI procedure introduced in Section 4.4 and the Three-Invariance Replacement Protocol defined in Section 4.2. It specifies the phrase-span taxonomy, candidate filtering rules, non-overlapping selection procedure, target-token selection rule, nonce-replacement constraints, and perturbation skip diagnostics used to compute phrase attribution scores.

B.1 CANDIDATE PHRASE TYPES

Each caption is parsed with spaCy's dependency parser. From the parse, four candidate phrase types are extracted:

The four-type taxonomy is calibrated to the perturbation axes that compositional benchmarks evaluate: object-presence (`raw_np`, `compact_np`), spatial and inter-object relations (`relation_np`), and predicate-object structures (`predicate_np_fragment`). The set of admissible prepositions for `relation_np` is {with, in, on, at, under, over, behind, near, inside, outside, by}; we deliberately exclude *of*, whose construction is overwhelmingly possessive rather than spatial-relational.

Table 5: Candidate phrase types used by CPI.

Type	Definition	Example
`raw_np`	spaCy noun chunk in its entirety	"a black dog"
`compact_np`	nominal head with its left-modifier subtree	"black dog"
`relation_np`	noun phrase with a selected prepositional dependent	"dog on the grass"
`predicate_np_fragment`	VBG/VBN predicate subtree containing a nominal anchor	"riding a horse"

B.2 Candidate Filtering Rules

Raw candidate spans frequently overlap and vary in linguistic quality. We apply a sequence of filtering rules—removing spans with no alphabetic content, exceeding eight content words or sixty characters, occupying more than 60% of the caption, dominated by function words, beginning with adposition or conjunction, headed by an adposition, comprising bare *be*-fragments, or lacking a nominal anchor. Surviving candidates are ranked by a heuristic quality score:

$$q(p) \;=\; \alpha_{\text{type}}(p) + \beta \cdot \text{cw_ratio}(p) - \gamma \cdot \text{length_penalty}(p), \tag{7}$$

with type-priority bonuses α_{type}: `relation_np` 3.0, `compact_np` 2.4, `predicate_np_fragment` 1.8, `raw_np` 1.2. This score is used only to rank overlapping candidate spans before perturbation; the final sample-level CPI score is computed solely from the resulting PAS values.

B.3 Non-overlapping Selection

A greedy non-overlapping selection traverses candidates in decreasing $q(p)$ order, accepting a candidate only if it shares no token with previously accepted spans. The result is a set of mutually disjoint, high-quality spans per caption.

```
candidates <- all surviving candidates
ranked <- sort candidates by q(p) descending
selected <- []
for c in ranked:
    if no token of c overlaps any token of any s in selected:
        selected.append(c)
return selected
```

B.4 Target Token Selection

For each accepted span, we identify a single nominal target token w_{ij}. For `relation_np` and `predicate_np_fragment`, the target is the rightmost noun or proper noun within the span—a choice grounded in the linguistic observation that relation and predicate fragments place their primary semantic load on the rightmost nominal. For `raw_np` and `compact_np`, the target is the span's syntactic root if nominal, otherwise the rightmost noun.

B.5 Three-Invariance Replacement Constraints

The replacement r_{ij} is generated deterministically from the lemma of w_{ij} via an onset-vowel-coda combinator that traverses a fixed phonotactic table. To implement the Three-Invariance Replacement Protocol, the generated candidate must satisfy the following constraints:

1. **Identity exclusion**: $r_{ij} \neq w_{ij}$.
2. **Alphabetic validity**: r_{ij} contains a sufficient density of alphabetic characters to constitute a valid pseudo-word.
3. **Substring exclusion**: neither r_{ij} is a substring of w_{ij} nor vice versa, ruling out spelling-proximity bias.
4. **Capitalization preservation**: the surface capitalization pattern of w_{ij} is mirrored.
5. **Inflection preservation**: plural and possessive surface forms are reproduced.

6. **Subtoken-count invariance**: $\#\text{subtok}_{\text{CLIP}}(r_{ij}) = \#\text{subtok}_{\text{CLIP}}(w_{ij})$ exactly.
7. **Numeric exclusion**: w_{ij} containing digits is skipped.

These constraints target distinct sources of bias in the PAS estimate. Subtoken-count invariance controls changes in OpenAI CLIP-BPE sequence length and positional structure. Lexico-semantic nullness prevents the replacement from acting as a meaningful synonym, hypernym, or visually grounded alternative to the original target. Surface-form preservation prevents capitalization, number, or possessive changes from becoming the dominant source of the similarity drop. Thus, the replacement is designed to remove lexical content while keeping tokenization and surface form as fixed as possible.

If no admissible r_{ij} is found, or if the resulting perturbed caption coincides with another already generated for the same sample or is judged ill-formed (e.g., disproportionately shortened), the candidate is skipped. We record skip reasons for diagnostic purposes.

B.6 Perturbation Examples

Table 6 shows representative perturbations produced by the replacement protocol. The examples illustrate that CPI changes the nominal target while preserving the surrounding phrase context and surface form.

Table 6: Representative controlled phrase perturbations.

Original phrase	Target	Replacement	Perturbed phrase
`a black dog`	`dog`	`[nonce]`	`a black [nonce]`
`dogs on the grass`	`grass`	`[nonce]`	`dogs on the [nonce]`
`riding a horse`	`horse`	`[nonce]`	`riding a [nonce]`
`John's bicycle`	`bicycle`	`[nonce]`	`John's [nonce]`

B.7 Skipped-Perturbation Reasons

Table 7: Skipped perturbation reasons. Percentages are computed over skipped perturbation rows.

Skip reason	Count	Percentage
`no_exact_subtoken_match_replacement`	113,775	94.84%
`target_has_digit`	5,591	4.66%
`no_target_token`	594	0.50%

The dominant source of skipped perturbations is the strict exact CLIP-BPE subtoken-count constraint. We keep this constraint because it removes a direct tokenization-length confound from the PAS computation. Candidates for which no exact-match nonce replacement can be found are skipped rather than forced. Skips are not uniformly distributed: heads that tokenize into many subtokens, typically long or rare nouns, are the ones most often left without an admissible nonce, so captions built around such nouns receive fewer interventions. Skipped rows are 2.67% of all candidate interventions (119,960 of 4,494,680), and mean-PAS aggregation makes a caption's score independent of how many of its candidates survive, so this bias affects which phrases are scored rather than how captions are ranked.

C Filtered Subset Statistics

Within the valid scored D_{align70} pool, CPI-kept captions are slightly longer on average and contain more phrases per caption than CPI-dropped captions, consistent with the intuition that captions with richer phrase structure provide more substantive compositional supervision.

The 99.79% coverage figure indicates that the strict replacement constraints still leave at least one admissible phrase intervention for nearly all processed samples in D_{align70}. The average valid sample admits 2.81 valid phrase interventions, providing sufficient coverage for mean-PAS aggregation.

Table 8: Dataset variant statistics. CPI-kept and CPI-dropped are computed within the valid scored D_{align70} pool used for Stage 2 selection.

Variant	Samples	Mean caption length	Mean # phrases	Valid-perturb coverage
Full	2,222,261	11.4	2.81	–
D_{align70}	1,557,650	11.6	2.83	99.79%
CPI-kept	1,111,131	11.8	2.97	–
CPI-dropped	443,189	11.1	2.50	–

Table 9: Perturbation coverage on D_{align70}.

Statistic	Value
Samples processed	1,557,650
Samples with ≥ 1 valid perturbation	1,554,320 (99.79%)
Samples with no valid perturbation	3,330 (0.21%)
Total valid perturbation rows	4,374,720
Skipped perturbation rows	119,960
Mean perturbations per valid sample	2.81
Median perturbations per valid sample	3.00
Maximum perturbations per valid sample	14

This coverage is measured after all parser- and perturbation-level validity checks. Invalid or unsupported candidates are skipped rather than forced, making the rule-based extraction pipeline conservative rather than brittle: unreliable perturbations are removed while nearly all processed samples retain at least one admissible intervention.

D Training and Evaluation Details

Reporting convention. Table 10 summarizes the run count and reporting format used by each experimental block. All $\pm$ values denote standard deviation over training seeds, computed after forming the reported aggregate metric; they are not standard errors and are not variation across benchmark subsets.

Table 10: Reporting convention across experimental blocks.

Experimental block	Runs	Format
Table 4 alignment sweep	Single seed	No $\pm$ (exploratory)
Table 1 main CLIP comparison	Five seeds	Mean $\pm$ std.
Table 18 matched controls / T-MARS	Two seeds	Mean $\pm$ std.
Table 2 CC12M scaling	Three seeds	Mean $\pm$ std.
Table 23 ViT-S/32	Three seeds	Mean $\pm$ std.
Table 23 ViT-L/14	Two seeds	Mean $\pm$ std.
Table 3 NegCLIP	Two seeds	Mean $\pm$ std.
Table 3 CE-CLIP	Two seeds	Mean $\pm$ std.
Appendix I ablations	Single seed	No $\pm$ (diagnostic)
Appendix I.4 Top-3 aggregation	Two seeds	Mean $\pm$ std.
Appendix J.6 structural controls	Single seed	No $\pm$ (diagnostic)
Appendix J.8 cross-scorer robustness	Two seeds	Mean $\pm$ std.
Appendix J.9 protocol robustness	Single seed	No $\pm$ (diagnostic)

Unless otherwise stated, comparisons within the same experimental block use the same architecture, initialization, optimizer, learning rate, maximum training steps, and random seeds; only the training subset or selection rule varies. The five-seed main CLIP comparison is reported separately from the additional matched-control, filtering-baseline, and scaling experiments, whose run counts are specified in Table 10. Exploratory sweeps and most diagnostic ablations are single-seed analyses; exceptions and repeated-seed robustness experiments are explicitly identified in Table 10.

Table 11: Default training configuration for the CC3M CLIP ViT-B/32 experiments. Experiments using different corpus scales or visual backbones are documented separately in the corresponding scaling appendix.

Item	Value
Architecture	CLIP ViT-B/32
Initialization	Fine-tuned from a scratch-pretrained CLIP checkpoint at 13,575 steps
Initial checkpoint dataset	Full data
Projection dim	512
Image resolution	224×224
Maximum text length	77 tokens
Tokenizer	OpenAI CLIP BPE
Optimizer	AdamW
$\beta_1, \beta_2, \epsilon$	0.9, 0.98, 10^{-8}
Learning rate	1.5×10^{-4}
Weight decay	0.2
Batch size (CLIP)	2048
Batch size (NegCLIP / CE-CLIP)	1024
Gradient accumulation	1
Maximum training steps	5,430
Random seeds	Five seeds for the main CLIP comparison; three seeds for CC12M and ViT-S/32; two seeds for the matched-control/T-MARS, NegCLIP, CE-CLIP, ViT-L/14, and cross-scorer robustness comparisons; single seed for exploratory sweeps and diagnostic ablations
Hardware	$4 \times$ NVIDIA RTX PRO 6000 Blackwell GPUs, 96GB memory per GPU

Evaluation benchmarks. We evaluate zero-shot classification on ImageNet-1K, CIFAR10, and CIFAR100; image–text retrieval on the Flickr30K Karpathy 1K test split using T2I and I2T R@$\{1, 5, 10\}$; linear-probe transfer on CIFAR10, CIFAR100, EuroSAT, and DTD; and compositional reasoning on SugarCrepe, SugarCrepe++, and VL-CheckList-VG. For VL-CheckList, we report the Visual Genome subset, which provides object, attribute, and relation probes with full image coverage in our setup.

D.1 Full Retrieval Results for the Main Comparison

Table 12 reports the complete five-seed Flickr30K retrieval results for the main CLIP curation comparison. The main text reports compact averages over the T2I and I2T directions in Table 1; the direction-specific R@1, R@5, and R@10 results are included here for completeness.

Table 12: Full Flickr30K retrieval results on the Karpathy 1K test split for the five-seed main CLIP curation comparison. Values report mean $\pm$ standard deviation over five training seeds. Bold marks the best value in each column.

Method	T2I R@1	T2I R@5	T2I R@10	I2T R@1	I2T R@5	I2T R@10
Full	10.31 ± 0.07	26.12 ± 0.25	35.55 ± 0.19	$\mathbf{14.26 \pm 0.26}$	34.04 ± 0.67	43.98 ± 1.15
Full-Random50	9.93 ± 0.21	25.69 ± 0.17	35.26 ± 0.25	13.68 ± 0.50	32.68 ± 0.31	42.76 ± 0.43
Align-only top50	10.31 ± 0.12	$\mathbf{27.58 \pm 0.12}$	$\mathbf{36.81 \pm 0.23}$	13.74 ± 0.71	33.32 ± 0.56	$\mathbf{44.68 \pm 0.28}$
CPI50	$\mathbf{10.37 \pm 0.22}$	26.60 ± 0.45	36.12 ± 0.27	13.88 ± 0.79	$\mathbf{34.70 \pm 0.35}$	44.46 ± 1.01

D.2 Linear-Probe Results

The linear-probe evaluation was conducted as a separate two-seed evaluation block. We report it here for completeness, but do not merge its uncertainty estimates with the five-seed repeated comparison in Table 1. Table 13 reports the per-dataset linear-probe results and their average.

Table 13: Linear-probe transfer results for the CC3M ViT-B/32 comparison. Per-dataset values report mean ± standard deviation over two seeds. The LP average is the arithmetic mean across CIFAR10, CIFAR100, EuroSAT, and DTD.

Method	CIFAR10 LP	CIFAR100 LP	EuroSAT LP	DTD LP	LP Avg
Full	66.32 ± 0.25	34.20 ± 0.01	68.59 ± 0.37	20.08 ± 2.07	47.30
Full-Random50	65.91 ± 0.17	33.61 ± 0.33	68.59 ± 0.37	$\mathbf{21.44 \pm 0.15}$	47.39
Align-only top50	66.32 ± 0.17	34.47 ± 0.11	$\mathbf{68.67 \pm 0.26}$	19.97 ± 1.92	47.36
CPI50	$\mathbf{66.73 \pm 0.01}$	$\mathbf{35.13 \pm 0.13}$	68.59 ± 0.37	$\mathbf{21.44 \pm 0.15}$	**47.97**

E Full Results for the Main CC3M Comparison

This appendix provides the full five-seed compositional breakdowns corresponding to the main CC3M comparison in Table 1. All values in this section report mean ± standard deviation over the same five training seeds used in the main repeated comparison.

The compact SugarCrepe aggregates reported in Table 1 are computed at the evaluation-example level over the corresponding benchmark subsets and therefore need not equal the unweighted mean of the subset-level accuracies reported below.

E.1 SugarCrepe Subset Breakdown

Table 14: SugarCrepe subset-level results for the five-seed main CC3M comparison. Values report mean ± standard deviation over five training seeds. Bold marks the best value in each row.

Subset	Full	Full-Random50	Align-only top50	CPI50
add_att	61.13 ± 0.60	61.42 ± 0.42	61.24 ± 0.97	$\mathbf{62.23 \pm 1.08}$
add_obj	58.10 ± 0.31	58.10 ± 0.48	57.45 ± 0.29	$\mathbf{59.44 \pm 0.41}$
replace_att	60.15 ± 0.38	60.58 ± 0.40	$\mathbf{61.73 \pm 0.79}$	61.12 ± 0.65
replace_obj	65.36 ± 0.24	$\mathbf{65.53 \pm 0.80}$	65.34 ± 0.24	65.39 ± 0.32
replace_rel	48.90 ± 0.48	48.68 ± 0.64	$\mathbf{50.33 \pm 0.75}$	50.18 ± 0.42
swap_att	52.04 ± 0.36	52.46 ± 1.56	$\mathbf{53.96 \pm 1.45}$	52.88 ± 0.78
swap_obj	51.43 ± 1.55	51.43 ± 2.33	$\mathbf{53.31 \pm 1.57}$	51.59 ± 0.68

The subset-level results show that the aggregate SugarCrepe gain does not correspond to uniform improvement across all perturbation types. CPI50 is strongest on the two addition subsets, while Align-only top50 is stronger on several replacement and swap subsets. The main-text overall score therefore summarizes a mixed subset-level pattern rather than universal dominance across every SugarCrepe category.

E.2 SugarCrepe++ Subset Breakdown

Table 15: SugarCrepe++ subset-level `acc-both` results for the five-seed main CC3M comparison. Values report mean ± standard deviation over five training seeds. Bold marks the best value in each row.

Subset	Full	Full-Random50	Align-only top50	CPI50
replace_att	49.01 ± 0.58	48.53 ± 0.69	48.60 ± 0.62	$\mathbf{49.52 \pm 0.57}$
replace_obj	$\mathbf{57.05 \pm 0.47}$	55.85 ± 0.52	56.44 ± 0.45	56.54 ± 0.33
replace_rel	35.16 ± 0.24	33.80 ± 0.58	$\mathbf{37.33 \pm 0.52}$	36.51 ± 0.21
swap_att	38.32 ± 0.54	37.30 ± 1.37	$\mathbf{39.28 \pm 1.12}$	37.99 ± 0.74
swap_obj	33.63 ± 1.40	$\mathbf{33.88 \pm 1.44}$	33.80 ± 0.67	31.84 ± 1.29

SugarCrepe++ likewise exhibits a mixed subset-level pattern. CPI50 is strongest on replace-attribute, whereas other subsets favor Full, Full-Random50, or Align-only top50. This is consistent with the main-text result that CPI50 improves the replacement aggregate while remaining close to, but not uniformly above, alignment-only filtering on the stricter pooled and macro SugarCrepe++ measures.

E.3 VL-CheckList-VG Aspect Breakdown

Table 16: VL-CheckList-VG aspect-level results for the five-seed main CC3M comparison. Values report mean ± standard deviation over five training seeds. Bold marks the best value in each column.

Method	Overall	Object	Attribute	Relation
Full	66.68 ± 0.19	70.84 ± 0.20	56.93 ± 0.28	37.90 ± 0.29
Full-Random50	67.08 ± 0.02	71.12 ± 0.04	56.96 ± 0.24	39.42 ± 0.45
Align-only top50	66.74 ± 0.05	70.79 ± 0.04	$\mathbf{57.47 \pm 0.03}$	38.65 ± 0.25
CPI50	$\mathbf{67.17 \pm 0.10}$	$\mathbf{71.18 \pm 0.11}$	57.38 ± 0.28	$\mathbf{39.63 \pm 0.23}$

The clearest CPI gain appears on the relation axis: CPI50 improves VL-CheckList-VG Relation from 38.65 ± 0.25 for Align-only top50 to 39.63 ± 0.23, while also yielding the highest Object and Overall scores. Attribute performance is similar but slightly lower than Align-only top50.

E.4 Paired Five-Seed Analysis

Because the main comparison uses matched random seeds across curation conditions, we additionally perform paired two-sided t-tests between CPI50 and Align-only top50 on two representative metrics highlighted in the main text.

Table 17: Paired five-seed comparison of CPI50 against Align-only top50. Differences are computed as CPI50 minus Align-only top50. The confidence intervals are 95% intervals for the mean paired difference.

Metric	Mean diff.	$t(4)$	p	95% CI
SugarCrepe Overall	$+0.405$	4.58	0.010	$[+0.159, +0.650]$
VL-CheckList Relation	$+0.981$	7.60	0.0016	$[+0.622, +1.339]$

The paired analysis indicates that the improvements on SugarCrepe Overall and VL-CheckList-VG Relation are consistent across the five matched training seeds rather than being driven by a single run. We restrict the statistical claim to these preselected representative metrics and do not treat the full set of benchmark and subset comparisons as independent hypothesis tests.

F Matched Controls and Competitive Filtering Baselines

Table 18: Matched controls and a competitive filter on CC3M (mean ± standard deviation over two seeds, reported separately from the five-seed comparison above). Random-from-Align70, Align70→T-MARS50, and Align70→CPI50 share the Stage-1 pool; the latter two share the budget and differ only in the Stage-2 rule. CPI adds $+0.59$ SC Overall and $+0.93$ VLC Relation over the random Stage-2 control, matches or exceeds Align70 on both with 28.7% fewer samples, and matches or exceeds T-MARS on every compositional column given the same pool. Full breakdowns are in Appendix F.

Method	Keep	IN-1K	SC Ovr	SC++ Pooled	VLC Rel	VLC Ovr
Full	100%	7.27±0.07	57.78±0.09	45.53±0.30	37.71±0.43	66.51±0.11
Align-only top50	50%	7.30±0.00	57.99±0.43	45.75±0.31	38.62±0.02	66.73±0.01
Align70	70%	7.58±0.03	58.35±0.01	45.79±0.24	39.31±0.41	67.26±0.03
Random-from-Align70	50%	7.23±0.04	57.90±0.26	45.43±0.21	38.69±0.17	67.11±0.01
T-MARS50	50%	7.45±0.00	58.29±0.08	**46.13±0.05**	39.10±0.05	66.85±0.02
Align70→T-MARS50	50%	**7.63±0.05**	58.25±0.00	45.59±0.07	39.20±0.43	66.92±0.13
Align70→CPI50 (Ours)	50%	7.43±0.01	**58.49±0.40**	45.60±0.03	**39.62±0.12**	**67.27±0.05**

This appendix provides the full benchmark breakdowns for the additional matched controls and competitive filtering baselines summarized compactly in Table 18. All results in this section report mean ± standard deviation over the same two training seeds (seeds 0 and 1) used for the main-text control comparison.

The primary Full and Align-only top50 baselines are already reported in full under the five-seed main comparison in Appendix E. We therefore focus here on the additional controls introduced for the two-stage pipeline: Align70, Random-from-Align70, T-MARS50, Align70→T-MARS50, and Align70→CPI50.

Align70 evaluates the Stage 1 candidate pool before second-stage selection. Random-from-Align70 randomly reduces that same pool to the final 50% data budget. T-MARS50 applies T-MARS directly to the full corpus at a 50% budget, whereas Align70→T-MARS50 and Align70→CPI50 start from the same Align70 candidate pool and use the same final 50% budget, differing only in the Stage 2 ranking rule.

F.1 General Transfer and Retrieval

Table 19: Full general-transfer and Flickr30K retrieval results for the additional filtering controls and baselines on CC3M. Values report mean ± standard deviation over two training seeds. T2I and I2T denote text-to-image and image-to-text retrieval, respectively.

Method	IN-1K	T2I R@1	T2I R@5	T2I R@10	I2T R@1	I2T R@5	I2T R@10	C10	C100
Align70	7.577±0.027	10.890±0.184	27.240±0.141	37.130±0.240	14.150±0.071	34.750±0.778	44.850±0.071	45.680±0.071	16.215±0.205
Random-from-Align70	7.232±0.037	10.840±0.198	26.860±0.198	36.220±0.028	14.150±0.636	33.950±0.212	44.800±0.566	46.120±0.085	15.400±0.170
T-MARS50	7.447±0.001	10.710±0.071	27.470±0.382	36.890±0.127	13.350±0.354	34.050±0.354	45.300±0.424	43.765±0.078	14.940±0.085
Align70→T-MARS50	7.627±0.049	10.900±0.085	27.720±0.170	37.510±0.127	13.300±0.424	35.800±0.283	45.350±0.071	44.450±0.071	13.835±0.233
Align70→CPI50 (Ours)	7.427±0.013	10.200±0.283	27.030±0.184	36.120±0.170	14.400±0.283	34.850±0.636	45.550±0.212	44.760±0.028	15.135±0.007

The general-transfer comparison is mixed across metrics. Align70→T-MARS50 yields the strongest ImageNet and Flickr text-to-image results among the matched Stage 2 filters, whereas Align70→CPI50 yields the strongest image-to-text R@1 result. The purpose of the comparison is therefore not to establish uniform dominance on general benchmarks, but to determine whether the Stage 2 selection rule changes the downstream trade-off under a fixed candidate pool and data budget.

F.2 SugarCrepe Breakdown

Table 20: Full SugarCrepe subset breakdown for the additional filtering controls and baselines. Values report mean ± standard deviation over two training seeds. Overall is the sample-weighted accuracy over all 7,511 benchmark examples.

Method	Overall	Add Att	Add Obj	Rep. Att	Rep. Obj	Rep. Rel	Swap Att	Swap Obj
Align70	58.348±0.009	61.561±1.226	58.511±0.377	61.358±0.628	65.103±0.043	50.284±1.006	53.904±0.212	51.020±1.732
Random-from-Align70	57.902±0.264	61.272±0.409	58.317±0.583	60.723±0.090	65.103±0.471	48.755±0.050	53.604±0.637	51.429±0.000
T-MARS50	58.288±0.075	62.139±0.817	57.590±0.171	61.231±0.090	64.952±0.000	51.067±0.201	52.703±0.637	55.510±0.000
Align70→T-MARS50	58.248±0.000	60.910±0.102	58.608±0.171	61.421±0.897	65.042±0.214	50.142±0.302	53.228±0.106	51.837±1.154
Align70→CPI50 (Ours)	58.488±0.395	61.127±0.613	59.336±0.583	60.470±0.269	65.587±0.128	50.462±0.251	52.553±1.274	51.837±0.577

The subset-level comparison is heterogeneous. T-MARS50 is particularly strong on add-attribute and swap-object, while Align70→CPI50 is higher on add-object, replace-object, and SugarCrepe Overall than the matched Align70→T-MARS50 condition. These results show that the aggregate differences are not produced by uniform gains across every perturbation type.

F.3 SugarCrepe++ Breakdown

Table 21: Full SugarCrepe++ strict `acc-both` breakdown for the additional filtering controls and baselines. Values report mean ± standard deviation over two training seeds. Overall is the pooled example-level `acc-both` metric used in Table 18.

Method	Overall	Rep. Att	Rep. Obj	Rep. Rel	Swap Att	Swap Obj
Align70	45.785±0.238	49.302±0.090	56.992±0.043	36.771±0.704	37.387±0.212	33.469±1.732
Random-from-Align70	45.428±0.208	49.365±0.359	56.235±0.086	35.811±0.251	38.664±0.319	33.469±1.732
T-MARS50	46.132±0.045	49.556±0.987	56.356±0.514	37.340±0.101	39.264±0.319	35.306±0.289
Align70→T-MARS50	45.585±0.074	49.810±0.269	55.993±0.000	36.558±0.302	38.138±0.425	33.878±2.309
Align70→CPI50 (Ours)	45.596±0.030	49.429±0.987	56.780±0.257	36.629±0.302	37.763±0.531	30.612±0.577

T-MARS50 obtains the highest pooled SugarCrepe++ score in this block. Within the matched Align70 candidate pool, however, Align70→T-MARS50 and Align70→CPI50 have nearly identical pooled scores (45.585 versus 45.596), while their subset-level strengths differ.

F.4 VL-CheckList-VG Aspect Breakdown

Table 22: VL-CheckList-VG aspect-level results for the additional filtering controls and baselines. Values report mean ± standard deviation over two training seeds.

Method	Overall	Object	Attribute	Relation
Align70	67.260±0.033	71.305±0.020	57.715±0.130	39.307±0.406
Random-from-Align70	67.114±0.013	71.218±0.036	57.564±0.011	38.688±0.170
T-MARS50	66.854±0.023	70.868±0.021	57.413±0.224	39.101±0.049
Align70→T-MARS50	66.922±0.128	70.936±0.094	57.410±0.061	39.203±0.429
Align70→CPI50 (Ours)	67.273±0.045	71.292±0.038	57.533±0.011	39.617±0.117

Relative to the matched Align70→T-MARS50 condition, Align70→CPI50 is higher on Object, Attribute, Relation, and Overall VL-CheckList-VG accuracy. The largest difference appears on the Relation axis. Random-from-Align70 further shows that simply restricting the training population to the Align70 pool is not sufficient to recover the same relation performance.

Overall, the matched-control results separate three effects: the quality of the Stage 1 alignment-filtered pool, the effect of reducing that pool to a 50% training budget, and the choice of Stage 2 ranking rule. The additional T-MARS comparison further shows that different filtering rules induce different general/compositional trade-offs rather than one method uniformly dominating every metric.

G Scaling Experiments

This appendix provides the extended results for the data- and model-scaling experiments summarized in Section 5.3. The CC12M and ViT-S/32 comparisons use three training seeds, while the ViT-L/14 comparison uses two training seeds. Within each scaling block, the compared curation strategies use matched model, initialization, optimization, and evaluation settings.

Table 23: Backbone scaling on CC3M at a matched 50% budget (ViT-S/32: three seeds; ViT-L/14: two seeds). Within a backbone the two rows differ only in the selected data. With ViT-L/14, CPI50 is higher on all six columns; with ViT-S/32, on retrieval, SC Overall, and VLC Relation. Full results are in Appendices G.2 and G.3.

Backbone	Method	IN-1K	F-T2I R@1	F-I2T R@1	SC Ovr	SC++ Both	VLC Rel
ViT-S/32	Align50	**7.10±0.04**	9.95±0.23	13.20±0.72	58.52±0.18	**46.60±0.33**	38.46±0.79
ViT-S/32	CPI50	6.85±0.02	**10.03±0.05**	**13.83±0.60**	**58.97±0.15**	46.08±0.17	**39.21±0.23**
ViT-L/14	Align50	73.31±0.03	72.06±0.06	88.95±0.07	81.02±0.02	69.96±0.48	53.91±0.60
ViT-L/14	CPI50	**73.44±0.06**	**73.02±0.14**	**89.75±0.21**	**81.34±0.07**	**70.59±0.12**	**54.87±0.54**

G.1 CC12M Scaling

We evaluate data-scale robustness on CC12M using the same fixed 5,430-step training schedule. Before any training, the selection signal itself replicates at the larger scale: valid-intervention coverage is 99.68% (CC3M: 99.79%), the alignment–PAS correlation is $r = 0.079$ (0.084), and the normalized mean-PAS separation between kept and dropped pairs is $0.462/-0.898$ ($0.479/-0.915$). The comparison includes the full corpus, the Stage 1 Align70 pool, matched 50% alignment-only and random baselines, and the final CPI50 subset selected from Align70. All values report mean ± standard deviation over three training seeds (seeds 0, 1, and 2).

General transfer and retrieval. CPI50 gives the strongest Flickr30K retrieval performance in both directions. ImageNet zero-shot is highest for Align70, while CIFAR10 and CIFAR100 favor the

Table 24: General-transfer results for CC12M at a fixed 5,430-step training schedule. Values report mean ± standard deviation over three training seeds.

Method	Keep	IN-1K	T2I R@1	I2T R@1	C10	C100
Full	100%	7.379±0.227	9.340±0.295	12.833±0.404	**36.233±0.947**	**9.647±0.700**
Align70	70%	**7.631±0.035**	9.413±0.076	13.100±0.917	35.373±1.419	9.320±0.416
Align50	50%	7.591±0.080	9.293±0.122	12.767±0.723	31.577±0.451	9.580±0.504
Random50	50%	7.377±0.057	9.260±0.125	13.067±0.493	35.767±1.231	9.280±0.469
CPI50 (Ours)	50%	7.541±0.018	**9.793±0.269**	**13.900±0.458**	34.883±1.226	9.240±0.115

full-data run. Thus, as at CC3M scale, the general benchmarks do not exhibit a uniform ordering across curation strategies.

Table 25: SugarCrepe Overall accuracy on CC12M. Overall is the sample-weighted accuracy over all 7,511 benchmark examples. Values report mean ± standard deviation over three training seeds.

Method	Keep	SC Overall
Full	100%	**61.612±0.523**
Align70	70%	61.031±0.402
Align50	50%	61.057±0.160
Random50	50%	61.128±0.253
CPI50 (Ours)	50%	61.434±0.078

SugarCrepe. Full achieves the highest SugarCrepe Overall accuracy. Among the 50%-budget subsets, CPI50 gives the highest mean accuracy.

Table 26: SugarCrepe++ aggregate results on CC12M. Both and Either are pooled example-level accuracies. Margin Mean and Margin Min summarize similarity margins. Values report mean ± standard deviation over three training seeds.

Method	Both	Either	Margin Mean	Margin Min
Full	47.271±0.454	74.438±0.476	0.024138±0.000943	0.005969±0.000466
Align70	47.558±0.648	74.178±0.450	0.024814±0.000378	0.006970±0.000508
Align50	47.474±0.437	74.129±0.340	0.025020±0.000316	0.006588±0.000376
Random50	46.913±0.390	74.073±0.216	0.023532±0.000371	0.005769±0.000380
CPI50 (Ours)	**48.126±0.194**	**74.529±0.306**	**0.025480±0.000140**	**0.007464±0.000285**

SugarCrepe++. CPI50 is highest on all four reported SugarCrepe++ aggregate metrics, including the strict Both criterion.

Table 27: VL-CheckList-VG aspect-level accuracy on CC12M. Values report mean ± standard deviation over three training seeds.

Method	Overall	Object	Attribute	Relation
Full	**68.037±0.114**	**72.033±0.163**	60.730±0.208	39.472±0.379
Align70	67.419±0.328	71.338±0.279	60.484±0.529	39.296±0.746
Align50	67.230±0.091	71.083±0.083	60.255±0.161	39.648±0.580
Random50	67.983±0.159	71.946±0.063	**60.810±0.176**	39.615±1.059
CPI50 (Ours)	67.259±0.108	71.074±0.084	60.144±0.271	**40.045±0.313**

VL-CheckList-VG accuracy.

VL-CheckList-VG margins. At CC12M scale, the overall pattern remains metric-dependent. CPI50 gives the strongest retrieval results, the strongest SugarCrepe++ aggregate results, and the

Table 28: VL-CheckList-VG mean similarity margins on CC12M. Values report mean ± standard deviation over three training seeds.

Method	Overall	Attribute	Object	Relation
Full	0.097793±0.002528	0.027675±0.000853	0.115389±0.002920	−0.010898±0.000507
Align70	0.098110±0.001319	0.027993±0.001127	0.115708±0.001444	−0.010598±0.000561
Align50	**0.098403±0.000847**	**0.028168±0.000701**	**0.115976±0.000938**	−0.010055±0.000814
Random50	0.096964±0.000205	0.027735±0.000553	0.114406±0.000329	−0.010904±0.000465
CPI50 (Ours)	0.098303±0.000458	0.026533±0.000622	0.115946±0.000493	**−0.010022±0.000366**

highest VL-CheckList Relation accuracy, while Full or other controls remain higher on several classification and non-relation VL-CheckList metrics. We therefore interpret the CC12M experiment as evidence that the CPI selection behavior persists at the larger corpus scale rather than as uniform improvement on every downstream metric.

G.2 ViT-S/32 Backbone Scaling

We next evaluate the same 50%-budget Align50 versus CPI50 comparison with a CLIP-style ViT-S/32 visual encoder. Both methods use the same ViT-S/32 initialization and training protocol and differ only in the selected training subset. Results report mean ± standard deviation over three training seeds.

Table 29: General-transfer results for ViT-S/32 on CC3M. Values report mean ± standard deviation over three training seeds.

Method	IN-1K	T2I R@1	I2T R@1	C10	C100
Align50	**7.103±0.038**	9.947±0.230	13.200±0.721	38.420±0.164	12.640±0.053
CPI50 (Ours)	6.849±0.022	**10.027±0.046**	**13.833±0.603**	**40.263±0.106**	**13.337±0.086**

General transfer.

Table 30: SugarCrepe subset-level results for ViT-S/32. Values report mean ± standard deviation over three training seeds.

Method	Overall	Add Att	Add Obj	Rep. Att	Rep. Obj	Rep. Rel	Swap Att	Swap Obj
Align50	58.523±0.181	**60.308±0.221**	58.341±0.318	**62.014±1.151**	65.456±0.642	**52.181±0.351**	53.954±0.378	45.850±0.943
CPI50 (Ours)	**58.971±0.154**	59.875±0.167	**59.441±0.292**	61.802±0.000	**66.465±0.160**	51.422±0.444	**54.104±0.173**	**49.388±1.472**

SugarCrepe.

Table 31: SugarCrepe++ results for ViT-S/32. Overall metrics are reported together with the five strict `acc-both` subset accuracies. Values report mean ± standard deviation over three training seeds.

Method	Cap. 1	Cap. 2	Both	Either	Rep. Att	Rep. Obj	Rep. Rel	Swap Att	Swap Obj
Align50	58.826±0.206	**62.378±0.246**	**46.598±0.330**	74.606±0.131	**50.423±0.366**	58.071±0.493	**37.150±0.082**	38.939±0.483	31.973±0.850
CPI50 (Ours)	**59.162±0.213**	62.168±0.413	46.079±0.172	**75.251±0.074**	48.477±0.000	**58.232±0.210**	35.989±0.188	**39.189±0.795**	**33.061±0.408**

SugarCrepe++.

VL-CheckList-VG. ViT-S/32 therefore shows a mixed but non-degenerate transfer pattern: CPI50 is higher on retrieval, CIFAR transfer, SugarCrepe Overall, and VL-CheckList Object, Relation, and Overall, while Align50 remains higher on ImageNet zero-shot and several SugarCrepe and SugarCrepe++ subsets.

Table 32: VL-CheckList-VG aspect-level results for ViT-S/32. Values report mean ± standard deviation over three training seeds.

Method	Overall	Object	Attribute	Relation
Align50	67.066±0.113	71.227±0.033	**56.939±0.079**	38.455±0.790
CPI50 (Ours)	**67.698±0.055**	**71.881±0.090**	56.909±0.067	**39.210±0.233**

G.3 ViT-L/14 Backbone Scaling

For the larger-backbone experiment, we fine-tune the Hugging Face `openai/clip-vit-large-patch14` pretrained checkpoint. Align50 and CPI50 use the same 50% data budget and optimization protocol. To limit disruption of the pretrained representation, the learning rate is 2.5×10^{-7}. Both conditions are evaluated over seeds 0 and 1 using the final checkpoint.

Table 33: General-transfer results for pretrained ViT-L/14 on CC3M. Values report mean ± standard deviation over two training seeds.

Method	IN-1K	T2I R@1	I2T R@1	C10	C100
Align50	73.307±0.030	72.060±0.057	88.950±0.071	95.925±0.092	80.315±0.049
CPI50 (Ours)	**73.444±0.057**	**73.020±0.141**	**89.750±0.212**	**96.140±0.226**	**80.775±0.276**

General transfer.

Table 34: SugarCrepe subset-level results for pretrained ViT-L/14. Values report mean ± standard deviation over two training seeds.

Method	Overall	Add Att	Add Obj	Rep. Att	Rep. Obj	Rep. Rel	Swap Att	Swap Obj
Align50	81.015±0.019	**80.203±0.613**	87.318±0.378	81.282±0.449	**95.823±0.342**	66.536±0.352	63.439±0.319	60.408±0.577
CPI50 (Ours)	**81.341±0.066**	79.697±0.103	**87.997±0.240**	**81.599±0.718**	95.763±0.000	**67.070±0.000**	**63.964±0.424**	**61.020±0.866**

SugarCrepe.

Table 35: SugarCrepe++ results for pretrained ViT-L/14. Overall metrics are reported together with the five strict `acc-both` subset accuracies. Values report mean ± standard deviation over two training seeds.

Method	Cap. 1	Cap. 2	Both	Either	Rep. Att	Rep. Obj	Rep. Rel	Swap Att	Swap Obj
Align50	78.516±0.089	78.096±0.624	69.960±0.475	86.651±0.238	70.876±0.090	93.220±0.342	56.543±0.905	48.348±0.212	45.919±0.866
CPI50 (Ours)	**78.716±0.045**	**78.947±0.252**	**70.591±0.119**	**87.072±0.089**	**71.574±0.180**	**93.584±0.000**	**57.041±0.100**	**49.475±0.743**	**47.551±0.288**

SugarCrepe++.

Table 36: VL-CheckList-VG aspect-level results for pretrained ViT-L/14. Values report mean ± standard deviation over two training seeds.

Method	Overall	Object	Attribute	Relation
Align50	81.135±0.284	85.135±0.255	70.790±0.091	53.907±0.598
CPI50 (Ours)	**81.281±0.135**	**85.180±0.100**	**70.913±0.120**	**54.867±0.535**

VL-CheckList-VG. For ViT-L/14, CPI50 is higher than Align50 on all five reported general metrics, SugarCrepe Overall, all four SugarCrepe++ overall metrics and all five strict-both subsets, and all four VL-CheckList-VG aspects. The SugarCrepe subset-level results remain mixed, with Align50 higher on add-attribute and replace-object.

H PLUG-IN FULL TABLES

The main text reports compact full-data and CPI-curated comparisons for NegCLIP and CE-CLIP in order to emphasize the central plug-in claim: changing only the training subset can improve objective-side compositional methods. This appendix reports the complete plug-in control results across Full, Random50, Align70, Align50, and CPI50 settings for each objective, together with per-subset SugarCrepe, SugarCrepe++, and VL-CheckList-VG aspect breakdowns.

These results serve two purposes. First, they contextualize the main-text gains against size-matched and alignment-based controls. Second, they show whether the same data-side effect persists when the objective already includes compositional training signals.

Table 37: Plug-in control summary for objective-side compositional methods, including the complete data controls used for each objective.

Objective	Data	SC Ovr	SC++ Ovr	VLC Rel	VLC Ovr
NegCLIP	Full	58.91 ± 0.10	43.46 ± 0.07	37.72 ± 1.13	67.46 ± 0.23
NegCLIP	Random50	59.12 ± 0.29	42.99 ± 0.03	39.78 ± 0.44	67.60 ± 0.08
NegCLIP	Align70	59.10 ± 0.32	44.01 ± 0.01	39.42 ± 0.08	68.11 ± 0.07
NegCLIP	Align50	59.00 ± 0.16	44.36 ± 0.06	40.35 ± 0.29	67.92 ± 0.01
NegCLIP	**CPI50**	$\mathbf{59.53 \pm 0.62}$	43.76 ± 0.22	$\mathbf{41.56 \pm 0.23}$	$\mathbf{68.49 \pm 0.14}$
CE-CLIP	Full	60.20 ± 0.48	35.46 ± 0.95	32.72 ± 0.09	70.24 ± 0.41
CE-CLIP	Random50	60.86 ± 0.39	37.14 ± 0.30	34.95 ± 1.15	70.41 ± 0.07
CE-CLIP	Align70	59.47 ± 0.58	35.39 ± 0.95	34.89 ± 0.89	69.93 ± 0.72
CE-CLIP	Align50	$\mathbf{61.49 \pm 0.33}$	$\mathbf{37.66 \pm 0.82}$	34.01 ± 0.07	70.17 ± 0.03
CE-CLIP	**CPI50**	61.07 ± 0.43	36.60 ± 0.04	$\mathbf{37.64 \pm 3.38}$	$\mathbf{70.72 \pm 0.47}$

H.1 NEGCLIP FULL RESULTS

We also report a routed Align70 diagnostic. Unlike CPI50, this setting does not discard the lower-CPI portion of Align70. Instead, all Align70 samples participate in the standard contrastive loss, while the NegCLIP auxiliary hard-negative loss is applied only to the top CPI-scored samples whose count matches CPI50. The remaining lower-CPI samples are used only with the standard contrastive objective and are not routed through the auxiliary hard-negative construction. This tests whether CPI can route objective-side compositional training toward samples with stronger phrase-level support, rather than only selecting a smaller training subset.

Table 38: NegCLIP full SugarCrepe subset breakdown across curation settings. Bold marks the best value in each column.

Setting	rep_obj	rep_att	rep_rel	add_obj	add_att	swap_obj	swap_att
Full	$\mathbf{66.07 \pm 0.64}$	60.47 ± 1.53	49.40 ± 0.45	58.95 ± 0.86	62.79 ± 0.92	56.94 ± 1.44	55.93 ± 0.53
Align70	65.68 ± 0.77	61.10 ± 0.09	$\mathbf{51.32 \pm 0.65}$	59.29 ± 0.38	61.42 ± 1.23	57.35 ± 0.29	54.50 ± 0.42
Align50	65.65 ± 0.64	60.79 ± 1.26	50.78 ± 0.91	57.98 ± 0.45	62.28 ± 1.02	$\mathbf{58.57 \pm 1.44}$	$\mathbf{57.66 \pm 1.06}$
Random50	65.59 ± 0.13	61.42 ± 0.36	49.96 ± 0.45	59.70 ± 0.14	61.92 ± 0.51	55.10 ± 0.00	56.46 ± 1.70
CPI-routed Align70	65.59 ± 0.13	$\mathbf{62.06 \pm 0.18}$	51.17 ± 0.05	59.12 ± 0.00	61.56 ± 0.41	56.73 ± 1.73	54.43 ± 0.32
CPI50	65.59 ± 1.16	61.93 ± 0.72	50.50 ± 0.10	$\mathbf{60.43 \pm 0.96}$	$\mathbf{63.15 \pm 1.63}$	56.12 ± 0.87	55.41 ± 0.21

Table 39: NegCLIP $\times$ all configurations—full SugarCrepe++ subset breakdown (acc-both).

Setting	rep_obj	rep_att	rep_rel	swap_obj	swap_att
Full	54.87 ± 0.47	46.70 ± 0.18	32.33 ± 0.15	34.90 ± 1.44	37.99 ± 0.64
Align70	55.27 ± 0.34	47.97 ± 0.36	33.11 ± 0.35	37.55 ± 0.58	36.79 ± 1.06
Align50	54.66 ± 0.43	47.14 ± 0.63	$\mathbf{34.74 \pm 0.55}$	$\mathbf{36.94 \pm 0.87}$	$\mathbf{38.51 \pm 1.38}$
Random50	54.66 ± 0.26	46.57 ± 0.54	31.47 ± 0.15	35.10 ± 0.00	37.09 ± 0.11
CPI-routed Align70	55.24 ± 0.04	$\mathbf{49.43 \pm 0.45}$	33.93 ± 0.70	35.92 ± 1.15	37.39 ± 0.21
CPI50	$\mathbf{55.39 \pm 0.86}$	46.89 ± 0.09	33.14 ± 0.60	33.88 ± 1.15	37.24 ± 1.06

Table 40: NegCLIP × all configurations—VL-CheckList-VG aspect breakdown.

Setting	Object	Attribute	Relation	Overall
NegCLIP + Full	71.82 ± 0.11	56.42 ± 0.23	37.72 ± 1.13	67.46 ± 0.23
NegCLIP + Align70	72.33 ± 0.07	57.13 ± 0.07	39.42 ± 0.08	68.11 ± 0.07
NegCLIP + Align50	71.97 ± 0.05	57.45 ± 0.08	40.35 ± 0.29	67.92 ± 0.01
NegCLIP + Random50	71.72 ± 0.15	56.51 ± 0.01	39.78 ± 0.44	67.60 ± 0.08
NegCLIP + CPI-routed Align70	72.08 ± 0.05	56.61 ± 0.19	38.69 ± 0.31	67.80 ± 0.00
NegCLIP + CPI50	$\mathbf{72.47 \pm 0.15}$	$\mathbf{57.68 \pm 0.21}$	$\mathbf{41.56 \pm 0.23}$	$\mathbf{68.49 \pm 0.14}$

The routed Align70 diagnostic separates two possible uses of CPI: subset selection and objective routing. Its results suggest that CPI can also be used to decide where to apply hard-negative compositional losses, although this diagnostic is secondary to the main CPI50 subset-replacement comparison.

H.2 CE-CLIP Full Control Results

Tables 41, 42, and 43 report CE-CLIP control results across Full, Random50, Align70, Align50, and CPI50 settings. These results provide the full plug-in comparison for the CE-CLIP objective.

Table 41: CE-CLIP full SugarCrepe subset breakdown across curation settings.

Setting	rep_obj	rep_att	rep_rel	add_obj	add_att	swap_obj	swap_att
Full	64.98±0.64	59.52±0.36	45.88±1.41	68.16±1.75	65.39±2.55	52.24±0.00	52.25±0.85
Align70	64.07±1.16	59.58±1.17	46.59±0.60	66.15±0.27	65.17±3.68	52.45±0.29	51.13±1.59
Align50	65.35±0.39	**62.69±0.36**	47.33±1.66	**68.96±0.14**	67.27±1.74	**55.92±0.58**	**53.30±1.91**
Random50	**65.89±0.39**	60.60±0.27	47.76±1.56	68.23±0.21	66.55±1.53	**55.92±2.31**	49.47±1.17
CPI50	65.10±0.30	61.74±0.99	**48.19±0.35**	67.39±0.79	**69.00±0.51**	55.71±2.02	51.65±1.49

Table 42: CE-CLIP × all configurations—full SugarCrepe++ subset breakdown (acc-both).

Setting	rep_obj	rep_att	rep_rel	swap_obj	swap_att
CE-CLIP + Full	48.27±0.30	40.10±0.72	22.72±1.16	36.33±2.31	29.88±0.85
CE-CLIP + Align70	46.97±1.11	39.97±1.26	23.47±0.40	37.55±0.58	28.98±2.55
CE-CLIP + Align50	47.73±0.90	**44.42±1.08**	**24.57±0.35**	40.20±0.29	**31.38±1.49**
CE-CLIP + Random50	**48.40±0.64**	41.81±0.81	24.40±0.10	**43.47±0.29**	27.63±0.42
CE-CLIP + CPI50	47.31±1.50	41.56±0.27	**24.57±0.55**	40.41±1.15	29.13±0.42

Table 43: CE-CLIP × all configurations—VL-CheckList-VG aspect breakdown.

Setting	Object	Attribute	Relation	Overall
CE-CLIP + Full	**75.56±0.45**	59.21±0.42	32.72±0.09	70.24±0.41
CE-CLIP + Align70	74.87±0.79	60.39±0.97	34.89±0.89	69.93±0.72
CE-CLIP + Align50	75.23±0.03	**60.92±0.08**	34.01±0.07	70.17±0.03
CE-CLIP + Random50	75.49±0.21	59.07±0.32	34.95±1.15	70.41±0.07
CE-CLIP + CPI50	75.51±0.13	59.46±0.02	**37.64±3.38**	**70.72±0.47**

For CE-CLIP, the clearest gains from CPI-curated data appear on relation-sensitive evaluation. CPI50 achieves the best VL-CheckList-VG Relation and Overall scores, while remaining competitive with Random50, Align70, and Align50 on SugarCrepe and SugarCrepe++ aggregates.

H.3 Cross-objective Summary

Table 44: Cross-objective summary of representative metrics. This table complements the objective-specific compact tables in the main text by showing high-level metrics across CLIP, NegCLIP, and CE-CLIP. Bold marks the best value within each objective family.

Method	Keep	IN-1K	F R@1 Avg	SC R/A Agg	SC Ovr	SC++ Rep	SC++ Ovr	VLC Ovr	VLC Rel
CLIP-Full	100%	7.27	12.16	58.01	57.78	47.01	45.53	66.51	37.71
CLIP-Random50	50%	7.06	11.86	58.01	57.56	45.90	44.30	67.06	39.15
CLIP-CPI50	50%	**7.43**	**12.30**	**58.84**	**58.49**	**47.61**	**45.60**	**67.27**	**39.62**
NegCLIP-Full	100%	7.05	**12.18**	59.28	58.91	44.96	43.46	67.46	37.72
NegCLIP-Random50	50%	6.59	11.48	59.54	59.12	44.53	42.99	67.60	39.78
NegCLIP-CPI50	50%	**7.15**	12.04	**60.07**	**59.53**	**45.51**	**43.76**	**68.49**	**41.56**
CE-CLIP-Full	100%	5.41	**8.51**	61.30	60.20	37.26	35.46	70.24	32.72
CE-CLIP-Random50	50%	5.52	8.23	62.20	60.86	**38.28**	**37.14**	70.41	34.95
CE-CLIP-CPI50	50%	**5.56**	8.03	**62.22**	**61.07**	37.82	36.60	**70.72**	**37.64**

I Additional Ablation Studies

This appendix reports the full ablation tables supporting the design choices summarized in Section 6. We first compare alternative sample-level aggregation functions for phrase-attribution scores, then provide a SugarCrepe subset-level breakdown, and finally vary the second-stage pruning ratio inside the alignment-filtered pool. These experiments justify two default choices used throughout the paper: mean PAS as the sample-level CPI score and the matched 50% final-retention setting as the operating point for main comparisons.

I.1 Score-Aggregation Ablation

Table 45 compares the mean phrase-attribution score against max-based, negative-mass, and combined aggregation variants. All variants use the same Stage-1 alignment-filtered pool and the same final retention budget. Mean-only aggregation provides the strongest SugarCrepe overall accuracy, motivating its use as the default CPI score.

Table 45: Score-aggregation ablation. All variants use the same Stage-1 alignment-filtered pool and the same final retention budget. Bold entries mark column-best values. Mean-only aggregation provides the strongest SugarCrepe overall accuracy, motivating its use as the default CPI score.

Variant	Components	IN-1K	T2I R@1	I2T R@1	C10 ZS	C100 ZS	SC Overall
Mean only	μ	**7.436**	10.40	**14.60**	45.70	15.14	**58.77**
Max only	$\delta^{\max}$	7.332	**11.00**	14.00	43.86	14.72	58.53
Negative-mass only	ν	7.244	10.34	12.30	44.97	15.10	58.25
Mean + Max	$\mu, \delta^{\max}$	7.268	10.42	13.30	46.29	**15.43**	58.67
Mean + Neg	μ, ν	7.236	10.88	13.40	46.31	15.06	58.11
Max + Neg	$\delta^{\max}, \nu$	7.276	10.12	13.30	44.37	14.83	58.22
Full triple	$\mu, \delta^{\max}, \nu$	7.228	10.60	12.70	**46.36**	15.16	58.70

I.2 SugarCrepe Subset Breakdown by Aggregation Variant

Table 46 reports the SugarCrepe subset-level breakdown for the score-aggregation ablation. Although several variants improve individual subsets, these gains do not translate into a stronger aggregate compositional score. The mean-only aggregation achieves the strongest overall SugarCrepe accuracy and is therefore used as the default CPI score.

We also observe a mild accuracy–margin dissociation: while mean-only attains the highest SugarCrepe accuracy, the full triple aggregation attains the highest weighted margin. This suggests that μ, $\delta^{\max}$, and ν may encode partly complementary aspects of phrase contribution. We leave learned or calibrated combinations of these statistics for future work.

Table 46: SugarCrepe subset breakdown by aggregation variant. Bold marks the best value in each column.

Variant	rep_obj	rep_att	rep_rel	add_obj	add_att	swap_obj	swap_att
Mean only	**65.68**	60.28	**50.64**	59.75	61.56	52.24	**53.45**
Max only	64.89	**62.31**	49.86	**60.09**	61.42	49.80	51.95
Mean+Max	65.19	61.29	50.36	59.46	62.72	**54.69**	51.80
Triple	65.44	61.68	49.79	59.60	**63.15**	51.43	52.55
Negmass only	64.59	60.15	49.00	59.89	62.72	52.65	52.10

I.3 Second-Stage Drop-Ratio Sweep

Table 47 varies the Stage-2 drop ratio inside D_{align70}. Moderate phrase-level pruning improves compositional accuracy, while more aggressive pruning reduces general representation quality and eventually harms compositional performance. The matched-50% operating point therefore represents a balance between compositional gain and general retention rather than an arbitrary choice.

Table 47: Second-stage drop-ratio sweep inside D_{align70}. Moderate phrase-level pruning improves compositional accuracy, while more aggressive pruning reduces general representation quality and eventually harms compositional performance. Bold numeric entries mark column-best values; the bold setting name denotes the matched-budget operating point used in the main experiments.

Setting	Final Keep	IN-1K	F R@1 Avg	LP Avg	SC Overall	SC++ Rep
Align70-only	70%	7.558	12.61	47.59	58.34	47.82
CPI drop10	63%	7.510	12.55	47.61	58.56	**47.94**
CPI drop20	56%	7.482	**12.62**	**47.65**	58.71	47.84
CPI drop28.7 (matched 50%)	**50%**	7.436	12.50	47.59	**58.77**	47.71
CPI drop35	45.5%	7.378	12.30	47.42	58.45	46.83
CPI drop45	38.5%	7.241	11.95	47.10	57.92	46.20

Takeaway. The ablations suggest that CPI benefits from a distributed phrase-support aggregation rule: mean PAS rewards captions whose phrase-level support is distributed across multiple candidates, whereas max-based aggregation can overemphasize a single salient phrase and negative-mass aggregation can over-penalize noisy but useful captions. This supports our use of mean PAS for the main CPI subset.

I.4 Mean PAS versus Top-3 PAS Aggregation

A potential concern with mean aggregation is that samples with many valid phrase interventions may dilute a small number of strong responses. We therefore compare the default mean-PAS aggregation with a Top-3 variant that averages only the three largest normalized phrase-level drops for each sample. If a sample contains fewer than three valid interventions, all available interventions are used.

Both variants start from the same Align70 candidate pool, use the same final 50% data budget of 1,111,131 samples, and share the same CLIP ViT-B/32 training configuration, initialization, objective, optimizer, schedule, training steps, and evaluation protocol. The only difference is the sample-level Stage-2 aggregation rule and the resulting selected sample identities.

The concern applies only to a minority of the pool: the median number of valid interventions per sample is 3, while 23.20% of samples have at least 4 interventions, 8.33% have at least 5, and only 0.82% have at least 7. The maximum of 14 valid interventions occurs for a single sample.

The two aggregation rules therefore induce very similar rankings and final subsets, although approximately 5.1% of selected membership differs between them.

Top-3 aggregation is slightly higher on Flickr R@1, whereas mean PAS is higher on ImageNet, Flickr R@5, and CIFAR zero-shot average. Thus, restricting aggregation to the strongest three interventions does not provide a uniform general-transfer advantage.

Table 48: Selection-level agreement between the default mean-PAS aggregation and the Top-3 PAS alternative.

Metric	Mean PAS vs. Top-3 PAS
Pearson correlation	0.9817
Spearman correlation	0.9699
Selected-set overlap	94.92%
Jaccard overlap	90.33%
Samples selected by only one rule	56,437 per rule

Table 49: General-transfer comparison between mean PAS and Top-3 PAS. Values report mean $\pm$ standard deviation over two seeds where available. Flickr retrieval averages the two retrieval directions, and CIFAR ZS averages CIFAR10 and CIFAR100.

Metric	Mean PAS	Top-3 PAS
ImageNet-1K ZS	7.43±0.01	7.27±0.02
Flickr R@1 Avg	12.30±0.00	12.49±0.07
Flickr R@5 Avg	30.94±0.32	29.97±0.19
CIFAR ZS Avg	29.95	29.03±0.09

Table 50: Compositional comparison between mean PAS and Top-3 PAS. Values report mean $\pm$ standard deviation over two seeds.

Metric	Mean PAS	Top-3 PAS
SugarCrepe Overall	58.49±0.40	58.45±0.03
SugarCrepe++ Replace Avg	47.61±0.14	46.78±0.10
SugarCrepe++ Both	45.60±0.03	45.12±0.01
VLC Object	71.29±0.04	71.32±0.08
VLC Attribute	57.53±0.01	57.97±0.03
VLC Relation	39.62±0.12	38.69±0.14
VLC Overall	67.27±0.05	67.22±0.08

The broad aggregate metrics are similar: SugarCrepe Overall differs by only 0.04 points and VL-CheckList Overall by 0.05 points. However, mean PAS is higher on the stricter SugarCrepe++ Replace and Both metrics and on VL-CheckList Relation, while Top-3 is slightly higher on VLC Object and Attribute.

Taken together, these results indicate that CPI is not highly sensitive to the specific choice between mean and Top-3 aggregation: the rankings are strongly correlated and the final subsets overlap by 94.92%. At the same time, Top-3 does not provide a consistent downstream advantage and is weaker on several stricter phrase-sensitive metrics. We therefore retain mean PAS as the default aggregation, since it uses all available interventions and more directly favors examples whose measured sensitivity is distributed across multiple phrases.

J Non-redundancy and Structural Heuristic Analyses

This appendix provides the full analyses supporting Section 6. We first examine how the normalized mean-PAS score relates to global alignment through correlation and kept-versus-dropped distribution analyses. CPI is not intended to be independent of alignment: it measures whether global image–text similarity is sensitive to the caption's constituent phrase substitutions after coarse alignment filtering. We therefore test whether the normalized mean-PAS score remains distinct from global alignment, while allowing selected phrase-supported captions to be slightly more globally aligned on average.

We further test whether PAS can be explained by text-encoder sensitivity alone using a text-only embedding-displacement control, and calibrate nonce substitutions against real-word controls to assess whether the measured drops reflect controlled phrase replacement rather than token-form artifacts alone. We also verify that the saved CPI-kept subset is exactly recovered by sorting the scored D_{align70} pool by the normalized mean-PAS ranking score. Finally, we compare CPI against simple structural caption heuristics based on caption length and extracted phrase count.

J.1 Cross-signal correlation

Table 51 shows that the normalized mean-PAS score has a positive but weak correlation with global alignment within the scored D_{align70} pool used for Stage 2 selection. This weak association indicates that CPI is not simply re-ranking samples by global alignment. At the same time, CPI is not intended to be independent of alignment: phrase-supported captions can also be globally well matched. Global alignment asks whether an image and caption are similar as a whole, whereas CPI asks whether the measured image–text score is sensitive to controlled substitutions of the caption's constituent phrases.

J.2 Text-only sensitivity control

This subsection provides the full text-only sensitivity diagnostic summarized in Section 6. A possible concern is that nonce replacement may perturb the CLIP text encoder independently of the image, so that PAS could reflect text-side sensitivity rather than image-conditioned phrase sensitivity. We test this by comparing the image-conditioned drop

$$\Delta_j = s(I, T) - s(I, \tilde{T}_j)$$

with a text-only embedding displacement

$$d_j^{\text{text}} = 1 - \cos\Big(f_T(T), f_T(\tilde{T}_j)\Big),$$

where f_T is the CLIP text encoder. This control removes the image from the comparison and measures how much the controlled nonce substitution moves the caption representation by itself.

We compute this diagnostic on 50K phrase interventions sampled from D_{align70}. As shown in Table 52, the correlation between Δ_j and d_j^{text} is weakly negative (Pearson $r = -0.215$, Spearman $\rho = -0.204$). Thus, larger nonce-induced text-embedding displacement does not explain larger image-conditioned similarity drops. This supports interpreting PAS as a relative image-conditioned controlled-substitution signal rather than a purely textual nonce artifact, while remaining consistent with our caution that individual phrase drops may be negative and should not be read as grounding labels.

Table 51: Correlation between global alignment and the normalized mean-PAS score on the scored D_{align70} pool used for Stage 2 selection.

Pair	Pearson r	Spearman ρ
alignment vs. normalized mean PAS	0.084	0.088

Table 52: Text-only sensitivity control on 50K phrase interventions sampled from D_{align70}. Text-only displacement is weakly and negatively correlated with image-conditioned PAS, suggesting that PAS is not explained by nonce-induced text-encoder sensitivity alone.

Statistic	Value
Pearson $r(\Delta_j, d_j^{\text{text}})$	-0.215
Spearman $\rho(\Delta_j, d_j^{\text{text}})$	-0.204
Mean Δ_j	-0.0086
Std. Δ_j	0.0184
Mean d_j^{text}	0.0842
Std. d_j^{text}	0.0511

J.3 Real-word calibration of nonce substitutions

This subsection provides the real-word calibration summarized in Section 6. A second concern is that nonce replacement may behave like a generic OOD perturbation: the similarity drop could arise simply because the replacement is an unnatural token, rather than because the removed phrase content matters for the image–text match. To calibrate this effect, we compare the nonce substitution against two real-word controls on 100K phrase interventions sampled from D_{align70}. For the same target token, we construct: (i) the original nonce replacement used by CPI, (ii) a coarse same-category real-word replacement, and (iii) a cross-category distractor replacement. The real-word controls are generated from a deterministic hand-audited category lexicon, and all replacements preserve the CLIP-BPE subtoken count of the original target.

As shown in Table 53, nonce drops are ordered between coarse same-category real-word replacements and cross-category distractors. Paired comparisons confirm both gaps: nonce exceeds same-category replacements by 0.0051 on average, while cross-category distractors exceed nonce by 0.0030 on average ($p < 10^{-16}$ for both paired tests). This ordering suggests that nonce-based PAS is not explained by a generic OOD perturbation alone: if nonce effects were purely generic OOD artifacts, nonce drops would be expected to match or exceed the cross-category distractor control. At the same time, nonce drops do not collapse to either real-word control: they are significantly larger than coarse same-category replacements and significantly smaller than cross-category distractors. We therefore interpret PAS as a controlled-substitution sensitivity signal rather than a pure real-word semantic substitution score or a grounding certificate.

J.4 Subset construction sanity check

As a sanity check, we verify that the saved CPI-kept subset is exactly recovered by sorting the scored D_{align70} pool by the normalized mean-PAS ranking score and retaining the top-ranked samples. Using the original row-level sample identifier, the recovered top-k set matches the saved CPI-kept training subset exactly. This confirms that the kept-versus-dropped analysis below is computed from the same selection rule used for training. All statistics in this subsection are computed over the scored subset of D_{align70} for which the normalized mean-PAS ranking score is available.

J.5 Kept-versus-dropped score distributions

Table 54 compares the samples retained and removed by CPI within the scored D_{align70} pool used for Stage 2 selection. CPI-kept samples have substantially higher normalized mean PAS than CPI-dropped samples (0.479 vs. -0.915), while also having slightly higher global alignment on average (0.25310 vs. 0.24836). The corresponding raw mean PAS values are 0.0946 and 0.0152, respectively. This does not mean that CPI collapses to alignment filtering. Rather, it reflects that phrase-supported

Table 53: Real-word calibration of controlled substitutions on 100K phrase interventions from D_{align70}. Coarse same-category and cross-category replacements are generated from a deterministic hand-audited category lexicon. All replacements preserve the CLIP-BPE subtoken count of the original target.

Replacement	Mean Δ	Std.	Median Δ	$P(\Delta > 0)$ (%)
Coarse same-category real word	0.0177	0.0210	0.0126	82.49
Nonce	0.0227	0.0230	0.0183	87.00
Cross-category distractor	0.0257	0.0243	0.0210	89.23

Table 54: Kept-versus-dropped score distribution within the scored D_{align70} pool used for Stage 2 selection. Normalized mean PAS corresponds to the mean-PAS ranking score used for CPI selection.

Population	Global alignment	Raw mean PAS	Normalized mean PAS
CPI-kept	0.25310	0.0946	0.479
CPI-dropped	0.24836	0.0152	-0.915

captions tend to be globally well matched, while CPI defines the Stage 2 selection rule through controlled-substitution phrase sensitivity.

Together, the correlation and kept-versus-dropped analyses show that CPI is alignment-aware but not alignment-equivalent. The weak global correlation in Table 51 indicates that normalized mean PAS is not simply an alignment re-ranking, while the kept-versus-dropped means show that the selected samples are nevertheless slightly more globally aligned on average. Global alignment removes coarse image–text mismatches in the first stage, whereas CPI reinterprets the remaining global match at the phrase level: it asks whether the measured score is sensitive to controlled substitutions of object, attribute, and relation phrases, or whether the match is sustained by more diffuse or dominant cues. This distinction is important because two image–text pairs can both be globally aligned while differing in how much their constituent phrases support the match under controlled intervention.

J.6 Structural caption heuristic baselines

To test whether CPI reduces to simple caption-level structural heuristics, we construct two additional baselines inside the same Stage-1 alignment-filtered pool D_{align70}. CaptionLengthHigh ranks samples by caption length and retains the top-scoring samples required to reach the same final 50% budget as CPI50. PhraseCountHigh ranks samples by the number of extracted candidate phrases before perturbation and applies the same final retention budget. These baselines are not distribution-matched controls; rather, they are strong structural heuristics designed to test whether CPI's gains can be explained by caption verbosity or phrase density alone.

These structural heuristic baselines are reported as single-run diagnostics under the same evaluation setting. They are intended to compare selection rules, not to replace the repeated-seed main results in Table 1.

Table 55: Compact structural heuristic controls inside D_{align70}. All values are single-seed diagnostic results and are intended to compare selection rules within the structural-control setting, not to replace the repeated-seed main comparisons in Table 1. Caption length and phrase count are competitive proxies for caption richness, but CPI50 better preserves general transfer while remaining strongest on SugarCrepe overall and VL-CheckList-VG Relation.

Method	IN-1K	F R@1 Avg	SC Overall	SC++ Overall	VLC Relation
CaptionLengthHigh	7.020	12.09	58.19	**46.31**	37.88
PhraseCountHigh	7.150	11.99	57.64	45.32	38.41
CPI50	**7.436**	**12.50**	**58.77**	45.62	**39.70**

Table 55 provides the compact structural-control summary, and Tables 56–61 report the full benchmark results. Overall, structural heuristics are competitive, especially CaptionLengthHigh on Sug-

arCrepe++. However, CPI50 remains strongest on ImageNet zero-shot, the Flickr R@1 average, SugarCrepe overall, and VL-CheckList-VG Relation. These results suggest that CPI captures an image-conditioned controlled-substitution signal beyond simple caption verbosity or phrase density, while caption length remains a complementary signal for some replacement-sensitive evaluations.

Table 56: General benchmark results for structural caption heuristic baselines. All values are single-run diagnostic results.

Method	IN-1K	T2I R@1	T2I R@5	T2I R@10	I2T R@1	I2T R@5	I2T R@10	C10 ZS	C100 ZS
Full	7.413	10.260	26.310	35.530	14.05	34.10	43.20	45.685	14.100
Full-Random50	7.060	10.000	25.560	35.180	13.30	32.30	42.40	44.920	15.020
Align-only top50	**7.477**	10.225	**27.655**	**37.015**	13.10	33.85	44.85	42.490	14.100
CaptionLengthHigh	7.020	10.180	27.200	36.600	14.00	34.20	43.30	45.320	14.520
PhraseCountHigh	7.150	**10.580**	27.040	36.060	13.40	33.20	43.90	**46.570**	**15.710**
CPI50	7.436	10.400	27.160	36.240	**14.60**	**35.30**	**45.70**	45.700	15.230

Table 57: SugarCrepe summary for structural caption heuristic baselines.

Method	SC Replace Avg	SC Add Avg	SC Replace/Add Agg	SC Overall
Full	55.89	59.76	58.63	57.69
Full-Random50	56.14	59.68	58.64	57.68
Align-only top50	**56.60**	58.67	58.70	57.92
CaptionLengthHigh	56.39	59.78	59.22	58.19
PhraseCountHigh	54.82	60.12	59.06	57.64
CPI50	56.44	**60.66**	**59.61**	**58.77**

Table 58: SugarCrepe subset-level results for structural caption heuristic baselines.

Method	rep_obj	rep_att	rep_rel	add_obj	add_att	swap_obj	swap_att
Full	65.25	59.52	48.86	58.24	61.27	53.88	51.95
Full-Random50	65.31	60.66	47.87	57.37	61.99	52.65	54.20
Align-only top50	65.19	**61.55**	49.43	57.37	59.97	51.43	**55.41**
CaptionLengthHigh	64.95	60.53	**51.07**	58.29	61.27	53.47	51.95
PhraseCountHigh	64.41	60.79	49.86	58.10	**62.14**	48.16	50.90
CPI50	**65.68**	60.28	50.64	**59.75**	61.56	52.24	53.45

Table 59: SugarCrepe++ summary metrics for structural caption heuristic baselines. SC++ Overall follows the acc-both metric and is not the unweighted average of the subset-level accuracies in Table 60.

Method	SC++ Replace Avg	SC++ Overall
Full	46.81	45.32
Full-Random50	46.26	44.29
Align-only top50	46.65	45.69
CaptionLengthHigh	**48.26**	**46.31**
PhraseCountHigh	47.56	45.32
CPI50	47.71	45.62

Table 60: SugarCrepe++ subset-level results for structural caption heuristic baselines.

Method	rep_obj	rep_att	rep_rel	swap_obj	swap_att
Full	57.08	48.22	35.14	**35.51**	37.84
Full-Random50	55.15	48.98	34.64	31.43	36.94
Align-only top50	56.23	47.59	37.48	34.29	37.69
CaptionLengthHigh	56.11	49.87	**38.33**	32.24	**39.18**
PhraseCountHigh	56.30	49.49	36.20	31.43	37.54
CPI50	**56.60**	**50.13**	36.42	30.20	38.14

Table 61: VL-CheckList-VG results for structural caption heuristic baselines.

Method	Overall	Object	Attribute	Relation
Full	66.514	70.683	56.669	37.712
Full-Random50	67.062	71.147	56.748	39.145
Align-only top50	66.731	70.779	57.489	38.624
CaptionLengthHigh	66.843	71.003	57.500	37.877
PhraseCountHigh	66.760	70.858	57.157	38.411
CPI50	**67.305**	**71.319**	**57.541**	**39.700**

Overall, these diagnostics show that caption structure is a competitive but incomplete proxy: within this structural-control comparison, CPI remains strongest on SugarCrepe overall and VL-CheckList-VG Relation, supporting an image-conditioned controlled-substitution signal beyond caption length or phrase count alone.

J.7 Construct-Validity Analysis of CPI

The preceding analyses show that CPI is not reducible to global alignment, text-only encoder displacement, nonce-token out-of-distribution effects, or simple caption-level structural heuristics. We next examine three additional alternative explanations for the CPI signal: (i) that all phrase scores primarily reflect a single context-invariant object-salience signal, (ii) that high mean PAS is driven by one exceptionally strong intervention, and (iii) that mean aggregation is effectively equivalent to selecting by the maximum phrase response.

These analyses are not intended to establish that CPI directly measures compositional binding or grounding. Rather, they test whether several simpler explanations are consistent with the observed intervention-level statistics. Unless otherwise noted, the analyses in this subsection use the full phrase-scored CC3M-derived corpus rather than restricting to the Stage-1 D_{align70} pool.

J.7.1 Context-specific non-redundancy of phrase sensitivity

We first test whether CPI scores measured in different phrase contexts reduce to the same underlying object-relevance signal. The phrase taxonomy is divided into two groups: NP-context interventions (`compact_np`, `raw_np`) and relation/predicate-context interventions (`relation_np`, `predicate_np_fragment`). Importantly, both groups replace a *nominal target*; this comparison therefore does not contrast object-word interventions with relation-word interventions. Instead, it asks whether sensitivity to the same class of nominal intervention changes with the surrounding syntactic context.

Across the full phrase-scored corpus, 2,216,936 samples have valid phrase-level scores, of which 309,455 contain at least one intervention from both groups. For each such sample, we compute the mean PAS separately for NP context and relation/predicate context and compare the resulting sample rankings.

The two context-specific scores have near-zero linear and rank correlation, and their Top-50% selected sets overlap by only 52.88%, close to the 50% overlap expected from independent half-set selection. Thus, samples ranked high under NP-context interventions are not generally the same samples ranked high under relation/predicate-context interventions.

Table 62: Context-specific CPI non-redundancy analysis. Correlation is computed over the 309,455 samples containing both NP-context and relation/predicate-context interventions. Top-50% overlap measures the fraction of one selected half also selected by the other score; two independent 50% selections have an expected overlap of 50%.

Comparison	Pearson	Spearman	Top-50% overlap
NP context vs. Relation/Predicate context	0.0259	0.0592	52.88%

We additionally compare each phrase-sensitivity score with the Stage-1 global alignment score.

Table 63: Relationship between context-specific CPI scores and global image–caption alignment. Top-50% overlap compares the selected half under each CPI score with the alignment-selected half.

CPI score	Spearman with alignment	Top-50% overlap
NP-context PAS	0.2302	57.38%
Relation/predicate-context PAS	0.1651	55.05%
Mean PAS	0.2692	58.55%
Max PAS	0.1946	56.27%

All four scores retain only weak rank association with global alignment. Together, these results argue against the simplest explanation in which CPI is determined solely by a context-invariant notion of replaced-object importance or by global alignment itself. The result does not imply that relation/predicate-context PAS directly measures relational binding: the perturbed token remains nominal in all cases. Rather, the same class of nominal intervention induces substantially different rankings depending on its surrounding phrase context.

J.7.2 Distributed sensitivity beyond a single dominant intervention

A second possibility is that high mean PAS is produced by one highly salient object or phrase whose replacement causes an extreme similarity drop. To test this alternative while controlling for obvious sample-level confounders, we construct matched bins based on intervention count, global image–caption alignment, and caption length. Intervention counts are grouped as exactly 2, 3, 4, or at least 5 valid interventions. Within each matched bin, samples are split into the upper and lower halves by mean PAS.

After forming the matched bins over the full phrase-scored corpus, the resulting analysis contains approximately 1.99M samples across 800 bins. Alignment and caption length are nearly identical between the CPI-high and CPI-low groups: the mean differences are only $+0.000621$ in alignment and -0.055 words in caption length.

Table 64: Matched distributed-sensitivity analysis. Entries report CPI-high minus CPI-low after matching on intervention count, global alignment, and caption length.

Statistic	CPI-high − CPI-low
Alignment score	+0.000621
Caption length	−0.055 words
Maximum PAS	+0.094447
Second PAS	+0.051540
Third PAS	+0.034938
Minimum PAS	+0.042546
Positive-intervention ratio	+18.69 pp
All-positive rate	+33.22 pp
Top-1 positive share	−9.21 pp
Positive second/first ratio	+11.24 pp

High-CPI samples therefore differ not only in their strongest intervention. Their second, third, and minimum responses are also higher, the fraction of positive interventions increases by 18.69

percentage points, and the probability that all interventions are positive increases by 33.22 points. At the same time, the relative contribution of the strongest positive intervention decreases by 9.21 points.

The same direction holds within every intervention-count group.

Table 65: Matched CPI-high minus CPI-low differences stratified by the number of valid interventions.

Interventions	Second PAS	Minimum PAS	All-positive	Top-1 share	Second/first ratio
2	$+0.0566$	$+0.0566$	$+30.92$ pp	-8.25 pp	$+12.90$ pp
3	$+0.0511$	$+0.0366$	$+35.55$ pp	-9.62 pp	$+10.84$ pp
4	$+0.0457$	$+0.0283$	$+35.87$ pp	-10.48 pp	$+9.59$ pp
5+	$+0.0393$	$+0.0225$	$+30.76$ pp	-10.08 pp	$+7.64$ pp

This pattern is inconsistent with a simple single-dominant-intervention explanation. After controlling for intervention count, global alignment, and caption length, high-CPI examples exhibit elevated responses beyond the maximum intervention and lower concentration on the strongest response. This supports an interpretation of mean PAS as capturing broadly distributed phrase sensitivity. It does not, by itself, establish compositional binding: multiple independently important objects could also produce positive responses across several nominal interventions.

J.7.3 Mean aggregation versus maximum-response selection

Finally, we test whether mean PAS simply acts as a softened version of maximum PAS. Because most captions contain only a small number of valid interventions, the two statistics are naturally correlated; correlation alone is therefore insufficient to determine whether they prioritize the same examples.

Using the same intervention-count, alignment, and caption-length matched bins as above, we independently select the Top-50% of samples by mean PAS and by maximum PAS. We then distinguish samples selected by both rules, selected only by mean PAS, selected only by maximum PAS, or selected by neither.

Table 66: Overlap between mean-PAS and maximum-PAS Top-50% selections. Overlap is the fraction of the mean-selected set also selected by max PAS; Jaccard is the intersection-over-union of the two selected sets.

Interventions	Mean–Max overlap	Jaccard
2	86.02%	75.46%
3	82.18%	69.75%
4	80.07%	66.77%
5+	77.98%	63.91%
Overall	83.12%	71.12%

Although the two aggregation rules overlap substantially, they are not equivalent. Across the matched analysis, 827,246 samples are selected by both rules, while 167,991 samples are selected exclusively by mean PAS and another 167,991 exclusively by maximum PAS. The overlap also decreases as the number of valid interventions increases.

We directly compare the two exclusive groups below.

Mean-only samples have a *smaller* maximum response than Max-only samples, but substantially larger second, third, and minimum responses. They are also much more likely to have consistently positive interventions, while concentrating substantially less of their positive response on the single strongest intervention. Thus, mean-based CPI does not merely reproduce maximum-response ranking: it preferentially retains a substantial set of examples whose sensitivity is distributed more consistently across multiple phrase interventions.

Summary and scope of the construct-validity analyses. The three analyses address complementary alternative explanations. Context-specific rankings argue against a single context-invariant

Table 67: Mean-only versus Max-only exclusive selection analysis. Entries report Mean-only minus Max-only. The two groups are matched on intervention count, global alignment, and caption length.

Statistic	Mean-only − Max-only
Alignment score	$+0.000644$
Caption length	-0.015 words
Intervention count	0.000
Mean PAS	$+0.025386$
Maximum PAS	-0.044277
Second PAS	$+0.063375$
Third PAS	$+0.056466$
Minimum PAS	$+0.072752$
Positive-intervention ratio	$+32.18$ pp
All-positive rate	$+65.89$ pp
Top-1 positive share	-37.39 pp
Positive second/first ratio	$+59.03$ pp

object-salience signal; the matched high-versus-low analysis argues against single-intervention dominance; and the exclusive mean-versus-max comparison shows that mean aggregation retains examples that would not be prioritized by the strongest intervention alone.

Taken together, CPI is not readily explained by a single context-invariant object-salience signal, by a greater number of available interventions, or by one extreme phrase response. Instead, high-CPI examples exhibit context-dependent and broadly distributed phrase sensitivity, and mean-based aggregation retains a substantial set of examples characterized by consistently positive responses across multiple interventions.

These results narrow the interpretation of CPI but do not establish that its similarity drop is an exclusive or direct measure of compositional binding or grounding. In particular, several independently salient objects can also produce distributed nominal-intervention responses, and scorer-specific visual or lexical biases are not eliminated by these analyses.

J.8 Cross-Scorer Robustness with SigLIP

CPI uses phrase-level similarity drops from a pretrained vision–language scorer to rank examples. A natural concern is therefore that the resulting subset may depend strongly on the particular CLIP scorer used to compute PAS. To test this, we replace the original OpenAI CLIP ViT-B/16 phrase scorer with a SigLIP ViT-B/16 scorer while leaving the downstream training pipeline unchanged.

Both scoring conditions start from the same Stage-1 Align70 candidate pool and retain the same final 50% data budget. The downstream model remains the same CLIP ViT-B/32 model with the same initialization, contrastive objective, optimizer, scheduler, learning rate, batch size, maximum number of training steps, evaluation code, and random seeds. Thus, the only changes are the Stage-2 phrase-level scorer and the identities of the samples selected by its resulting ranking.

J.8.1 Selection-level agreement

We first compare the sample rankings produced by CLIP- and SigLIP-based PAS.

Table 68: Cross-scorer agreement between the original CLIP-based CPI ranking and a SigLIP-based CPI ranking computed over the same Stage-1 candidate pool.

Metric	CLIP vs. SigLIP
Kendall τ	0.385
Top-50% selected-set overlap	80.6%

The Kendall correlation of 0.385 shows that the two scorers do not induce the same fine-grained ordering over the candidate pool. Nevertheless, 80.6% of the final selected samples are shared between the two 50%-budget subsets. Because each rule keeps 71.3% of D_{align70} (50% of the corpus out of a

70% pool), two independent random selections would already share 71.3% of their members; the observed overlap exceeds chance by 9.3 points, and the decisive evidence is therefore the downstream replication in Tables 69–72 rather than the overlap itself. Thus, CPI is not scorer-invariant at the level of exact sample ranking, but its selection and its downstream effect are shared across the two vision–language scorers.

J.8.2 Downstream performance of SigLIP-selected data

We next train the same CLIP ViT-B/32 model on the subset selected using SigLIP-based PAS. All values below report mean ± standard deviation over seeds 0 and 1.

Table 69: General-transfer performance of the CLIP model trained on the SigLIP-scorer CPI50 subset.

Metric	SigLIP-scorer CPI50
ImageNet ZS	7.238±0.023
Flickr T2I R@1	10.730±0.325
Flickr I2T R@1	13.550±0.212
CIFAR10 ZS	45.140±0.509
CIFAR100 ZS	14.930±0.127

Table 70: VL-CheckList-VG performance of the CLIP model trained on the SigLIP-scorer CPI50 subset.

Setting	Overall	Object	Attribute	Relation
SigLIP-scorer CPI50	67.046±0.054	71.093±0.060	57.336±0.188	39.155±0.060

For reference, the original CLIP-scorer CPI50 condition obtains 39.62±0.12 on VL-CheckList Relation under the corresponding two-seed comparison. The SigLIP-selected subset is slightly lower on this metric, but relation-sensitive performance does not collapse after replacing the Stage-2 scorer.

Table 71: SugarCrepe subset-level accuracy for the CLIP model trained on the SigLIP-scorer CPI50 subset.

Setting	Add Att	Add Obj	Rep. Att	Rep. Obj	Rep. Rel	Swap Att	Swap Obj
SigLIP-scorer CPI50	60.694±0.409	59.312±0.069	60.406±0.359	65.255±0.257	49.680±0.051	52.628±1.805	54.490±0.288

Table 72: SugarCrepe++ aggregate performance for the CLIP model trained on the SigLIP-scorer CPI50 subset.

Setting	Caption 1	Caption 2	Both	Either
SigLIP-scorer CPI50	57.978±0.476	61.236±0.327	45.438±0.699	73.775±0.104

The SigLIP-selected subset therefore remains usable across general-transfer and compositional evaluations under an otherwise identical downstream training configuration. In particular, VL-CheckList Relation remains at 39.155±0.060, SugarCrepe Replace-Relation reaches 49.680±0.051, and the strict SugarCrepe++ Both metric remains 45.438±0.699.

Interpretation and scope. These results weaken the explanation that CPI's downstream behavior is specific to the exact ranking induced by the original CLIP ViT-B/16 scorer. Changing the scorer substantially alters the fine-grained ranking ($\tau = 0.385$), yet 80.6% of the final selected subset remains shared and the newly selected data retains non-trivial general-transfer and composition-sensitive performance.

This experiment does not show that CPI is free from scorer bias. CLIP and SigLIP are both pretrained vision–language models and may share dataset, lexical, or visual biases. Nor does the experiment establish that SigLIP-based scoring is better than CLIP-based scoring. The supported conclusion is

narrower: CPI is not exclusively dependent on the exact sample ranking of a single CLIP scorer, and its data-selection behavior exhibits substantial cross-scorer robustness.

J.9 Protocol Robustness and Phrase-Extraction Validity

CPI relies on several implementation choices when constructing phrase-level interventions, including the phrase extractor, target-token selection rule, nonce realization, and exact CLIP-BPE length matching. We therefore test whether the resulting data selection and downstream utility are overly sensitive to these choices. We additionally perform a manual audit of the extracted phrase spans themselves.

J.9.1 Phrase-extractor robustness

The default implementation uses a spaCy-based phrase extractor. To test whether CPI depends strongly on a particular dependency parser, we replace this component with a Stanza dependency-based extractor while keeping the target-token rule, nonce-generation protocol, BPE-length constraint, image–text scorer, aggregation rule, Stage-1 pool, and final data budget fixed.

Selection-level stability is evaluated on a fixed 100K subset of the Stage-1 Align70 pool containing samples for which both extractors produce valid CPI scores.

Table 73: Phrase-extractor robustness. The selected-set overlap is measured at the same Stage-2 operating budget.

Variant	Spearman vs. Default	Selected-set overlap	Mean CPI
Default spaCy	–	–	0.07495
Stanza dependency extractor	0.902	94.76%	0.07511

Despite changes in dependency structure and phrase boundaries, the two extractors produce a high sample-level rank correlation and retain 94.76% of the same selected examples. Their average CPI scales are also nearly identical.

We further train the same CLIP ViT-B/32 model on the subsets selected by each extractor. These are single-seed diagnostic runs.

Table 74: Downstream comparison of the default spaCy extractor and the Stanza dependency-based alternative. All values are single-seed diagnostics.

Variant	IN-1K	T2I R@1	I2T R@1	C10	C100	SC Ovr	SC++ Both	VLC Ovr
Default spaCy	7.392	11.160	13.000	45.400	15.630	58.328	45.533	67.175
Stanza	7.374	11.120	13.100	45.330	15.680	58.361	45.490	67.160

The downstream differences are minor and metric-dependent. General-transfer differences are at most 0.1 points in the reported metrics, SugarCrepe Overall differs by only 0.033 points, SugarCrepe++ Both by 0.043 points, and VL-CheckList Overall by 0.015 points. Thus, CPI is not tied to the exact phrase segmentation produced by spaCy, although the extractor still affects a small fraction of selected membership.

J.9.2 Target-token and nonce-protocol sensitivity

We next vary the target-token selection rule and nonce-generation protocol. Four variants are compared:

- **Default**: default target rule, phonotactic nonce, exact CLIP-BPE length matching;
- **Syntactic-head**: syntactic-head target selection, with the default nonce and exact BPE matching;
- **Hash-char**: default target selection with a hash-character nonce, while preserving exact BPE length;

- **BPE-unmatched**: default target selection and phonotactic nonce without exact BPE-length matching.

Because the protocols differ slightly in phrase coverage, we restrict the comparison to a common universe containing 5,498,225 phrase interventions over 2,187,108 samples for which all four protocols produce valid scores. The same Align70 pool and final Stage-2 budget are then used for every variant, yielding 1,098,010 training samples per condition.

Table 75: Selected-set overlap with the default CPI protocol. All variants use the same common phrase universe, Stage-1 pool, and final training-data budget.

Variant	Intersection	Overlap with Default	Jaccard
Syntactic-head	1,065,616	97.05%	94.27%
Hash-char	969,522	88.30%	79.05%
BPE-unmatched	985,359	89.74%	81.39%

Changing the target-token rule leaves most selected membership unchanged, whereas nonce realization and BPE matching alter roughly 10–12% of the selected set. These components therefore affect CPI ranking in practice, rather than being inert implementation details.

We next evaluate the resulting single-seed training subsets.

Table 76: Protocol-sensitivity results. All values are single-seed diagnostics under the same CLIP ViT-B/32 training configuration.

Variant	IN-1K	T2I R@1	I2T R@1	SC Ovr	SC++ Both	VLC Attr	VLC Rel	VLC Ovr
Default	7.392	11.160	13.000	58.328	45.533	57.239	39.216	67.175
Syntactic-head	7.364	10.660	13.300	58.101	45.091	57.270	38.886	66.865
Hash-char	7.180	10.580	12.900	58.155	45.133	57.316	38.646	66.716
BPE-unmatched	7.332	10.520	13.600	58.115	44.608	56.902	38.702	67.118

The default protocol gives the highest SugarCrepe Overall, SugarCrepe++ Both, VL-CheckList Relation, and VL-CheckList Overall in this single-seed comparison, but the alternative protocols remain close on most metrics. The largest compositional drop appears when removing exact BPE matching on SugarCrepe++ Both ($45.533 \rightarrow 44.608$).

These results should not be interpreted as a statistically significant ranking among protocol variants. Rather, they show that target-token choice, nonce realization, and BPE matching do alter sample membership, while the overall downstream utility of CPI remains comparatively stable under reasonable alternatives. Exact BPE-length matching is therefore best viewed primarily as a control against tokenization-induced variation rather than as a component that uniformly improves every downstream metric.

J.9.3 Manual phrase-extraction validity audit

Finally, we manually audit whether the extracted spans are well formed and correspond to their intended extraction types. We perform two complementary audits: a uniform random sample of 500 spans from the full extracted-span population, and a type-balanced audit containing 125 examples from each of the four extraction types.

The uniform audit yields 95.60% validity, while reweighting the type-balanced results according to the actual extraction distribution gives a similar population estimate of 95.31%. The lower validity of the rare `relation_np` and especially `predicate_np_fragment` types is masked by their low frequency in the overall population and should therefore be interpreted separately rather than using the balanced 89.40% figure as a population-level estimate.

Overall, approximately 95.3–95.6% of spans in the extracted population are judged well formed and consistent with their intended phrase type.

Table 77: Manual phrase-extraction validity audit. The type-balanced estimate intentionally oversamples rare and more difficult phrase types; the population-reweighted estimate applies the observed type-level validities to their actual extraction frequencies.

Audit	Samples	Valid	Invalid	Validity
Uniform random	500	478	22	95.60%
Type-balanced	500	447	53	89.40%
Population-reweighted estimate	–	–	–	95.31%

Table 78: Type-level results from the balanced phrase-extraction audit.

Extraction type	Population freq.	Audited	Valid	Invalid	Validity
`compact_np`	89.76%	125	120	5	96.00%
`raw_np`	3.69%	125	119	6	95.20%
`relation_np`	4.87%	125	111	14	88.80%
`predicate_np_fragment`	1.68%	125	97	28	77.60%

K Theoretical Details and Proofs

This appendix proves the three statements of Section 4 and records what each one does and does not license. Notation follows Section 4: $s(I,T) = \langle f_I(I), f_T(T)\rangle$ with unit-norm encoders, $\mathcal{T}_j T = T[w_j \leftarrow r_j]$ with $r_j \in \mathcal{R}(w_j)$, $\Delta_j = s(I,T) - s(I,\mathcal{T}_j T)$, and $\mu = \frac{1}{K}\sum_j \Delta_j$.

K.1 Proof of Lemma 1 (alignment decomposition)

For each j, $s(I,T) = s(I,\mathcal{T}_j T) + \Delta_j$ by definition of Δ_j. Averaging the K identities gives $s(I,T) = \frac{1}{K}\sum_j s(I,\mathcal{T}_j T) + \frac{1}{K}\sum_j \Delta_j = h + \mu$. □

The lemma is an identity and therefore holds for any scorer, any phrase extractor, and any replacement rule; it carries no assumption. Its content is the corollary stated in the main text: because a pair-level filter observes only $h + \mu$, it cannot distinguish a pair whose similarity survives content removal ($\mu \approx 0$) from one whose similarity is carried by content ($\mu > 0$) when the two sums coincide. Equation 2 is thus not a posited objective but the reason a second measurement is needed. Which replacement rule makes μ interpretable is the subject of Proposition 2; whether μ and s are empirically distinct is the subject of Appendix J ($\mathrm{corr}(s,\mu) = 0.084$ on D_{align70}).

K.2 Proof of Proposition 2 (first-order sensitivity)

Fix (I,T) and a phrase j with head w_j and replacement $r_j \in \mathcal{R}(w_j)$. Let $\mathbf{u} = f_I(I)$ and let $\mathbf{e}_{w_j}, \mathbf{e}_{r_j} \in \mathbb{R}^m$ be the input embedding blocks of the two tokens. Because $|\tau(r_j)| = |\tau(w_j)|$, the two blocks have the same dimension m and occupy the same positions in the input sequence, so the map

$$g(\mathbf{e}) \;:=\; \big\langle \mathbf{u},\, f_T\big(T[\mathbf{e}_{w_j} \leftarrow \mathbf{e}]\big)\big\rangle, \qquad \mathbf{e} \in \mathbb{R}^m, \tag{8}$$

is well defined on the segment $\{\mathbf{e}_{w_j} + t(\mathbf{e}_{r_j} - \mathbf{e}_{w_j}) : t \in [0,1]\}$, with positional encodings and attention masks identical at every point of the segment. Under the smoothness assumption stated before Proposition 2, g is twice differentiable there with $\|\nabla^2 g\| \le M_j$. Write $\boldsymbol{\delta} = \mathbf{e}_{r_j} - \mathbf{e}_{w_j}$. Taylor's theorem with the Lagrange remainder gives, for some $t^* \in (0,1)$,

$$g(\mathbf{e}_{r_j}) - g(\mathbf{e}_{w_j}) = \nabla g(\mathbf{e}_{w_j})^\top \boldsymbol{\delta} + \tfrac{1}{2}\,\boldsymbol{\delta}^\top \nabla^2 g(\mathbf{e}_{w_j} + t^*\boldsymbol{\delta})\,\boldsymbol{\delta}. \tag{9}$$

By the chain rule $\nabla g(\mathbf{e}_{w_j})^\top \boldsymbol{\delta} = \langle \mathbf{u}, J_j \boldsymbol{\delta}\rangle$, where J_j is the Jacobian of f_T with respect to the block of w_j evaluated at T. Since $\Delta_j = g(\mathbf{e}_{w_j}) - g(\mathbf{e}_{r_j})$,

$$\Delta_j = -\langle \mathbf{u}, J_j\boldsymbol{\delta}\rangle - R_j, \qquad R_j := \tfrac{1}{2}\,\boldsymbol{\delta}^\top \nabla^2 g(\mathbf{e}_{w_j} + t^*\boldsymbol{\delta})\,\boldsymbol{\delta}, \qquad |R_j| \le \tfrac{1}{2} M_j \|\boldsymbol{\delta}\|^2. \tag{10}$$

Now use the embedding decomposition assumed before Proposition 2, $\mathbf{e}_w = \boldsymbol{\phi}(\sigma(w)) + \mathbf{c}_w$. Surface-form invariance $\sigma(r_j) = \sigma(w_j)$ cancels the $\boldsymbol{\phi}$ terms, and the nonce carries content $\mathbf{c}_{r_j} = \boldsymbol{\varepsilon}_{r_j}$, so $\boldsymbol{\delta} = \boldsymbol{\varepsilon}_{r_j} - \mathbf{c}_{w_j}$ and

$$\Delta_j = \langle \mathbf{u}, J_j \mathbf{c}_{w_j}\rangle - \langle \mathbf{u}, J_j \boldsymbol{\varepsilon}_{r_j}\rangle - R_j, \tag{11}$$

which is Equation 4.

Relaxed invariances. If $|\tau(r_j)| \neq |\tau(w_j)|$, the blocks differ in dimension, the positions of every later token shift, and g is no longer a function on a fixed segment; the difference $s(I,T) - s(I,\mathcal{T}_j T)$ then contains the response of f_T to a change of sequence length and positional encoding, which depends on $|\tau(r_j)| - |\tau(w_j)|$ and not on $\mathbf{c}_{w_j}$. If $\sigma(r_j) \neq \sigma(w_j)$, the term $\langle \mathbf{u}, J_j(\boldsymbol{\phi}(\sigma(r_j)) - \boldsymbol{\phi}(\sigma(w_j)))\rangle$ survives the cancellation. If r_j is a real word with content $\mathbf{c}_{r_j}$, the artifact term becomes $\langle \mathbf{u}, J_j \mathbf{c}_{r_j}\rangle$, whose sign and size depend on the semantic relation between r_j and the image. Each of the three terms has a sign determined by the replacement rather than by the phrase, so there is no reason to expect it to cancel when averaged over the phrases of a caption. □

What the proposition licenses. Δ_j is a first-order response of the fixed scorer to removing the lexical content of one head. It is not a statement about where w_j appears in the image, and it can be negative when $\langle \mathbf{u}, J_j \mathbf{c}_{w_j}\rangle < 0$, i.e., when the content of w_j pulls the caption embedding away from the image under the fixed scorer (stylistic, stock-caption, or text-prior phrases). CPI ranks by relative μ and does not require positive drops. The artifact term is bounded empirically rather than assumed small: text-only displacement is negatively correlated with Δ_j (Appendix J.2), and nonce drops lie strictly between same-category and cross-category real-word controls (Appendix J.3), which is the ordering $\langle \mathbf{u}, J_j \mathbf{c}_{\text{same}}\rangle > \langle \mathbf{u}, J_j \boldsymbol{\varepsilon}\rangle > \langle \mathbf{u}, J_j \mathbf{c}_{\text{cross}}\rangle$ predicted by the relaxed-invariance analysis when the same-category word shares content with w_j and the cross-category word does not.

K.3 Proof of Proposition 3 (first-order Shapley term)

For $S \subseteq [K]$ let T_S be the caption in which every head outside S is replaced by its nonce (admissibility is checked per token, so T_S is well defined), and let $v(S) = s(I, T_S)$. Then $T_{[K]} = T$ and $T_{[K]\setminus\{j\}} = \mathcal{T}_j T$, so $\Delta_j = v([K]) - v([K] \setminus \{j\})$, the marginal contribution of phrase j to the grand coalition.

Every set function admits the Möbius (Harsanyi) expansion $v(S) = \sum_{R \subseteq S} d(R)$ with $d(\emptyset) = v(\emptyset)$ and $d(R) = \sum_{Q \subseteq R} (-1)^{|R|-|Q|} v(Q)$ (Shapley, 1953). Substituting,

$$\Delta_j = \sum_{R \subseteq [K]} d(R) - \sum_{R \subseteq [K]\setminus\{j\}} d(R) = \sum_{R \ni j} d(R). \tag{12}$$

The Shapley value of player j is the average marginal contribution over orderings, and in dividend form equals $\varphi_j = \sum_{R \ni j} d(R)/|R|$. Subtracting,

$$\Delta_j - \varphi_j = \sum_{R \ni j} \Big(1 - \tfrac{1}{|R|}\Big) d(R) = \sum_{R \ni j,\, |R| \geq 2} \Big(1 - \tfrac{1}{|R|}\Big) d(R), \tag{13}$$

since the $|R| = 1$ term vanishes. Summing $\Delta_j = \sum_{R \ni j} d(R)$ over j counts each R exactly $|R|$ times: $\sum_j \Delta_j = \sum_{R \neq \emptyset} |R|\, d(R)$. Efficiency of the Shapley value gives $\sum_j \varphi_j = v([K]) - v(\emptyset) = \sum_{R \neq \emptyset} d(R)$. Hence

$$\sum_j \Delta_j = \big(v([K]) - v(\emptyset)\big) + \sum_{|R| \geq 2} (|R| - 1)\, d(R), \tag{14}$$

and dividing by K yields Equation 5. If $d(R) = 0$ for all $|R| \geq 2$, both correction sums vanish, $\Delta_j = \varphi_j$, and $\mu = (v([K]) - v(\emptyset))/K$. □

What the proposition licenses. Mean PAS is the exact per-phrase share of the similarity carried by lexical content over the fully hollowed caption whenever phrases do not interact, and otherwise differs from it by a weighted sum of interaction dividends. CPI computes K singleton interventions, not 2^K coalitions, and does not compute $v(\emptyset)$; the interaction term is therefore not estimated but tested indirectly. Three observations bear on it. Mean aggregation outperforms max-only and negative-mass aggregation (Appendix I.1), which is what one expects if the first-order term, and not a single dominant coalition, carries the signal. High-μ captions show higher second, third, and minimum PAS and a lower top-1 share in matched comparisons (Appendix J.7.2), i.e., their sensitivity is distributed across phrases. Top-3 aggregation selects 94.92% of the same pairs (Appendix I.4). None of these observations bounds $d(R)$ for $|R| \geq 2$; for the median caption with three interventions, the full expansion costs eight text forward passes, so estimating $d(R)$ directly is an extension of the cached pipeline rather than a new one.

K.4 Operational interpretation of PAS

Three sign cases arise in practice. $\Delta_j > 0$: removing the head lowers similarity; the scorer is sensitive to the phrase. $\Delta_j \approx 0$: the similarity is sustained by the rest of the caption or by an image–text prior. $\Delta_j < 0$: replacement increases similarity, which under Equation 4 means the content of w_j points away from the image for this scorer. Negative drops are common in web captions and are retained in the mean; they lower μ and thereby rank the caption as more hollow, which is the intended behavior. No sign case certifies visual presence or absence.

L Computational Cost

Table 79 reports the measured cost of the CPI curation pipeline using cached image embeddings. We first compute and store image embeddings once, and then reuse this cache for Stage-1 alignment scoring and Stage-2 counterfactual delta scoring. Thus, the reported GPU cost corresponds to our cached image-embedding implementation rather than repeatedly re-encoding images for every perturbation.

Table 79: Measured computational cost of the CPI curation pipeline on the CC3M-derived corpus using cached image embeddings. Wall-clock time is measured for each component; GPU-hours are wall-hours multiplied by the number of GPUs, and CPU process-hours are wall-hours multiplied by the number of CPU processes.

Component	Wall time	GPU-hours	CPU process-hours
Image embedding cache construction	0.203 h	0.811	–
Stage-1 alignment scoring from cache	0.037 h	0.149	–
Phrase extraction + perturbation generation	0.728 h	–	5.824
Merge perturbation shards	0.024 h	–	0.024
Stage-2 delta scoring from cache	0.186 h	0.742	–
Merge delta-scored shards	0.034 h	–	0.034
Total	1.211 h	1.702	5.882

Overall, CPI curation requires approximately 1.21 hours of sequential wall-clock time, 1.70 GPU-hours, and 5.88 CPU process-hours in our cached image-embedding implementation. The cache is built once per corpus and reused across subsequent curation and scoring runs, so this offline cost can be amortized across CLIP, NegCLIP, and CE-CLIP training experiments.

M Extended Discussion and Limitations

What the results say about web captions. Three empirical facts together describe the corpus rather than the method. First, alignment saturates at the 70% pool (Table 4), so roughly the top two thirds of CC3M pairs are matches by the scorer's account and further sorting by the scorer is sorting noise. Second, within that pool, the global alignment score s and the content component μ of Lemma 1 are only weakly correlated (Appendix J.1), so a large share of well-aligned captions keep their similarity when their content words are removed. Third, the pairs that lose the most similarity under content removal are the ones whose inclusion moves relation and object-presence metrics (Section 5). Read together: a sizable fraction of the supervision that pair-level filters retain is carried by scaffolding and priors rather than by the caption's nouns, and models trained on it learn correspondingly less about which named things are in the image.

Implications for different readers. For practitioners who run a filtering pipeline, CPI is a second stage, not a replacement: keep the existing scalar for coarse removal, cache image embeddings, and add K text forward passes per caption; on CC3M the addition costs 1.70 GPU-hours (Appendix L) and reaches Align70-level compositional quality with 28.7% less data. For developers of compositional objectives, the plug-in results (Appendix H) indicate that a hard-negative loss and phrase-supported positives add rather than duplicate; the routed-Align70 diagnostic suggests that CPI can also decide *where* an auxiliary loss is applied. For benchmark designers, the split between presence-type gains (add-object, relation) and order-type non-gains (swap) is a reminder that these axes are separable at

the data level and should be reported separately. For researchers on data attribution, CPI is a singleton ablation on a fixed scorer; the full coalition expansion of Proposition 3 is computable for typical captions and would directly test the interaction term we treat as residual.

Limitations as sharpened claims. *Nominal interventions only.* The admissible set $\mathcal{R}(w)$ of Equation 3 is defined through a surface-form signature for nominal tokens. Extending it to prepositions and verbs requires signatures for those classes and a nonce generator that respects them; until then CPI measures context-conditioned sensitivity to nominal substitution, which is exactly the quantity that moved relation metrics in every setting we tested. *No word-order sensitivity.* Any functional of the sensitivity profile is invariant to permutations that preserve the token multiset, so swap probes are outside what selection by presence can improve. The correct experiment is therefore not a different selection rule but the combination with an order-sensitive objective, which Table 3 and Appendix H.1 report. *Scorer circularity.* CPI uses a CLIP-family scorer to select data for training a CLIP-family model, and a scorer with limited compositionality is part of the problem being addressed. The SigLIP replication (Appendix J.8) bounds this: two scorers with different data and losses select 80.6% of the same pairs and both improve over Align50. The bound is not a proof of independence, and both scorers were trained on web alt-text; a scorer with stronger relational sensitivity is the natural next input to the same pipeline. *Rule-based extraction.* The extractor is a dependency parser with heuristic filters; 95.3–95.6% of spans are well formed on audit, and a different parser (Stanza) selects 94.76% of the same pairs (Appendix J.9). Parser errors thus change the selection by a few percent and produce only small downstream changes, at most 0.10 points across the reported metrics. *Benchmark artifacts.* VL-CheckList Relation admits text-only shortcuts, which is why SugarCrepe is the primary compositional test and why the Relation gain is interpreted as relation-argument presence. SugarCrepe++ defines no overall score; both aggregation conventions are reported and neither is favored in the text. *Scale and training regime.* All runs fine-tune from a checkpoint for 5,430 steps on CC3M or CC12M. Whether the saturation point, the hollow fraction, and the relation gain persist under from-scratch pretraining on DataComp- or LAION-scale corpora is untested here; the curation cost is linear in the number of captions, so the test is affordable, and the claim in this paper is confined to the regime we ran. *Comparison scope.* T-MARS is compared under a matched pool and budget. DFN and Sieve require a learned filter or a captioner tuned to other corpora and operating points, and a faithful comparison would need those to be re-tuned at CC3M scale; we do not report one.

N Qualitative Examples

This appendix complements the non-redundancy analysis in Section 6 with sample-level qualitative examples. Figure 3 visualizes disagreement cases between alignment-only filtering and CPI-based selection. These examples are intended to illustrate the measured controlled-substitution signal, not to certify visual grounding.

CPI kept / Align50 dropped. The top row of Figure 3 shows samples retained by CPI but removed by Align50. These examples have lower global alignment scores but high normalized mean PAS, indicating that their measured image–text scores are sensitive to constituent phrase substitutions.

Align50 kept / CPI dropped. The bottom row shows the opposite disagreement case. These samples receive higher global alignment scores and are therefore retained by Align50, but CPI assigns low normalized mean PAS. In these cases, the image–caption pair is globally plausible, yet the caption contains subjective, web-style, or text-heavy content whose phrase substitutions do not substantially reduce the measured image–text score. CPI drops these samples because the measured score is comparatively insensitive to the caption's constituent phrase substitutions.

Together, these examples illustrate why CPI is not equivalent to stricter alignment filtering. Alignment determines whether an image–caption pair is globally plausible, whereas CPI asks whether the measured image–text score is sensitive to the caption's constituent phrases under controlled substitution. As a result, CPI can retain samples that alignment-only filtering would discard and remove globally plausible but phrase-insensitive pairs that alignment-only filtering would retain.

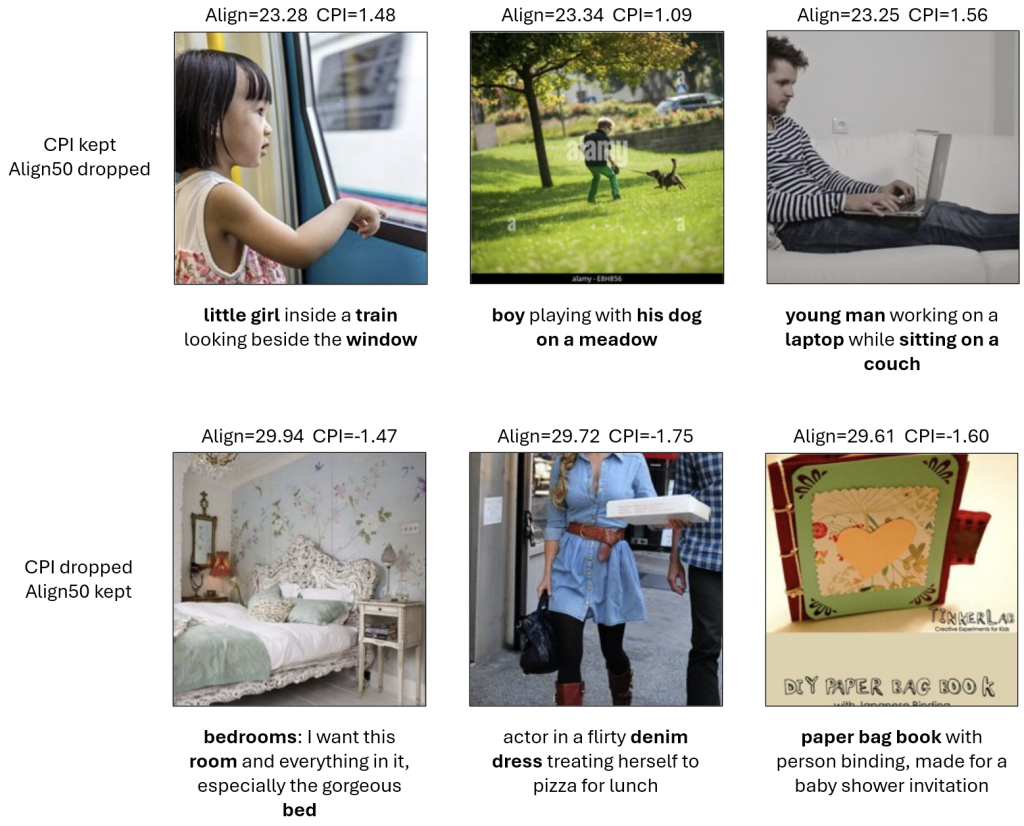


Figure 3: Representative qualitative examples where CPI and Align50 make different curation decisions. Top: samples kept by CPI but dropped by Align50. Although their global alignment scores fall below the Align50 threshold, their captions contain concrete visual phrases whose controlled substitutions produce high normalized mean PAS. Bottom: samples kept by Align50 but dropped by CPI. These pairs are globally plausible, but their captions contain subjective, web-style, or text-heavy descriptions whose constituent phrase substitutions produce low normalized mean PAS. Bold text marks representative caption phrases for readability and is not intended as a grounding annotation. Scores report raw global alignment and normalized mean PAS.