CoRe: Jointly Learning Contextual Representations for Understanding and Generation

Anonymous authors
Paper under double-blind review

Abstract

Understanding and generation place distinct demands on learned representations, yet these demands need not conflict. We introduce CoRe, which jointly learns contextual representations that support understanding, provide conditioning, and serve as the space for generation. Both images and text are generated through continuous flows in the representations used for understanding. Trained from scratch on CC12M, the 383M-parameter model matches same-data contrastive controls on global understanding and reaches COCO-30K FID 10.10, comparable to much larger generators with pretrained components. Stronger language-pretrained systems retain an advantage in prompt following and caption fidelity. Matched interventions show that generating directly in the representation used for understanding improves compositional generation over a separate projected space, even as the latter reconstructs more accurately. A more flexible readout also recovers most of the dense-prediction gap. Together, these results show that understanding and generation can coexist in one contextual representation, with their interaction governed largely by how information is retained, organized, and made accessible.

1 Introduction

"What I cannot create, I do not understand." —Richard Feynman

Feynman's remark describes a demanding test for a learned representation. Generating an image of *one red hat* requires knowing what a hat is, what red is, and what "one" means, and rendering all three coherently; describing that image requires recovering the same facts in language. Generation is not the opposite of understanding but one of its most exacting uses. This suggests a representation-level question: *can the same contextual representation support both understanding an observation and generating it?*

A common view is that the two objectives compete: understanding favors abstract semantics, whereas generation requires fine detail, motivating decoupled encoders or separate codebooks (Wu et al., 2025; Qu et al., 2025; Deng et al., 2025). Yet many "details"—color, texture, boundaries, spatial relations—also matter for understanding (Cimpoi et al., 2014; Geirhos et al., 2018; Zhou et al., 2021), and reconstruction need not conflict with semantics when the representation has sufficient capacity (Ma et al., 2026a). We therefore separate two questions that are often conflated: *what information a representation retains*, and *how that information is organized* for downstream readout (Bengio et al., 2013; Xu et al., 2020). Generation objectives encourage retention of the content needed to recover observations; understanding objectives encourage semantic organization by bringing related observations together. These constraints need not require separate representations.

We build this view into CoRe (Figure 1). Modality-specific encoders map images and text to contextual patch and token representations. Each modality uses the same contextual representation in three roles: it is read for understanding, conditions generation of the other modality, and is itself the space in which that modality is generated. Global readouts are aligned contrastively, while generation proceeds through continuous trajectories in the contextual representation spaces. Because these representations are learned jointly and have no externally defined latent targets, generative learning is anchored primarily to observables—pixels and token identities—with representation regression used only to support sampling. Thus understanding and generation jointly shape one learned representation rather than communicating through a separate generative latent.

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.

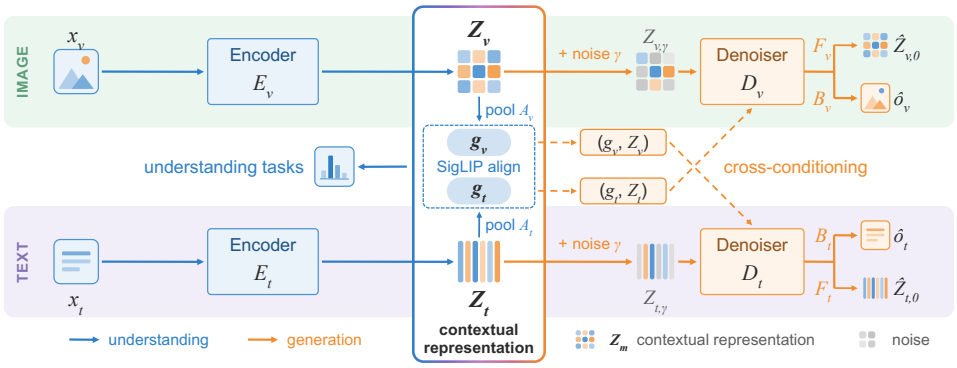

Figure 1: **CoRe uses contextual representation for understanding, conditioning, and generation.** For each modality $m \in \{v, t\}$, the encoder E_m produces a contextual representation Z_m: a sequence of patch representations for images or token representations for text. Attention pooling extracts g_m, with g_v and g_t aligned by SigLIP for understanding. The pair (g_m, Z_m) conditions generation of the other modality. Generation operates in the same representation space: given the noised representation $Z_{m,\gamma}$, the denoiser D_m feeds a representation head F_m that predicts $\widehat{Z}_{m,0}$ for sampling, and an observable head B_m that predicts pixels or token logits $\widehat{o}_m$ for generative supervision.

Two aspects make this construction revealing for studying shared representations. First, *language is generated in the continuous space in which it is understood.* Unlike diffusion over lookup embeddings or separately learned latents (Li et al., 2022b; Dieleman et al., 2022; Chen et al., 2026; Meshchaninov et al., 2026), CoRe generates the contextual token representation that is also pooled for semantic alignment; the same token can therefore take different representations in different contexts. Second, *images are generated in a jointly learned contextual patch space*, rather than in pixels, a pretrained VAE latent, or a frozen representation (Rombach et al., 2022; Yu et al., 2024; Zheng et al., 2026). In both modalities, contextual representations support understanding, condition generation of the other modality, and serve as the space for their own generation, allowing one flow formulation and one conditioning mechanism to serve both directions.

We keep the setting deliberately controlled. The image and text encoders and two denoisers, 383M parameters in total, are trained jointly from scratch on CC12M with public LLaVA-NeXT recaptions and original alt-text (Emporium, 2024); no pretrained language model, text encoder, or VAE is part of the inference model. CoRe matches a same-data, same-step contrastive control on global understanding and reaches COCO-30K FID 10.10, comparable to much larger generators using pretrained T5-XXL encoders and VAEs (Li et al., 2026; Fan et al., 2025). Its prompt following remains below MiniT2I, which uses a pretrained FLAN-T5-Large encoder (Wang et al., 2026), and its caption fidelity trails systems trained on substantially more or task-specific caption data. More importantly, controlled interventions show what sharing changes: generating directly in the representation for understanding improves compositional generation over a matched separate generation space, while a more flexible readout recovers most of the dense-prediction gap.

Contributions. (i) *One representation, three roles.* We introduce CoRe, which learns contextual representations that support understanding, provide conditioning, and serve as the space for generation (Section 3). (ii) *Understanding and generation can be complementary.* We show that learning both capabilities through the same representations preserves global understanding and benefits compositional generation (Sections 4 and 5). (iii) *Joint training reorganizes representations while retaining recoverable detail.* We find that joint training reorganizes contextual representations while leaving much of their dense information recoverable, locating much of the apparent tension in how information is accessed rather than whether it is retained (Section 5).

2 Related Work

Should understanding and generation share a representation? Unified multimodal models combine understanding and generation through shared sequence models over discrete tokens (Wang et al., 2022; 2024b), mixed autoregressive and diffusion objectives (Zhou et al., 2025; Deng et al.,

2025; Xie et al., 2025), diffusion over both modalities (Swerdlow et al., 2025; Yang et al., 2026; Li et al., 2025), or visual interfaces engineered for both uses (Wu et al., 2025; Chen et al., 2025c; Qu et al., 2025; Ma et al., 2026a; Yan et al., 2026). These systems usually build on pretrained language models, VAEs, or vision encoders; we ask a narrower, representation-level question with every component learned from random initialization.

Which space generation runs in. Pretrained semantic features improve diffusion as alignment targets or as the latent space itself (Yu et al., 2024; Zheng et al., 2026), and aligning or jointly tuning an autoencoder latent makes it easier to generate from (Yao et al., 2025; Leng et al., 2025). Conversely, generative and reconstructive training yields transferable features and can sharpen understanding (Chen et al., 2020; He et al., 2022), including encoders trained through generative decoders (Preechakul et al., 2022; Hudson et al., 2024; Chen et al., 2025d; Li et al., 2023) or by captioning (Tschannen et al., 2023); UniDiffuser learns the conditionals and marginals our routes train, but over frozen latents (Bao et al., 2023). DREAM (Li et al., 2026) jointly optimizes contrastive understanding and masked generation over VAE tokens with a frozen T5 encoder. In these approaches generation runs in pixels, an autoencoder latent, or a frozen or separately regularized feature space; in CoRe it runs, in both modalities, in the representation that understanding reads.

Continuous diffusion for language. Continuous diffusion language models corrupt token embeddings or learned latents and decode categories (Li et al., 2022b; Dieleman et al., 2022; Chen et al., 2026; Meshchaninov et al., 2026); their continuous space serves generation only. In CoRe the continuous text sequence is also the text representation for understanding, pooled for alignment and used to condition images. Continuity also makes flow-map and shortcut distillation applicable, compressing a continuous trajectory into a few jumps (Boffi et al., 2024; Frans et al., 2025).

3 CoRe: jointly learned contextual representations

CoRe learns one contextual representation for each modality and uses it in three ways: it is read for understanding, conditions generation of the other modality, and is itself the space in which that modality is generated (Figure 1). The key design choice is therefore not merely to share an encoder, but to make understanding and generation operate on the same learned representation.

One contextual representation per modality. For a paired image and caption (x_v, x_t) and modality $m \in \{v, t\}$,

$$h_m = E_m(x_m), \qquad Z_m = \mathrm{LN}(P_m h_m) \in \mathbb{R}^{N_m \times d}, \qquad g_m = A_m(Z_m) \in \mathbb{R}^{d_g}. \tag{1}$$

Here E_m is a transformer encoder over image patches or text tokens, P_m projects its features tokenwise to width d, and A_m attention-pools the sequence into a global representation g_m. The non-affine layer normalization fixes the scale of each token relative to the noise used below. Because E_m uses self-attention, Z_m is *local in index but contextual in content*: each image entry corresponds to a patch but depends on the full image, and each text entry corresponds to a token but depends on the full caption. The global representation g_m is read from this same sequence. We never align Z_v and Z_t token to token; only their pooled readouts are aligned.

Generation runs in the contextual representation. We corrupt the contextual representation itself with variance-preserving Gaussian noise, parameterized by the log noise-to-signal ratio γ:

$$Z_{m,\gamma} = \alpha_\gamma Z_m + \sigma_\gamma \varepsilon, \qquad \alpha_\gamma^2 = \mathrm{sigmoid}(-\gamma), \qquad \sigma_\gamma^2 = \mathrm{sigmoid}(\gamma), \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{2}$$

The clean endpoint of this trajectory is therefore exactly the contextual representation used for understanding. A modality-specific denoiser predicts both the underlying observation and the clean representation:

$$H_m = D_m(Z_{m,\gamma}, \gamma, C_m), \qquad \widehat{o}_m = B_m(H_m), \qquad \widehat{Z}_{m,0} = F_m(H_m). \tag{3}$$

The observable head B_m predicts RGB pixels for images and vocabulary logits for text, while the representation head F_m estimates the clean representation used by the sampler.

Observable targets anchor a moving representation. The contextual representation Z_m changes throughout training because its encoder is learned jointly with the denoiser, thus has no fixed ground-truth coordinates. If latent regression were the only objective, both the prediction and its target could move together toward an easier solution (Assran et al., 2023). We therefore anchor generation

primarily to fixed observables—pixels and token identities—and use regression to the detached Z_m only as auxiliary supervision for sampling. This lets the representation evolve while keeping its generative meaning tied to what is actually observed.

Cross, self, and null conditioning. To generate modality m from the other modality $\bar{m}$, we condition on its contextual representation and pooled readout, $C_m = (g_{\bar{m}}, Z_{\bar{m}})$: $g_{\bar{m}}$ enters through adaptive normalization, while $Z_{\bar{m}}$ supplies cross-attention keys and values. During training, each denoiser uses one of three routes: *cross* ($p = 0.7$), *self* (0.2), or *null* (0.1). For fixed encoders, paired observations induce a joint distribution over contextual representations,

$$p(Z_v, Z_t) = p(Z_v \mid Z_t)\, p(Z_t) = p(Z_t \mid Z_v)\, p(Z_v), \tag{4}$$

and the cross and null routes train the corresponding conditional and marginal distributions in both directions; classifier-free guidance combines them at inference (Ho & Salimans, 2022). The self route replaces the local condition with a null input while retaining the target modality's own pooled representation, teaching how much of that modality can be generated from its global summary alone. Thus, the same contextual representations define both the cross-modal conditionals and the unimodal marginals of the learned joint distribution.

Joint objective. The same contextual representations are learned jointly for alignment and generation. We optimize

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{img}} + 0.5\,\mathcal{L}_{\text{text}}, \qquad \mathcal{L}_{\text{align}} = \text{SigLIP}(\bar{g}_v, \bar{g}_t), \quad \bar{g}_m = \frac{g_m}{\|g_m\|_2}. \tag{5}$$

The alignment loss encourages paired images and captions to agree through their pooled semantic representations, while leaving the contextual representations themselves to be learned jointly (Zhai et al., 2023). Generation instead anchors each contextual representation to its observable:

$$\begin{aligned}
\mathcal{L}_{\text{img}} &= \mathbb{E}_{\gamma\sim q_v,\, r}\Big[\mathcal{L}_{\text{MSE}}(\widehat{o}_v, x_v) + \mathbf{1}[\gamma < \gamma_{\text{p}}]\big(0.1\,\mathcal{L}_{\text{LPIPS}} + 0.01\,\mathcal{L}_{\text{P-DINO}}\big) + 0.25\,\mathcal{L}_{\text{rep},v}\Big], \\
\mathcal{L}_{\text{text}} &= \mathbb{E}_{\gamma\sim q_t,\, r}\Big[\mathcal{L}_{\text{CE}}(\widehat{o}_t, x_t) + 0.25\,\mathcal{L}_{\text{rep},t}\Big],
\end{aligned} \tag{6}$$

where the expectations are over sampled noise levels γ and conditioning routes r, and

$$\mathcal{L}_{\text{rep},m} = \frac{1}{N_m d}\,\|F_m(H_m) - \text{sg}(Z_m)\|_F^2\,. \tag{7}$$

Pixels and token identities provide the fixed observable targets; the detached representation regression is auxiliary supervision for the clean-representation prediction used by the sampler. For images, frozen VGG16 and DINOv2 provide low-noise perceptual supervision during training only (Ma et al., 2026b; Simonyan & Zisserman, 2014; Oquab et al., 2023). Thus, alignment specifies *which paired observations should agree semantically*, while generation specifies *what their contextual representations must keep recoverable*. The visual noise proposal q_v is fixed logit-normal (Esser et al., 2024); the text proposal q_t follows LangFlow's learned noise schedule (Chen et al., 2026). Remaining implementation details are given in Appendix A.

Inference. A single encoder pass produces the contextual representation Z_m and its pooled readout g_m. Understanding tasks read these representations directly. For cross-modal generation, the same source representation conditions the denoiser, while the target modality starts from Gaussian noise and is iteratively denoised in its contextual representation space. At the final step, B_m maps the denoiser features to RGB pixels or token logits. Thus no separate generative latent, such as a VAE latent, is introduced for generation. The same contextual representation is read for understanding, conditions the other modality, and defines the space in which that modality is generated.

4 Experiments: one representation, three uses

We test whether one contextual representation per modality supports understanding, image-to-text, and text-to-image generation, and whether generation from it can be compressed to a few steps. Each benchmark uses a fixed evaluation protocol for all models we evaluate, including released external models.

Training setup. All base models are trained from random initialization for 125K steps at batch size 8192 on CC12M images (Changpinyo et al., 2021) with the public LLaVA-NeXT recaptions

(Emporium, 2024) and the original alt-text, one caption sampled per image presentation (*CC12M-R*). CoRe has 383M parameters in total.

Controlled variants. We compare variants sharing the same training data and encoder architecture. *Contrastive only* trains the 143M encoders only with the SigLIP objective, providing a baseline without generative training; *no contrastive loss* drops only the alignment loss while retaining all other training losses, testing the role of explicit image–text alignment; *frozen contrastive → generation* freezes the contrastive-only encoders and trains the two denoisers for another 125K steps, testing generation from fixed contrastive representations. CoRe + Gram/iBOT adds Gram anchoring and iBOT-style self-distillation on encoder patch features after 24K steps (Siméoni et al., 2025; Zhou et al., 2021), testing whether these regularizers improve dense spatial prediction. We also apply these regularizers to *Contrastive only*, allowing comparison with and without generative training under the same structural regularization.

Fine-tuning. For text-to-image generation we fine-tune on a public 120K synthetic mixture (Wang et al., 2026). The final recipe (♯ rows) first updates all CoRe + Gram/iBOT parameters on CC12M-R at 512^2 for 10K steps, then trains only the image denoiser on the mixture for 8K steps, preserving the first-stage model's understanding and captioning.

Across the following tables, blue denotes published results, teal denotes our evaluations of released models or outputs, and black denotes our models and controls. Blank entries indicate unreported results, while "–" indicates evaluations that were not run or are not applicable. Architecture and training details are provided in Appendix A, and evaluation protocols in Appendix B. Appendices C–E provide external comparison details, additional controlled experiments, and extended results, respectively; Appendix F describes distillation.

4.1 Understanding

Table 1 evaluates global recognition, cross-modal matching, and dense prediction.

Global classification. ImageNet-1k zero-shot classification uses all 50K validation images and an 80-prompt ensemble; full-train linear probes (90 epochs, LARS; You et al., 2017) use the pooled representation g or average-pooled encoder features h. At equal data and training steps, CoRe reaches 50.63% zero-shot and 73.48% linear-probe accuracy on g, against 49.51% and 72.34% for the contrastive-only control. With Gram/iBOT, CoRe reaches 50.90% and 74.31%, respectively, both within 0.5 percentage points of the best results among the CC12M-trained references (COSMOS: 51.4% zero-shot; LaSLIP: 74.4% linear probe).

Retrieval and compositional matching. COCO Karpathy-5K retrieval (Lin et al., 2014) evaluates retrieval in both directions between 5K images and 25K captions; we report R@1 in each direction and mean recall (mR), averaged over R@1, R@5, and R@10 in both directions. SugarCrepe (Hsieh et al., 2023) tests whether image–text similarity prefers a correct caption over its compositional perturbations, reporting accuracy averaged over seven official splits. CoRe reaches 70.34% mR on COCO retrieval and 80.6% accuracy on SugarCrepe, compared with 69.75% and 79.4%, respectively, for the contrastive-only control. Together with the classification results, these comparisons suggest that adding generation preserves global recognition and image–text matching performance.

Dense spatial prediction. To test whether local semantic information remains accessible to a linear classifier, a linear 1×1 head on frozen encoder features h or contextual representations Z is trained on full-resolution ADE20K (Zhou et al., 2017) and scored on its 2K validation images. Compared with contrastive-only training, joint learning lowers mIoU by 1.6 on h (36.12 to 34.53) and by 5.6 on Z (29.04 to 23.42). Section 5 shows that a simple nonlinear MLP probe closes most of the gap on Z, suggesting that much of the relevant information remains present but is organized differently.

4.2 Image-to-text generation

Table 2 evaluates captions on three axes: caption quality and hallucination, image-specific discrimination, and detailed coverage. Our non-distilled models generate one full caption per image using a fixed 64-step DDPM sampling protocol.

Correctness and hallucination. We evaluate captions on COCO Karpathy-5K, with five reference captions per image. SPICE (Anderson et al., 2016) measures semantic agreement with the references.

Table 1: **Understanding** (higher is better). g: pooled representation; h: encoder features; Z: contextual representations. ZS: zero-shot classification; LP: linear probing; I→T and T→I: image-to-text and text-to-image retrieval R@1; mR: mean recall over R@1/5/10 in both directions. Classification, retrieval, and SugarCrepe results are percentages; ADE20K reports mIoU. *Inherited*: pretrained components the system uses. CC12M-R: VLM-recaptioned CC12M (LLaVA-NeXT for ours, DreamLIP for COSMOS); CC12M-L: LLM-rewritten captions. Published LP/ADE results use each model's own probe protocol. [a]DREAM-B at σ=.45. [b]Published CC12M CLIP of Fan et al. (2023) / released SLIP-repository CC12M CLIP (Mu et al., 2022) evaluated under our protocol; retrieval and SugarCrepe results are from the latter. FLUID-L as reimplemented by DREAM. ⋆Equal to the contrastive-only encoders. ♯Final fine-tune, evaluated at 512^2 with 1.2M additional positional parameters; the resolution differs from the 256^2 rows, so the higher zero-shot accuracy cannot be attributed to fine-tuning alone.

				ImageNet-1k			COCO retrieval			ADE20K		Sugar-
Model	Params	Inherited	Data	ZS↑	LP g ↑	LP h ↑	I→T↑	T→I↑	mR↑	h ↑	Z ↑	Crepe↑
CC12M-trained references												
LaSLIP-B/16 (Fan et al., 2023)	150M	none	CC12M-L	49.7		74.4						
COSMOS-B/16 (Kim et al., 2025)	150M	none	CC12M-R	51.4			64.2	48.9	74.6			
CLIP-B/16[b] (Radford et al., 2021)	150M	none	CC12M	40.2 / 37.41		70.3	32.90	22.50	48.51			72.1
SLIP-B/16 (Mu et al., 2022)	150M	none	CC12M	41.71			37.58	26.78	53.14			72.1
FLUID-L (Fan et al., 2025)	∼5.4B	T5-XXL, SD-VAE	CC12M			48.1				22.1		
DREAM-B (Li et al., 2026)	∼5.2B	T5-XXL, SD-VAE	CC12M			69.5[a]						
DREAM-L (Li et al., 2026)	∼5.5B	T5-XXL, SD-VAE	CC12M			72.7				36.8		
Controls (same training data and encoder architecture)												
Contrastive only	143M	none	CC12M-R	49.51	72.34	72.51	60.16	41.75	69.75	36.12	29.04	79.4
+ Gram/iBOT	143M	none	CC12M-R	50.63	73.98	74.06	61.90	43.65	70.87	38.74	31.99	79.6
No contrastive loss	383M	none	CC12M-R	1.18	70.22	71.72	0.40	0.19	1.51	33.89	22.28	53.8
Frozen contrastive → generation	383M	none	CC12M-R	49.51⋆	72.34⋆	72.51⋆	60.16⋆	41.75⋆	69.75⋆	36.12⋆	29.04⋆	79.4⋆
CoRe	383M	none	CC12M-R	**50.63**	**73.48**	**72.82**	**60.78**	**43.12**	**70.34**	**34.53**	**23.42**	**80.6**
CoRe + Gram/iBOT	383M	none	CC12M-R	**50.90**	**74.31**	**73.62**	**61.08**	**44.06**	**70.88**	**36.63**	**24.05**	**80.0**
+ fine-tuning (final)♯	384M	none	CC12M-R	51.49	–	–	62.84	45.34	72.02	37.30	22.25	80.6

Table 2: **Image-to-text generation.** Our non-distilled models use one sample per image with 64-step DDPM (128 NFE; NFE: denoiser network evaluations); distilled models use the budgets shown in their row labels. CHAIR_i and TrueMatch R@1 are percentages; TrueMatch reports results for bags of 3/5/7 images. [a]BLIP decoding uses beam size 3 and at most 20 tokens. [b]Reranks 64 sampled captions with CLIP. [c]COCO-finetuned open_clip release (Ilharco et al., 2021). [d]Trained on 5.4B generated annotations; <CAPTION> on COCO, <DETAILED_CAPTION> elsewhere. ♯Final fine-tune with image inputs at 512^2. CIDEr-D is length-sensitive (ours: 43 words on average; reference captions: about 10), and CoRe scores substantially below autoregressive captioners on this metric.

			COCO 5K				TrueMatch	DetailCaps	CompreCap	DOCCI
Model	Params	Inherited	SPICE↑	CIDEr-D↑	CHAIR_i↓	RefCLIP-S↑	R@1 3/5/7↑	CAPTURE↑	unified↑	CAPTURE↑
External captioners										
BLIP-base[a] (Li et al., 2022a)	224M	ViT-B, BERT	17.53	94.39	3.2	.808	42.9/28.7/25.8	23.03	33.56	25.83
NDLB (Gaur et al., 2024)	244M	CLIP-B/32, GPT-2					67.7/58.7/57.9			
MAGVLT-B (Kim et al., 2023)	447M	VQ-GAN	14.3[b]							
CoCa-L/14[c] (Yu et al., 2022)	638M	none (LAION-2B)	23.45	123.05	3.0	.832	52.2/40.4/38.2	28.81	38.70	30.89
Florence-2-base[d] (Xiao et al., 2024)	230M	DaViT, BART	24.66	133.02	2.6	.834	53.1/41.9/42.2	46.92	57.72	48.13
LLaVA-1.5-7B (Liu et al., 2024a)	7.1B	Vicuna, CLIP-L						49.44		
Cambrian-1-8B (Tong et al., 2024)	∼11.0B	Llama-3, 4 enc.					60.6/53.3/53.1			
InstructBLIP-13B (Dai et al., 2023)	14.2B	Vicuna, EVA-ViT-g					53.7/42.7/42.1			
LLaVA-NeXT-34B (Liu et al., 2024b)	∼34.8B	Yi, CLIP-L					57.9/46.9/47.9		58.64	
Controls										
No contrastive loss	383M	none	13.86	8.66	10.1	.741	57.0/49.0/47.6	44.43	42.56	45.84
Frozen contrastive → generation	383M	none	10.06	3.63	16.5	.736	58.3/46.3/43.9	43.73	41.11	44.46
CoRe	383M	none	**13.53**	**9.10**	**11.2**	**.740**	**58.9/53.3/48.4**	**43.62**	**42.84**	**45.66**
CoRe + Gram/iBOT	383M	none	**13.50**	**9.13**	**11.2**	**.740**	**60.6/51.9/47.9**	**44.15**	**41.72**	**45.52**
+ fine-tuning (final)♯	384M	none	13.91	9.45	10.4	.743	–	–	–	–
distilled, 1 jump (2 NFE)	383M	none	12.05	6.69	11.8	.732	59.7/46.7/45.5	44.41	41.04	45.57
distilled, 8 jumps (9 NFE)	383M	none	13.66	11.32	11.9	.737	58.9/50.0/48.4	44.86	42.19	46.54

CHAIR_i (Rohrbach et al., 2018) measures the percentage of object mentions that are hallucinated. RefCLIPScore (Hessel et al., 2021) measures agreement with both the image and the references using CLIP embeddings. CoRe reaches SPICE 13.53 with 11.2% hallucinated object mentions; its SPICE score trails several captioners trained on larger or COCO-specific caption datasets. The final fine-tuning recipe reaches SPICE 13.91 and reduces hallucinated object mentions to 10.4%.

Image-specific detail. Reference agreement can reward generic captions, so TrueMatch (Gaur et al., 2024) asks a frozen CLIP ViT-B/32 listener to pick each caption's source image from a bag of 3, 5, or 7 visually similar images. CoRe reaches R@1 of 58.9%, 53.3%, and 48.4%, respectively, above

Table 3: **Text-to-image generation and reconstruction.** Our models use 256^2 unless marked. For our non-distilled models, GenEval/DPG use 128 steps (257 NFE), and FID uses 64 steps (128 NFE); distilled models use the budgets shown in their row labels. GenEval is reported on a 0–1 scale and DPG on a 0–100 scale. PSNR is in dB; rFID denotes reconstruction FID. [a]CLIP reranking of 32 samples. [b]FID from the CC12M-pretraining checkpoint at 512^2, before the 120K fine-tune. [♯]Final fine-tune at 512^2.

				Prompt following		FID		Reconstruction	
Model	Params	Inherited	Data	GenEval↑	DPG↑	COCO-30K↓	CC12M-50K↓	PSNR↑	rFID↓
External systems									
SD1.5 (Rombach et al., 2022)	1.1B	CLIP-L, VAE	LAION	.43	63.18				
SDXL (Podell et al., 2024)	3.5B	CLIP-L/bigG, VAE	internal	.55	74.65				
MAGVLT-B (Kim et al., 2023)	447M	VQ-GAN	CC3M/12M+			12.08[a]			
MiniT2I-B/32 (Wang et al., 2026)	601M	FLAN-T5-Large	CC12M-R	.505	78.3	13.69			
MiniT2I-B/16 (Wang et al., 2026)	599M	FLAN-T5-Large	CC12M-R+120K	.873	84.2	10.51[b]			
FLUID-L (Fan et al., 2025)	∼5.4B	T5-XXL, SD-VAE	CC12M			9.62	4.53		
DREAM-L (Li et al., 2026)	∼5.5B	T5-XXL, SD-VAE	CC12M			10.0	4.34		
SD-VAE (Rombach et al., 2022)	84M	–	autoencoder					25.04	3.05
SDXL-VAE (Podell et al., 2024)	84M	–	autoencoder					26.10	2.95
Controls									
No contrastive loss	383M	none	CC12M-R	.193	45.31	23.20	9.73	26.91	4.76
Frozen contrastive → generation	383M	none	CC12M-R	.400	70.12	12.44	7.99	19.26	11.58
CoRe	383M	none	CC12M-R	**.461**	**74.43**	**10.10**	**4.33**	**24.09**	**6.43**
CoRe + Gram/iBOT	383M	none	CC12M-R	**.448**	**73.49**	**10.20**	**4.37**	**24.03**	**6.24**
+ fine-tuning (final)[♯]	384M	none	+120K	**.729**	**77.79**	–	13.84	–	–
distilled, 1 jump (2 NFE)	383M	none	CC12M-R	.317	67.38	16.06	9.53	–	–
distilled, 8 jumps (9 NFE)	383M	none	CC12M-R	.428	72.09	12.45	6.07	–	–

LLaVA-NeXT-34B at every bag size and close to the much larger Cambrian-1-8B, which scores higher at sizes 3 and 7 and ties at size 5. NDLB is explicitly trained for caption self-retrieval and achieves higher scores.

Detailed coverage. DetailCaps-4870 (CAPTURE against three detailed references; Dong et al., 2024), CompreCap (object, attribute, and relation coverage with a fixed Llama-3-8B judge; Lu et al., 2025), and DOCCI (CAPTURE against one long human description; Onoe et al., 2024) ask whether a caption covers fine-grained content. CoRe scores 43.62, 42.84, and 45.66, respectively, above BLIP and CoCa across all three benchmarks. It trails Florence-2, trained on 5.4B generated annotations, and the reported results of much larger 7B–34B multimodal models with pretrained language and vision components.

4.3 Text-to-image generation

Table 3 evaluates prompt following, image-distribution quality, and image reconstruction.

Prompt following. GenEval (Ghosh et al., 2023) checks single objects, object pairs, counting, colors, relative position, and color attribution over 553 prompts with an object detector; DPG-Bench (Hu et al., 2024) evaluates entities, attributes, and relations through VQA questions over 1065 dense prompts. For both benchmarks, our non-distilled models generate four samples per prompt using a fixed sampler (128 steps, CFG 3), with prompts encoded by our own text encoder trained only on the training captions. Without fine-tuning, CoRe reaches .461 on GenEval and 74.43 on DPG, above SD1.5 on both and comparable to SDXL on DPG. It trails MiniT2I-B/32 (.505 and 78.3), which uses a pretrained FLAN-T5-Large text encoder. The final fine-tuning recipe raises CoRe + Gram/iBOT from .448 to .729 on GenEval and from 73.49 to 77.79 on DPG at 512^2, without sacrificing the first-stage model's understanding or captioning, since the second stage updates only the image denoiser.

Distribution quality. Single-sample FID compares one generation per prompt with real images: zero-shot on COCO-30K (one caption per validation image) and in-distribution on 50K CC12M training images with their original alt-text. The FID sampler is selected once on disjoint COCO development prompts (64 steps, 128 NFE). CoRe reaches FID 10.10 on COCO-30K and 4.33 on CC12M-50K, close to the reported 10.0 and 4.34 of DREAM-L, which uses a frozen T5-XXL encoder and an SD-VAE.

Recoverability. We encode 5K ImageNet validation images, add noise at the sampling schedule's lowest noise level, and reconstruct each image with one denoiser call conditioned on its own contextual representation. CoRe reaches 24.09 dB PSNR, within 1 dB of SD-VAE (25.04 dB). The no-contrastive-



An owl perched on a branch at night under a full moon | An oil painting of a lighthouse on a rocky coast during a storm | A watercolor painting of cherry blossoms along a quiet river | A blue ceramic mug to the left of a stack of old books | A small wooden cabin beside a lake surrounded by autumn trees, reflected in the water | A hot air balloon floating over rolling green hills | A purple butterfly resting on a sunflower | A desert landscape with sand dunes and a single cactus

Figure 2: **Text-to-image samples from CoRe** (final fine-tune, 512^2; sampler selected on held-out prompts). For each prompt, one of four seeds is chosen by the model's *own* image–text score, not an external judge; all seeds of all 24 prompts, including failures, are in Figure 6.

loss control reconstructs more accurately (26.91 dB) but follows prompts less accurately (GenEval .193), a contrast we analyze in Section 5.

4.4 Few-step generation

Because generation runs in a continuous contextual representation space, we can distill both image and text generation into a few steps while keeping the encoders fixed. We distill the CoRe + Gram/iBOT denoisers into shortcut students that jump between noise levels (Frans et al., 2025; Boffi et al., 2024), anchoring each student's final pixel or token predictions to the teacher's to preserve output quality. With eight jumps (9 NFE), the image student reaches GenEval .428, retaining approximately 96% of the 257-NFE teacher's score of .448 (Table 3). The eight-jump caption student retains the teacher's performance across all three detailed-caption benchmarks while reducing computation from 128 to 9 NFE (Table 2). On one H100 at batch size 1, image generation takes 173 ms instead of 4.75 s for the teacher, and caption generation takes 79 ms, excluding source encoding.

Summary. Together, these results show that CoRe learns contextual representations that preserve contrastive-level global understanding while supporting both image and text generation. On several benchmarks, it matches or exceeds larger baselines and systems that rely on pretrained components. Joint learning improves generation over frozen contrastive representations, and distillation accelerates both generation directions without changing the encoders.

5 How does joint learning shape representations?

Section 4 shows that one contextual representation can support understanding, conditioning, and generation. We now examine how joint learning shapes this representation, distinguishing *retention*—what task-relevant information remains recoverable—from *organization*—how that information is arranged and accessed.

Generating in the understanding representation improves compositionality. We compare two matched continuations of the same checkpoint (Table 4). In the *shared* arm, image generation uses the contextual representation Z_v used for understanding; in the *separate* arm, it uses a second projection of the same visual backbone, initialized as a copy of the original, while understanding continues to use Z_v. Although the separate representation reconstructs images better (25.03 vs. 23.38 dB), the shared representation achieves better compositional generation (GenEval .428 vs. .362; DPG 71.30 vs. 65.91), with similar global understanding. Better reconstruction therefore does not necessarily yield better conditional generation: generating in the representation used for understanding benefits the composition of text and visual concepts.

A small nonlinear readout recovers most of the dense-prediction gap. The clearest tension appears in dense visual prediction. On ADE20K, the jointly learned contextual representation performs substantially worse than the contrastive-only control under a linear readout, but a small MLP reduces the gap from 8.67 to 0.99 mIoU, with the same pattern under Gram/iBOT (Figure 3). The representations remain frozen throughout, indicating that much of the dense information remains recoverable through a simple nonlinear readout. The shared-versus-separate comparison shows the

Table 4: **Shared versus separate generation space.** Both arms continue the same 24K-step checkpoint for 16K steps, generating in either the understanding representation Z_v or a separate projection initialized as its copy. ADE20K uses the reduced-resolution diagnostic protocol.

Generated in	ZS	mR	SPICE	GenEval	DPG	PSNR	ADE lin.	ADE MLP
Shared (CoRe)	44.65	69.15	12.01	**.428**	**71.30**	23.38	15.4	26.0
Separate	44.66	69.42	12.13	.362	65.91	**25.03**	**21.9**	26.6

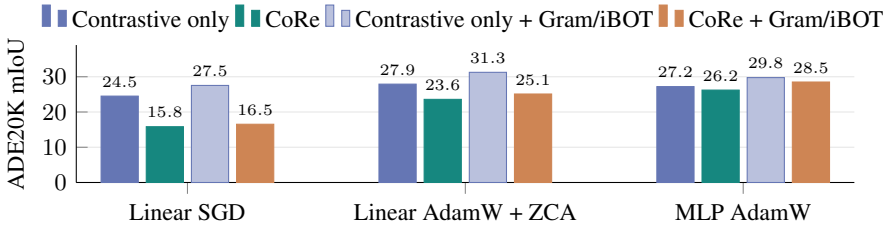

Figure 3: **Fixed representations, different readouts.** ADE20K mIoU of frozen contextual representations under a reduced-resolution diagnostic protocol. The gap to the contrastive-only control falls from 8.67 with a linear probe to 0.99 with a small MLP, and from 11.0 to 1.25 with Gram/iBOT on both models.

same pattern: a large linear-probe gap nearly disappears with an MLP (Table 4). These results locate much of the apparent tension in how information is accessed rather than whether it is retained.

Joint learning reorganizes visual and textual neighborhoods differently. We measure neighborhood *enrichment*: the fraction of a token's ten nearest neighbors sharing a label, divided by the corresponding fraction expected under random sampling. From the contrastive-only control to CoRe, same-image enrichment among visual patches falls from $83.3\times$ to $6.5\times$; same-class enrichment among patches from other images falls from $9.0\times$ to $2.0\times$. Text instead preserves local lexical structure: same-word and same-family enrichment across captions remain nearly unchanged ($51.4\times$ to $50.8\times$ and $5.88\times$ to $5.80\times$, respectively; same-family neighbors exclude the same word). Connecting each text token to its 12 nearest neighbors yields an undirected graph whose largest connected component contains, on average, 97% of sampled tokens for CoRe, compared with 68% for the contrastive-only control. Contextual meanings also remain distinguishable: leave-one-out nearest-neighbor classification within each word achieves 98.8% balanced sense accuracy, averaged over six polysemous words. Joint learning therefore produces different organizational changes across modalities: less source-centric visual neighborhoods, and preserved local lexical structure within a more connected text representation.

6 Conclusion

CoRe shows that understanding and generation can operate through the same contextual representation. For each modality, one representation is read for understanding, conditions generation of the other modality, and serves as the space for its own generation. In a controlled CC12M setting, with no pretrained language model, text encoder, or VAE in the inference model, CoRe matches same-data contrastive controls on global understanding while supporting competitive image generation. Prompt following and caption fidelity remain below systems that inherit stronger text encoders or substantially more caption data. More importantly, controlled comparisons show that generation benefits from jointly learning the representation and generator, that using the same representation for understanding and generation improves compositionality, and that most of the apparent dense-prediction gap can be recovered by a more flexible readout.

Limitations and outlook. Our goal is to isolate this representation-level question rather than to scale the system. Our text encoder sees only CC12M captions, which makes counting and spatial prompts hard. Larger scale and representation alignment (Yu et al., 2024) are compatible with the construction; we hope this setting helps to study what a representation must retain, and how it must be organized, to both understand and create.

AI USE STATEMENT

We used generative AI tools to help write code, and to polish writing. We did not use them to generate data sets. We also used these tools for broader literature discovery. The authors reviewed all AI-assisted work. Every reported number was checked against the evaluation outputs that produced it, citations against their publisher records, and code by running it. We take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Architecture and training details are provided in Appendix A, and evaluation protocols in Appendix B. Appendices C–E provide external comparison details, additional controlled experiments, and extended results, respectively; Appendix F describes distillation, Appendix G iterative refinement, and Appendix H the representation measurements of Section 5.

REFERENCES

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

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. 2022.

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

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. 2025:55458–55472, 2025d.

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. Langflow: Continuous diffusion rivals discrete in language modeling. *arXiv preprint arXiv:2604.11748*, 2026.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. volume 36, pp. 49250–49267, 2023.

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

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

Caption Emporium. conceptual-captions-cc12m-llavanext. `https://huggingface.co/datasets/CaptionEmporium/conceptual-captions-cc12m-llavanext`, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. volume 36, pp. 35544–35575, 2023.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. In *International Conference on Learning Representations*, volume 2025, pp. 100218–100231, 2025.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Manu Gaur, Darshan Singh, and Makarand Tapaswi. No detail left behind: Revisiting self-retrieval for fine-grained image captioning. *arXiv preprint arXiv:2409.03025*, 2024.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. 2018.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. volume 36, pp. 52132–52152, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

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*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. volume 36, pp. 31096–31116, 2023.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23115–23127. IEEE, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip, 2021.

Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, and Zeynep Akata. Cosmos: Cross-modality self-distillation for vision language pre-training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14690–14700. IEEE, 2025.

Sungwoong Kim, Daejin Jo, Donghoon Lee, and Jongmin Kim. Magvlt: Masked generative vision-and-language transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23338–23348. IEEE, 2023.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. pp. 18262–18272, 2025.

Chao Li, Tianhong Li, Sai Vidyaranya Nuthalapati, Hong-You Chen, Satya Narayan Shukla, Jianpeng Cheng, Yonghuan Yang, Jun Xiao, Xiangjun Fan, Aashu Singh, et al. Unifying contrastive and generative objectives for visual understanding and text-to-image generation. *arXiv preprint arXiv:2603.02667*, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022a. URL `https://proceedings.mlr.press/v162/li22n.html`.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. pp. 36115–36125, 2026.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2142–2152, 2023.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. volume 35, pp. 4328–4343, 2022b.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2779–2790. IEEE, 2025.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19618–19627. IEEE, 2025.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. volume 38, pp. 129274–129297, 2026a.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. Pixelgen: Improving pixel diffusion with perceptual supervision. *arXiv preprint arXiv:2602.02493*, 2026b.

Viacheslav Meshchaninov, Alexander Shabalin, Egor Chimbulatov, Nikita Gushchin, Ilya Koziev, Alexander Korotin, and Dmitry Vetrov. How to train your latent diffusion language model jointly with the latent space. *arXiv preprint arXiv:2605.07933*, 2026.

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

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pp. 291–309. Springer, 2024.

OpenDatasets. dalle-3-dataset. Hugging Face dataset, 2024. URL `https://huggingface.co/datasets/OpenDatasets/dalle-3-dataset`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10609–10619. IEEE, 2022.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2545–2555. IEEE, 2025.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. volume 37, pp. 87310–87356, 2024.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. volume 36, pp. 46830–46855, 2023.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pp. 23318–23340. PMLR, 2022.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. volume 37, pp. 121475–121499, 2024a.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. Minit2i: A minimalist baseline for text-to-image generation, 2026. URL `https://peppaking8.github.io/#/post/minit2i`.

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

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829. IEEE, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. 2020.

Jinbo Yan, Limeng Qiao, Jie Qin, Junyan He, Feize Wu, and Guanglu Wan. Unispace: Unified visual representation and scalable multimodal modeling. *arXiv preprint arXiv:2608.08676*, 2026.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. volume 38, pp. 138867–138907, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. 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? 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. 2026:35791–35820, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. 2025:6446–6469, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. 2021.

Appendix overview. Each section answers one question raised by the main text. A: how the model is built and trained, and what it costs. B: how each benchmark is evaluated. C: how the external systems compare in components, data, and protocol. D: which interventions establish the causal claims (conditioning routes, objective balance, shared versus separate generation space, fine-tuning). E: the full capability results behind the three main tables. F: few-step distillation. G: iterative refinement through the text encoder. H: the representation measurements of Section 5.

A Architecture, training, and resource accounting

Encoders and the contextual representation. The visual tower is a 12-layer ViT (width 768, 12 heads) on 16×16 patches of images in $[0, 1]$, with no mean/std normalization and learned absolute positions (no class token). The text tower has 12 layers, width 512, and eight heads over a learned embedding table; token identities follow the T5 SentencePiece tokenizer (Raffel et al., 2020), whose 32,100 entries are padded to a 32,128-way embedding and output layer. Blocks are pre-norm with GELU MLPs (ratio 4) and QK-RMSNorm. Each tower projects to width 256 and applies non-affine LayerNorm, giving Z_m. Eight learned queries attention-pool Z_m; their outputs are averaged and layer-normalized to the 512-d global readout g_m. Alignment uses the L2-normalized g_m; conditioning uses the layer-normalized output. Padding is excluded from encoder attention, pooling, cross-attention keys, and the token cross-entropy; the text denoiser's self-attention is unmasked, so pad slots are part of the language trajectory.

Denoisers and heads. The visual denoiser has width 768, 12 blocks, and 12 heads; the language denoiser has width 512, six blocks, and eight heads. Both use adaLN-zero modulation from an MLP of a sinusoidal embedding of $\gamma/4$ plus the projected global condition, and cross-attention to the local condition. The visual observable head B_v reshapes the final 16×16 grid of trunk features and renders RGB with a small convolutional network: a 3×3 convolution to 128 channels, GroupNorm, SiLU, and a residual block, then four stages of nearest $\times2$ upsampling with convolution and residual blocks ($128\to64\to32\to32\to32$ channels), a final convolution to three channels, and a sigmoid (1.42M parameters). The text observable head is an untied linear layer to vocabulary logits. The representation heads F_m are sibling linear layers to width 256. The tensor in which image generation runs is the same Z_v used for local representation readout.

Language length and termination. A caption occupies a fixed canvas of 256 token slots: the tokenizer appends an end-of-sequence token and pads with zeros. All slots are noised and denoised in parallel; the cross-entropy covers valid tokens and the end token, and the representation regression covers all slots. At inference, each slot is read by argmax and everything after the first end token is discarded; captions without an end token are flagged. The text ladder terminates at $\gamma \approx 4.3$, and the token logits are taken from the conditional branch of the last denoiser call; guidance acts on the clean-representation estimate only.

Complete objective. The implemented base objective is Equations 5–6, i.e. $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{img}} + 0.5\,\mathcal{L}_{\text{text}}$ with representation-regression weights 0.25 (image) and $0.5 \times 0.25 = 0.125$ (text), plus the text-schedule fit $0.1\,\mathcal{L}_{\text{sched}}$ on detached losses, which updates only the schedule parameters. $\mathcal{L}_{\text{representation},m}$ is a mean over batch, positions, and channels. SigLIP uses a learnable temperature initialized at $1/0.07$ (capped at 100) and bias initialized at -10, with negatives gathered across devices. Clean-representation targets are detached; the noisy input and source condition carry gradients into the encoders. Schedule fitting uses detached losses. The LPIPS term uses torchvision VGG16 features at the five LPIPS taps with channel unit-normalization and a spatial mean, averaged over taps and without LPIPS's learned channel weights; its weight 0.1 therefore corresponds to a rescaled, uncalibrated LPIPS. The P-DINO term averages the patchwise cosine distance between last-layer DINOv2-B features of predicted and target images (an 18×18 grid at 256 pixels). The perceptual weights (0.1, 0.01) and gate ($t \geq 0.3$, i.e. $\gamma < 1.695$) follow PixelGen (Ma et al., 2026b). Both perceptual terms are evaluated on a subsample of 1024 images per training step (512 during fine-tuning).

The visual noise proposal is a truncated Gaussian $\gamma \sim \mathcal{N}(1.6, 1.6^2)$ on $[-12, 12]$, i.e. a logit-normal time distribution $\text{logit}(t) \sim \mathcal{N}(-0.8, 0.8^2)$ as in SD3, PixelGen, and JiT (Esser et al., 2024; Ma et al., 2026b; Li & He, 2026); it is fixed. The language proposal is LangFlow's learned truncated-Gumbel

loss-curve schedule (Chen et al., 2026), initialized at location 0 and scale 2; γ is drawn by stratified inverse-CDF sampling. Self-conditioning is disabled in the base models; its projection stays exactly zero. The frozen-encoder self-conditioning continuations are specified in Appendix G.

Optimization and data. Training uses the CC12M images (Changpinyo et al., 2021), obtainable from the dataset's official URLs. Each image carries its alt-text and two synthetic captions from a public release (Emporium, 2024): a long description generated by LLaVA-NeXT (Llama-3-8B) and a short version of at most 30 words condensed by Llama-3-8B. Each training example draws one of the three captions with probabilities 0.45/0.35/0.20 (long/short/alt-text), renormalized over the captions present; MiniT2I trains on the long captions only, so our caption mix differs from its setting. AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), learning rate 1.4×10^{-3} with 4K linear warm-up and cosine decay to zero, gradient clipping at 1.0, EMA decay 0.999, global batch 8192, 125K training steps. The effective weight decay on encoder and denoiser matrices is 0.2; biases, normalization parameters, the patch embedding, the token embedding, positional tables, null codes, and the SigLIP scale/bias are excluded. Random resized crops use area $[0.5, 1]$; horizontal flips are gated by caption compatibility.

Structural intervention. CoRe + Gram/iBOT resumes the joint run at training step 24K (optimizer, schedule, and EMA included) and continues to 125K with two added terms on backbone patch features: Gram anchoring toward the model's own 24K weights (weight 2) and an iBOT-style masked self-distillation loss with an EMA teacher (weight 1; 8192 prototypes, temperatures 0.1/0.04, block masking of 10–50% of patches, a separate AdamW at 10^{-3}) (Siméoni et al., 2025; Zhou et al., 2021). The iBOT head (8.4M parameters) and teacher are dropped at inference. Both terms act at full weight from training step 24,001, without warm-up.

Fine-tuning. Every fine-tune starts from the CoRe + Gram/iBOT EMA weights with a fresh optimizer, with Gram and iBOT off. The synthetic data is MiniT2I's 120K mixture (Wang et al., 2026): BLIP3o-60k (Chen et al., 2025a), DALL·E 3 images (OpenDatasets, 2024), and ShareGPT-4o-Image images (Chen et al., 2025b), each with its captions.

Final fine-tune at 512^2 (the $^\sharp$ rows of the main tables). The positional tables are interpolated bicubically to a 32×32 grid (1024 image tokens; 384M parameters), and the visual noise proposal is shifted by $\ln 4$ in log noise-to-signal ratio for the fourfold token count, as in SD3 (Esser et al., 2024). A first stage trains all parameters for 10K steps on CC12M-R images whose short side is at least 512 pixels (batch 5120, learning rate 1.1×10^{-4}, 250 warm-up steps). The second stage then trains only the image denoiser and its null codes on the 120K mixture for 8K steps (batch 5120, learning rate 1.1×10^{-4}, 100 warm-up steps), with both encoders, the text denoiser, and the text null codes frozen and the alignment loss switched off. Because nothing that understanding or captioning reads is updated, both are measured once on the first-stage model at its native 512^2 input.

Controlled fine-tunes at 256^2 (Appendix D.4). 5000 training steps on the 120K mixture at batch 8192, i.e. MiniT2I-B/16's sample budget (41M pair presentations, about 340 passes over the mixture), learning rate 1.4×10^{-4} with 20 warm-up steps and cosine decay, EMA decay 0.999 (the evaluated EMA retains 0.7% of the initial weights). One run trains all parameters; the other freezes both encoders, switches the alignment loss off, and trains the two denoisers and their null codes.

A.1 Parameters and external resources

At inference the model uses only its own encoders and denoisers. Training additionally uses frozen VGG16 and DINOv2-B for decoded-image supervision (Simonyan & Zisserman, 2014; Oquab et al., 2023); these are training-only components, but they are not detached from the representation: their gradients pass through the image denoiser into the noisy representation, which stays attached to the image encoder (weights 0.1 and 0.01, only at $\gamma < 1.695$). The T5 tokenizer defines token identities, but the embedding table and text encoder are learned from scratch. Synthetic captions carry knowledge from the LLaVA-NeXT captioner that produced them; we use the released captions (Emporium, 2024) as they are and did not re-audit them against the evaluation sets. Evaluation uses separate pretrained scorers—Inception for FID, DINOv2 and CLIP for feature distances, CLIP ViT-B/32 for TrueMatch and CLIPScore, Mask2Former for GenEval, mPLUG for DPG, and parsers, matchers, and a Llama-3-8B judge for detailed captions—only for scoring (Appendix B).

Table 5: **Parameter accounting for the full base model.** Counts in millions. The renderer is part of the visual denoiser. Training-only components (perceptual networks, iBOT head and teacher, Gram anchor) and evaluators are excluded.

Component	Parameters (M)
Visual tower, projection, and pooling	87.229
Text tower, projection, and pooling	55.731
Visual denoiser and heads	182.321
of which convolutional renderer	1.423
Text denoiser and heads	57.845
Total, including null codes, SigLIP scale/bias, and schedule parameters	383.128
Total at 512^2 (1024-entry positional tables in the visual tower and denoiser)	384.307

Table 6: **Training runs and exposure.** Every model we report, with what it trains and on what. Pair presentations count repeated examples.

Model	Initialization; trained parameters	Data	Steps × batch	Presentations
CoRe; contrastive only; no contrastive loss	scratch; all (encoders only for contrastive only)	CC12M-R, 256^2	125K × 8192	1.024B
CoRe + Gram/iBOT	CoRe at 24K; all, plus Gram/iBOT terms	CC12M-R, 256^2	101K × 8192	1.024B in total
Frozen contrastive → generation	contrastive-only encoders; denoisers	CC12M-R, 256^2	125K × 8192	2.048B in total
Final fine-tune, stage 1	CoRe + Gram/iBOT; all	CC12M-R, 512^2	10K × 5120	+51.2M
Final fine-tune, stage 2	stage 1; image denoiser and its null codes	120K mixture, 512^2	8K × 5120	+41.0M
Controlled fine-tunes (App. D.4)	CoRe + Gram/iBOT; all, or denoisers only	120K mixture, 256^2	5K × 8192	+41.0M
Shared / separate generation space (App. D.3)	CoRe at 24K; all	CC12M-R, 256^2	16K × 8192	+131M each

Our base training sees about 1.85 times DREAM's pair presentations, on recaptioned CC12M; unique data and exposure are different resource dimensions (Table 10). The visual denoiser costs 69.14 GFLOPs per forward pass at 256^2 (MiniT2I-B/32: 313). Wall-clock costs on one H100 are in Table 25.

Training compute. Table 7 counts model FLOPs (forward and backward, per training example at 256^2; `torch FlopCounterMode` on each checkpoint's own configuration, matrix multiplies, convolutions and attention, no activation recomputation). The encoders and the alignment loss cost the same 205.8 GFLOPs in every arm; CoRe's additional compute goes to the two denoisers (209.8 and 70.3) and the frozen perceptual networks (38.0), not to the encoders that understanding evaluates. A contrastive-only model would need about 318K steps instead of 125K to match CoRe's training FLOPs; we did not train such a FLOP-matched control.

Table 7: **Training FLOPs per example** (GFLOPs, forward + backward, 256^2). Perceptual terms run on one eighth of each batch, Gram anchoring and iBOT on one quarter; the costs are averaged over the batch.

Model	Encoders + SigLIP	D_v	D_t	VGG16 + DINOv2	Gram + iBOT	Total	Ratio
Contrastive only	205.8	–	–	–	–	205.8	1.00
CoRe	205.8	209.8	70.3	38.0	–	523.9	2.55
Contrastive only + Gram/iBOT	205.8	–	–	–	62.8	268.5	1.30
CoRe + Gram/iBOT	205.8	209.8	70.3	38.0	62.8	586.6	2.85

Which losses move the image encoder. Table 8 splits one training step's gradient on the image encoder (all 238 parameter tensors of E_v, P_v, A_v) by loss term, for CoRe + Gram/iBOT at 125K steps with its raw weights: two emulated training steps of 8192 real training pairs each, the trainer's loss weights, and training-distributed noise levels (the perceptual gate fired on 54% of the supervised images). The DINOv2 term does reach the encoder, with a nonzero gradient on every tensor, but carries 0.11% of its gradient norm (0.21% when the gate fires; 0.13% after Adam preconditioning with the checkpoint's own second moments) and is nearly orthogonal to the total. It acts mainly on the image denoiser, where it is 31% of the gradient norm. Near convergence the encoder's gradient is dominated by SigLIP and iBOT, and the direct image-generation terms (pixel MSE, VGG, and the clean-representation regression) together contribute about 5%. These are per-step shares at 125K steps, which are noise-dominated (the total gradient's cosine between the two steps is .12; a noise-free U-statistic estimate gives the same 0.12% DINOv2 share); they say nothing about early training, and the retention results rest on the comparisons across models, not on gradient shares.

Table 8: **Per-term gradient on the image encoder** (CoRe + Gram/iBOT, 125K steps; norm of each term's gradient as a fraction of the total gradient norm, and its cosine with the total; mean of two steps). Fractions sum to more than 100% because the terms partly cancel.

Loss term	Fraction of norm	Cosine with total
SigLIP alignment	92.4%	+.94
iBOT	33.3%	+.34
Text denoising (token cross-entropy, through cross-conditioning)	10.4%	+.20
Image clean-representation regression	2.6%	+.00
Gram anchoring	2.2%	−.01
Pixel MSE	1.4%	+.01
VGG perceptual	0.78%	−.02
Text clean-representation regression	0.48%	+.09
DINOv2 perceptual	0.11%	+.02

B EVALUATION PROTOCOLS

Each benchmark has one frozen protocol, applied unchanged to every row we evaluate; released external models go through the same scorers, image sets, and views.

B.1 UNDERSTANDING

ImageNet-1k. Zero-shot classification uses the entire 50K validation set, 1000 categories, and an 80-prompt ensemble. Full-train linear probing uses all 1,281,167 training images and evaluates on all 50K validation images. Three readouts are trained from the same frozen forward: average-pooled backbone h (768 dimensions), average-pooled contextual representation Z (256), and attention-pooled g (512, ending in LayerNorm). Following the DREAM/MAE recipe (Li et al., 2026; He et al., 2022), each head uses non-affine BatchNorm followed by a linear classifier, the LARS optimizer (SGD with momentum and a layer-wise adaptive learning rate; You et al., 2017) for 90 epochs, batch size 4096, base learning rate 0.1 scaled by batch/256, ten warmup epochs, and cosine decay. Random resized crops of size 256 are drawn from preprocessed 256-pixel image chunks. The primary score uses the last epoch. Inputs retain raw RGB $[0, 1]$ scale. Confidence intervals resample validation images; they characterize test-sample uncertainty for a fitted head.

Retrieval, compositional matching, and dense prediction. COCO retrieval uses the Karpathy test split: 5000 images and 25,000 reference captions. We report R@1/5/10 in each direction and their mean. SugarCrepe uses its seven full official splits (1652/788/1406/245/666/2062/692 items) and ARO the full VG-Attribution and VG-Relation splits (28,748/23,937 items) (Hsieh et al., 2023; Yuksekgonul et al., 2022); both score captions by global image–text similarity. The primary ADE20K evaluation uses full-resolution ADEChallengeData2016, with 20,210 training and 2000 validation images. The head is linear, as in DINOv3 (Siméoni et al., 2025): channel dropout (0.1), BatchNorm, and a 1×1 convolution, i.e. one 150-class linear classifier shared by all patch tokens; its logits are bilinearly upsampled to pixel resolution. It fits frozen h or Z for 40K iterations, batch size 16, and

512-pixel crops. Sliding-window evaluation uses crop/stride 512/341. Learning rate 10^{-4} follows DREAM's stated setting; 10^{-3} is a separately reported sensitivity test from the DINOv3 configuration (Li et al., 2026; Siméoni et al., 2025). The last iteration supplies the primary score.

The readout-accessibility diagnostic uses a twofold-downsampled ADE20K package, 8000 training images, 512-pixel input, and 40 epochs. It compares linear SGD, linear AdamW with training-only ZCA, and MLP AdamW on fixed features. This reduced-resolution study isolates readout effects; the primary full-resolution protocol supplies the external comparison.

B.2 Image generation and reconstruction

T2I sampling. GenEval and DPG use the native fitted visual ladder, 128 steps, CFG 3, and churn 1 at the model's native resolution (256^2; 512^2 for the final fine-tune). Each benchmark uses its prescribed prompts and seed multiplicity. Quantitative images receive no reranking. GenEval uses 553 prompts and four images per prompt, evaluated by its object-focused scorer (Ghosh et al., 2023). DPG uses 1065 prompts and four images per prompt with the mPLUG VQA evaluator (Hu et al., 2024). Overall DPG scores average the dependency-aware crop scores. Category columns use the last crop's category averages, while the overall score averages dependency-aware scores across all four crops. Each aggregation is reported under its released definition.

COCO FID uses 30K seeded sampled val2014 images, the first annotated caption for each image, and one generated image per prompt. Real references are the paired 30K images resized on the short side to 256 and center-cropped. Features come from pytorch-fid's Inception-2015 checkpoint (Heusel et al., 2017). FD-DINOv2 and FD-CLIP are feature-distance diagnostics with separate numerical scales. Real-versus-real estimates characterize finite-sample bias: for fixed-recipe runs a pessimistic half-split of the reference set (4.405 at 15K per side), and for development-selected runs a disjoint 10K-versus-30K draw (4.387). Perfectly matched feature distributions have zero FID.

Development-selected COCO recipe. A disjoint 5K development set selects among 64/128 steps, CFG peaks 2/3/4/5, and constant/linear guidance, with churn 1 throughout. The rule minimizes development FID, then maximizes CLIPScore among configurations within 0.2 FID of that minimum. This selects 64 steps with guidance increasing from 1 toward 5 linearly in noise-level progress (development FID 20.181; minimum 20.034). The core-model selection is applied unchanged to every other checkpoint, once on the 30K test prompts. Each prompt receives one generated image. The reported 128 NFE is exact: the first step runs at guidance weight 1 and therefore needs one branch, the remaining 63 steps need two, and one final conditional evaluation renders pixels ($1 + 2 \times 63 + 1$). End-to-end timing additionally includes the source encoder.

Sampler for the 512^2 models. A 54-recipe sweep (guidance 3–7, constant or rising, 32–128 steps, churn 0/1, three terminal noise levels) on a held-out GenEval-style development set—439 prompts from the official prompt generator with a different seed, every test prompt removed—run on a 512^2 fine-tune selects 64 steps, constant CFG 5, and churn 1; the top of the ranking is flat (CFG 5 and 6 within .001). On the test prompts it changes GenEval by less than .01 relative to the fixed recipe (final model: .737 vs. .729) and raises DPG by 1.6–2.0 (79.75 vs. 77.79). The tables keep the fixed recipe; the qualitative samples use the selected one.

One recipe per benchmark. Prompt-following benchmarks and FID therefore use different samplers, as in DREAM, which reports its best FID and CLIP score over constant and linear CFG schedules without stating a held-out selection set (Li et al., 2026); our FID recipe is chosen on disjoint development prompts and run once on the test prompts. The core model's choice is applied unchanged to every other checkpoint; re-selecting guidance per checkpoint would favor them slightly (development FID 19.86 at CFG 3 vs. 20.38 for + Gram/iBOT). Swapping the recipes costs each metric: the FID-selected recipe gives GenEval .430 and DPG 74.52 for CoRe (fixed recipe: .461 and 74.43), and the fixed recipe gives COCO-30K FID 13.87 (selected: 10.10).

Training-distribution image synthesis. CC12M evaluation draws 50K reference images and a disjoint 50K real control from our CC12M training images (CC12M has no held-out split; DREAM does not state whether its 50K are held out), using original alt-text prompts, as DREAM does, and the development-selected recipe. 512^2 generations are first brought to 256^2 with the references'

own transform (bicubic antialiased short-side resize and center crop). FID measures the training distribution; the disjoint real-versus-real estimate is 1.442.

Reconstruction. Reconstruction starts from the encoded image with low-noise corruption at $\gamma = -3.34$ (the visual ladder's terminal level), then predicts RGB with one denoiser call. Off-the-shelf VAEs are scored on the identical 5K images, preprocessing, and metrics, decoding their posterior mode. The 5K images are a fixed seeded draw from ImageNet-1k validation; each is resized on its short side to 256 (bicubic) and center-cropped to 256^2 in $[0, 1]$, with no training augmentation. PSNR is computed per image on $[0, 1]$ and averaged; rFID compares the 5K reconstructions with the same 5K preprocessed originals (Inception-2015). The main tables use the full-condition route, which supplies the same image's (g_v, Z_v) as global and cross-attention inputs in addition to the noised representation; the null route supplies learned null conditions. A 50K-image rFID is reported for CoRe and CoRe + Gram/iBOT. A stochastic contextual reconstruction and a deterministic autoencoder can be compared descriptively when sample counts and preprocessing are stated.

B.3 Image-to-text generation

A fixed development decision. The I2T recipe is selected on 1000 Karpathy validation images, hash-ordered after excluding all test IDs and images used by TrueMatch, CompreCap, and the COCO portion of DetailCaps. We compare Heun-64 (CFG 5) with ancestral DDPM-64 (CFG 5, $\eta = 1$), using each checkpoint's learned text ladder. Both disable self-conditioning and use argmax vocabulary readout of the conditional branch at the last ladder step. The decision maximizes the mean of SPICE and CAPTURE, subject to a CHAIR_s degradation of at most five points; ties within 0.1 favor Heun. DDPM obtains 23.73 versus 23.24 and is frozen for all checkpoints.

Noise is generated per sample on a CPU fp32 stream seeded from dataset, image ID, sample index, and a fixed experiment seed. Evaluation batch size is 64. COCO uses a 256-pixel center crop; fine-grained and detailed-caption benchmarks use long-side resizing and centered gray padding to preserve boundary content. Scores use the complete EOS-normalized text; raw token IDs and truncation flags are retained.

Caption quality and hallucination. COCO uses 5000 test images with all five references and the official pycocoevalcap scorers, including CIDEr-D and SPICE (Vedantam et al., 2015; Anderson et al., 2016). CHAIR_i is the fraction of mentioned object instances deemed hallucinated; CHAIR_s is the fraction of captions containing a hallucinated object (Rohrbach et al., 2018). CLIPScore uses ViT-B/32 and $2.5 \max(0, \cos)$; RefCLIPScore also incorporates references (Hessel et al., 2021). These scales are retained throughout. BLIP uses the released `blip-image-captioning-base` checkpoint (revision 82a3776), whose weights are converted from BLIP's pretrained ViT-B CapFilt-L model rather than its COCO caption fine-tune (Li et al., 2022a), with prompt "a picture of", beam size three, and at most 20 tokens as in BLIP's COCO evaluation.

Discrimination and detailed coverage. TrueMatch uses the official 254/104/93 bags of size 3/5/7 and the frozen OpenAI CLIP ViT-B/32 listener. Every bag image is a target once. The listener truncates at 77 tokens; ties count as hits following the official implementation. Our scorer reproduces the benchmark's released caption sets to within 0.1 points (HolisticCap 57.5/48.0/49.2 vs. 57.5/48.1/49.1).

DetailCaps-4870 uses three released references and CAPTURE 0.1.13 with FactualSceneGraph 0.5.0, its Flan-T5 parser, and MPNet matcher (Dong et al., 2024). A parser limit of 64 tokens per sentence motivates reporting truncation. CompreCap uses its 560 annotated images, spaCy parsing, MiniLM matching, and a Llama-3-8B-Instruct judge (Lu et al., 2025). We use greedy judge decoding and a single BOS token; the external comparison captions are rescored under the same settings. For CAPTURE, the rescored external captions match the benchmark's three-reference README values (its arXiv table reports two-reference values: CogVLM 60.20, ShareCaptioner 59.55, LLaVA-1.5-7B 49.69); for CompreCap, the paper's unified scores are 60.05 (GPT-4o) and 58.85 (LLaVA-NeXT-34B) versus our rescored 60.07 and 58.64. DOCCI-CAPTURE transfers the CAPTURE metric to one human reference per image (Onoe et al., 2024) and is labeled as a metric transfer.

C EXTERNAL COMPARISONS AND RESOURCE CONTEXT

Here we give component and data accounting for every external model in the main tables, complete prompt-following profiles, and the image-distribution protocols behind each FID. Each model row cites its original work. Blue: published by the cited source. Teal: open-source checkpoints or released outputs evaluated by us with our scorers. Parameter counts of released models were computed from tensor shapes in the official checkpoints (headers only), counting shared embeddings once.

C.1 MODEL COMPONENTS AND TRAINING DATA

We count every parameter the reported system needs at inference, including frozen text encoders, language models, and image codecs; text modules and codecs are listed separately but are included in the total. CoRe's 383M include both generative directions, its 114M text pathway, and its 1.42M renderer; its VGG/DINOv2 perceptual losses and Gram/iBOT head are training-only.

The GenEval .55 for SDXL uses the base pipeline without the refiner (the official GenEval generation script loads only the base model). DREAM uses the frozen `google/t5-v1_1-xxl` encoder and the SD kl-f8 ft-EMA VAE. UniSpace's 8B label is the active size per token of a model with two fully duplicated experts. The CoCa row is the COCO-finetuned open_clip release; our TrueMatch run of it reproduces the published value within 0.6.[1]

[1] Component sources: SD1.5, SDXL base, SD-VAE, T5-XXL, UniSpace, MiniT2I.

Table 9: **System components (M parameters).** Totals include every component needed at inference. ∼: computed from released tensor shapes or configurations.

Model	Total	Language component	Visual component or image codec
CoRe	383.1 (384.3 at 512^2)	55.7 encoder + 57.8 generator	87.2 encoder + 182.3 generator (incl. 1.42 renderer)
Understanding and captioning			
CLIP / LaCLIP / LaSLIP / SLIP-B/16 (Radford et al., 2021; Fan et al., 2023; Mu et al., 2022)	∼150	∼63	∼86 ViT-B/16
COSMOS-B/16 (Kim et al., 2025)	∼150	CLIP text tower	ViT-B/16
BLIP-base captioner (Li et al., 2022a)	224.0	137.9 BERT-base decoder with cross-attention	86.1 ViT-B/16
NDLB (Gaur et al., 2024; Mokady et al., 2021)	∼244	124.4 GPT-2 + 31.5 mapper	87.8 CLIP ViT-B/32
CoCa-L/14 (Yu et al., 2022)	∼638	331.7 (unimodal + multimodal)	306.7 ViT-L/14
Florence-2-base (Xiao et al., 2024)	230	BART-style encoder–decoder	DaViT
LLaVA-1.5-7B (Liu et al., 2024a)	7063	6740 Vicuna-7B	CLIP ViT-L/14-336
ShareCaptioner (Chen et al., 2024)	∼8401	7322 InternLM-7B	∼1079 EVA-ViT-g + Q-Former
Cambrian-1-8B (Tong et al., 2024)	∼11049	8030 Llama-3-8B	∼3018: four frozen encoders + SVA
InstructBLIP-13B (Dai et al., 2023)	∼14192	∼13017 Vicuna-13B	∼1176 ViT-g + Q-Former
CogVLM (Wang et al., 2024a)	∼17640	Vicuna-7B + visual expert	EVA-CLIP-E
LLaVA-NeXT-34B (Liu et al., 2024b)	∼34750	∼34390 Yi-34B	CLIP ViT-L/14-336
GPT-4o (Hurst et al., 2024)	undisclosed	undisclosed	undisclosed
Image generation			
MAGVLT-B (Kim et al., 2023)	447	integrated (371 transformer)	∼76 VQ-GAN
MiniT2I-B/32 / B/16 (Wang et al., 2026)	601 / 599	341 frozen FLAN-T5-Large	none (pixels; 260 / 258 denoiser)
SD1.5 (Rombach et al., 2022)	∼1066	∼123.1 CLIP-L text	∼83.7 VAE (∼859.5 U-Net)
SDXL base (Podell et al., 2024)	∼3469	∼123.1 + 694.7 CLIP-L + OpenCLIP-bigG	∼83.7 VAE (∼2567.5 U-Net)
DREAM-B / DREAM-L (Li et al., 2026)	∼5176 / 5482	66 contrastive text + ∼4762 frozen T5-XXL	∼83.7 SD kl-f8 ft-EMA (330 / 636 core, incl. text)
FLUID-L, DREAM impl. (Fan et al., 2025; Li et al., 2026)	∼5416	∼4762 frozen T5-XXL	∼83.7 SD-VAE (570 core)
UniSpace (Yan et al., 2026)	∼17200	integrated: two Qwen3-8B experts, ∼8190 each	∼836 PR-Qwen-ViT tokenizer + decoder

Table 10: **Training data and inherited models.** Presentations count repeated examples.

System	Task-training data	Exposure / inherited resources
CoRe	CC12M; LLaVA-NeXT long/short recaptions and alt-text	1.024B presentations; frozen VGG/DINOv2 perceptual losses (training only)
+ final fine-tune	CC12M-R at 512^2, then MiniT2I's 120K synthetic mixture	+51.2M, then +41.0M presentations (image denoiser only)
CLIP / LaCLIP / LaSLIP (Fan et al., 2023)	CC12M; LaCLIP/LaSLIP add LLaMA caption rewrites	35 epochs, batch 8192
COSMOS (Kim et al., 2025)	10.0M CC12M pairs with DreamLIP synthetic captions	32 epochs, batch 4096; EMA self-distillation
DREAM (Li et al., 2026)	11.3M CC12M pairs, alt-text	49 epochs ($\sim$0.554B presentations); frozen T5-XXL, SD-VAE
MiniT2I-B/32 (Wang et al., 2026)	CC12M long recaptions; then the 120K mixture	JAX reference configuration: 128M + 20.48M presentations; the PyTorch port uses twice the batch (256M + 41M); the blog does not state which produced its numbers; frozen FLAN-T5-Large; 512^2
MAGVLT (Kim et al., 2023)	CC3M, CC12M, SBU, VG ($\sim$17M)	pretrained VQ-GAN; CLIP reranking at evaluation
BLIP (Li et al., 2022a)	129M pairs with CapFilt bootstrapped captions	ImageNet-pretrained ViT, BERT initialization
NDLB (Gaur et al., 2024)	COCO captions, then self-retrieval fine-tuning	frozen CLIP ViT-B/32, pretrained GPT-2
InstructBLIP / Cambrian-1 / LLaVA	instruction mixtures (13 datasets / 2.5M + 7M / 558K + 665K; LLaVA-NeXT 558K + 760K)	pretrained LLMs and vision encoders
ShareCaptioner / CogVLM (Chen et al., 2024; Wang et al., 2024a)	100K detailed GPT-4V captions on InternLM-XComposer / a filtered 1.5B image–text pool plus grounding and instruction data	pretrained LLMs and vision encoders
SD1.5 (Rombach et al., 2022)	LAION subsets	$\sim$3.16B presentations; frozen CLIP-L text, VAE
SDXL (Podell et al., 2024)	undisclosed mixture	frozen CLIP-L and OpenCLIP-bigG text encoders, VAE
UniSpace (Yan et al., 2026)	generation, editing, and understanding mixture	$\sim$510M presentations; Qwen3-8B and a frozen visual tokenizer

C.2 Prompt following: external profiles and the MiniT2I gap

GenEval evaluates six constraints; DPG adds five question categories and a dependency-aware overall score. Published values keep their source precision; MiniT2I values follow its blog and JAX release. Our own profiles are in Table 19.

Table 11: **GenEval and DPG-Bench profiles of external systems.** GenEval tasks in percent, overall as a fraction. SD rows: official GenEval release (Ghosh et al., 2023) and Hu et al. (2024), Table 4; others: their papers or releases. Teal: the released MiniT2I-B/16 through our pipeline.

	GenEval							DPG-Bench					
Model	Single	Two	Count	Color	Position	Attr.	Overall	Global	Entity	Attribute	Relation	Other	Overall
SD1.5 (Rombach et al., 2022)	97	38	35	76	4	6	.43	74.63	74.23	75.39	73.49	67.81	63.18
SDXL (Podell et al., 2024)	98	74	39	85	15	23	.55	83.27	82.43	80.91	86.76	80.41	74.65
UniSpace (Yan et al., 2026)	98	92	69	88	83	73	.84	84.80	92.26	90.00	94.97	88.80	86.49
MiniT2I-B/32, CC12M pretraining (Wang et al., 2026)							.505						78.3
+ 120K fine-tuning							.826						82.3
MiniT2I-B/16, + 120K fine-tuning (Wang et al., 2026)							.873						84.2
our scorers, 512^2	99.7	95.7	79.1	92.8	81.5	78.3	.878						84.12
our scorers, downsampled to 256^2	99.7	95.2	76.9	93.4	82.8	80.0	.880						84.65

MiniT2I-B/32 and CoRe both use 256 image tokens and the same LLaVA-NeXT CC12M recaptions (Emporium, 2024). After CC12M pretraining, MiniT2I-B/32 reaches GenEval/DPG .505/78.3 and CoRe .461/74.43; after fine-tuning on the 120K mixture, MiniT2I-B/32 reaches .826/82.3 and our final fine-tune .729/77.79. Three measurements narrow down the gap. First, the released MiniT2I-B/16 run through our pipeline reproduces its published scores (.878/84.12 vs. .873/84.2) and gives the same scores when its images are downsampled to 256^2 (.880/84.65), so neither our scorers nor our output size penalize a model. Second, fine-tuning for MiniT2I-B/16's sample budget (5000 steps at batch 8192) reaches .732 at 256^2 (Appendix D.4), so fine-tuning exposure alone does not close the gap. Third, the final model's remaining deficit to MiniT2I-B/16 is concentrated in position (40.0 vs. 81.5), counting (60.3 vs. 79.1), and attribute binding (63.5 vs. 78.3), while single objects and colors are near ceiling for both (Tables 11 and 19)—the language-compositional tasks where a pretrained text encoder is most likely to help. The other differences are the frozen 341M FLAN-T5-Large text encoder versus our text encoder learned from captions alone, denoiser compute (313 vs. 69 GFLOPs per forward at 256 image tokens), caption mix, and the number of fine-tuning optimizer steps (40K vs. 8K; 20.48M vs. 41.0M presentations).[2]

C.3 Image-distribution comparisons

Single-sample FID depends on resolution, reference preprocessing, prompt field, and guidance selection, so each row keeps its protocol. DREAM does not state whether its CC12M evaluation images are held out or which CLIP backbone its similarity uses; our 100cos uses ViT-B/32. Numbers that select among several samples per prompt are not comparable and are omitted: DREAM-L with semantically aligned decoding (COCO 10.4, candidate count not stated for COCO) and OFA-L (10.5, COCO-finetuned, best of 24 by CLIP) (Li et al., 2026; Wang et al., 2022).

Fine-tuning on the synthetic mixture moves every model away from both natural-image distributions: our fine-tunes rise to 13–14 on CC12M-50K and 28–29 on COCO-30K (Table 16), and the released MiniT2I-B/16 scores 30.21 on COCO under our protocol. MiniT2I's reported FID is that of its pretraining checkpoint.

[2] MiniT2I blog; JAX and PyTorch releases.

Table 12: **Zero-shot COCO-30K and training-distribution CC12M-50K FID with protocol context.** Our rows: single sample per prompt, development-selected recipe (64 steps, 128 NFE), paired COCO references (short side 256, center crop), pytorch-fid Inception. External CC12M values from Li et al. (2026). [§]A different 30K COCO draw; on a shared 10K subset, 34.05 vs. 17.96 for CoRe + Gram/iBOT.

	COCO-30K	CC12M-50K		
Model	FID↓	FID↓	100cos	Protocol and resources
FLUID-L, DREAM impl. (Fan et al., 2025; Li et al., 2026)	9.62	4.53	30.0	CC12M, 256^2, single sample; T5-XXL + VAE
DREAM-L (Li et al., 2026)	10.0	4.34	29.4	CC12M, 256^2, single sample, 128 steps (64 steps: 4.57 / 29.1); best of two CFG schedules, per metric; T5-XXL + VAE
REPA, DREAM impl. (Yu et al., 2024)		4.42	29.9	published recipe
MiniT2I-B/16, CC12M pretraining (Wang et al., 2026)	10.51			CC12M recaptions, 512^2; references are precomputed 512^2 COCO statistics (pairing unstated)
+ 120K fine-tuning, our scorers	30.21[§]			released checkpoint, our pipeline
MAGVLT-B (Kim et al., 2023)	12.08			CC3M+CC12M+SBU+VG; CLIP reranking of 32 samples
MiniT2I-B/32, CC12M pretraining (Wang et al., 2026)	13.69			CC12M recaptions, 512^2, 256 image tokens
CoRe	10.10	4.33	29.33	256^2, 383M total
+ Gram/iBOT	10.20	4.37	29.30	same recipe and references
+ final fine-tune	–	13.84	–	512^2 generations brought to 256^2

D CONTROLLED INTERVENTIONS IN JOINT LEARNING

Every intervention below is evaluated on all three axes. Understanding: ImageNet zero-shot (ZS), COCO 5K retrieval mean recall (mR), the reduced-resolution ADE20K diagnostic on the contextual representation (MLP readout; the full-resolution protocol is reserved for the main models), and SugarCrepe (SC). Image-to-text: COCO 5K SPICE and CHAIR$_i$ with the frozen DDPM-64 recipe. Text-to-image: GenEval and DPG with the fixed 128-step recipe, COCO-30K FID with the development-selected recipe of the main tables, and ImageNet 5K reconstruction PSNR through the full-condition route.

D.1 WHICH CONDITIONING ROUTES SHAPE THE REPRESENTATION?

The self route conditions the target denoiser on its own global readout with a null local key. Removing it and redistributing its probability either proportionally or entirely to the null route leaves understanding and prompt following nearly unchanged but breaks full-condition reconstruction (Table 13): with the target's own (g_v, Z_v) supplied as condition, PSNR falls from 24.1 to 9–11 dB, while the null-conditioned reconstruction of the same models is unaffected (about 24.2 dB). Without the self route, the denoiser cannot use its own modality's representation as a condition, a failure invisible to benchmarks that condition on the other modality.

Table 13: **Conditioning-route ablation.** Route probabilities are (null, self, cross).

	Understanding				Image → text		Text → image			
Routes	ZS	mR	ADE (MLP)	SC	SPICE	CHAIR$_i$↓	GenEval	DPG	FID↓	PSNR
CoRe (.10, .20, .70)	50.63	70.34	26.22	80.6	13.53	11.2	.461	74.43	10.10	24.09
Self off, proportional (.125, 0, .875)	50.93	70.79	26.78	79.1	14.02	10.7	.439	73.30	10.13	9.45
Self off, to null (.30, 0, .70)	50.44	70.15	26.62	80.3	13.74	10.1	.428	72.80	10.61	10.93

D.2 How does the balance of objectives shape the representation?

Table 14 sweeps the alignment weight with the generative weights fixed, from no alignment (weight 0, the no-contrastive-loss control, which still trains both denoisers) through 0.25, 1, and 4, together with the contrastive-only extreme (no generative loss). Once alignment is present, global understanding is nearly flat across a 16-fold range of its weight (ZS 49.9–51.2). Prompt following peaks at weight 1, and captioning degrades at weight 4 (SPICE 12.98). Reconstruction improves monotonically as alignment weakens (26.91 dB without it), and prompt following collapses without it (GenEval .193): alignment is what lets a text representation place a prompt. The default weight of 1 is the only setting that is near the best on all three axes. Adding Gram/iBOT to the contrastive-only model shows that the structural terms help either objective: its full-resolution ADE20K mIoU rises to 38.74 at the backbone and 31.99 at the representation (Table 1), so the joint model's dense deficit persists under matched structural regularization, while its global readouts match (ZS 50.90 vs. 50.63, retrieval 70.88 vs. 70.87). COCO FID is nearly flat across aligned models (10.1–10.6) while prompt following varies (GenEval .427–.461), a reminder that distributional fidelity and prompt fidelity are distinct.

Table 14: **Objective balance.** Alignment weight λ_{align} with generative weights fixed; the contrastive-only control removes the generative losses instead. Gram anchoring alone (no iBOT, from 24K steps, $\lambda_{\text{Gram}} = 2$) leaves the dense and linear-probe readouts at the CoRe level (full-resolution ADE20K on h 34.72 vs. 34.53, ImageNet linear probe on h 72.34 vs. 72.82); the gains appear only with iBOT added (36.63, 73.62). An iBOT-only arm was not run.

	Understanding				Image → text		Text → image			
Objective	ZS	mR	ADE (MLP)	SC	SPICE	CHAIR$_i$↓	GenEval	DPG	FID↓	PSNR
No contrastive loss ($\lambda_{\text{align}} = 0$)	1.18	1.51	23.48	53.8	13.86	10.1	.193	45.31	23.20	26.91
$\lambda_{\text{align}} = 0.25$	51.20	71.28	26.49	79.8	14.14	10.2	.428	72.24	10.60	25.29
CoRe: $\lambda_{\text{align}} = 1$	50.63	70.34	26.22	80.6	13.53	11.2	.461	74.43	10.10	24.09
$\lambda_{\text{align}} = 4$	49.87	69.89	27.12	79.9	12.98	11.2	.427	71.80	10.17	22.71
Contrastive only (no generation)	49.51	69.75	27.21	79.4	–	–	–	–	–	–
CoRe + Gram/iBOT	50.90	70.88	28.51	80.0	13.50	11.2	.448	73.49	10.20	24.03
Contrastive only + Gram/iBOT	50.63	70.87	29.76	79.6	–	–	–	–	–	–

D.3 Does generation need the understanding representation itself?

The controls above show that co-adaptation substantially improves generation; they do not show that the space generation runs in must be the representation that understanding reads. We therefore train a matched pair that differs in one configuration key. Both arms continue the CoRe checkpoint at 24K training steps for 16K steps at batch 8192 and 256^2 (EMA initialization, fresh AdamW, learning rate 1.3×10^{-3} with 1000 warm-up steps and cosine decay, same data order and seed). In the *shared* arm the image denoiser generates $Z_v = \mathrm{LN}(P_v h_v)$, as in CoRe. In the *separate* arm it generates $\mathrm{LN}(P_v^{\text{gen}} h_v)$ through a second 768-to-256 projection (+0.2M parameters) initialized as a copy of P_v, so both arms start from the same function and differ only in where generation's gradient enters the encoder. Alignment, text conditioning, and every probe still read Z_v in both arms.

Generating in the understanding representation makes generation more compositional (Table 15). The shared arm is ahead by .066 GenEval and 5.4 DPG (prompt-bootstrap 95% intervals $[.043, .088]$ and $[4.40, 6.43]$; 10,000 paired resamples over prompts, stratified by GenEval task, none of which reverses the sign; this measures prompt sampling only, not training seeds; across three sampling seeds GenEval is $.4272 \pm .0029$ against $.3616 \pm .0024$, a gap of $.066 \pm .004$), with the largest GenEval gains on two objects (50.3 vs. 36.1), counting (30.0 vs. 20.0), and color attribution (19.5 vs. 12.0)—the tasks that need text–image binding—and a lead in every DPG category (entity 80.2 vs. 76.2, other 66.0 vs. 60.8). Global understanding does not care where generation's gradient enters: both arms gain the same over the initialization (ZS $+6.6$, retrieval $+3.3$ and $+3.5$, ImageNet k-NN $+5.8$ and $+5.7$) and tie within 0.3 on SugarCrepe, ARO, and captioning. The separate space learns the better reconstruction code (+1.65 dB, FD-DINOv2 57.0 vs. 78.8), but that code is less predictable from noise: at $\gamma = 4$ / 5 the clean-state estimate's cosine to its target is .48 / .28 against .56 / .34, and the mean denoising loss over training steps 8K–11K is 13% higher (0.0809 vs. 0.0713; both targets are normalized, but they differ, so the loss comparison is only indicative). Whether the shared target is also of lower rank depends on the measure: by participation ratio it is (21.8 against 32.8 for the

Table 15: **Shared versus separate generation space.** Both arms: 16K training steps from the same 24K-step checkpoint, 256^2, EMA weights. ADE20K: reduced-resolution diagnostic on the representation Z_v with linear SGD / linear AdamW + ZCA / MLP readouts (Appendix B). Rank: participation-ratio effective rank of Z_v on COCO patches. Reconstruction: ImageNet 5K, PSNR / FD-DINOv2. (b) Diagnostics: ImageNet validation-split probes on the pooled g_v (40/10 images per class; k-NN at $k = 20$, linear, 2048-unit MLP); ARO attribution / relation; linear ADE20K at the backbone and representation taps (a second run of the linear protocol); cosine between the clean-state estimate and the generation target at $\gamma = 4$ / 5; mean training denoising loss over steps 8K–11K.

(a)

	Global understanding			Dense representation		I→T	Image generation		
Generated in	ZS	mR	SC	ADE lin / ZCA / MLP	Rank	SPICE	GenEval	DPG	Reconstruction
24K-step initialization	38.07	65.88	–	–	28.4	–	–	–	–
Shared (CoRe)	44.65	69.15	79.8	15.4 / 22.9 / 26.0	21.8	12.01	**.428**	**71.30**	23.38 / 78.8
Separate	44.66	69.42	79.6	**21.9** / **25.4** / 26.6	30.0	12.13	.362	65.91	**25.03** / **57.0**

(b)

Generated in	ImageNet k-NN / lin / MLP	ARO	ADE tap study h / Z_v	$\cos(\widehat{Z}_0, Z_0)$	Denoising loss
24K-step initialization	49.10 / 49.27 / 58.45	–	26.0 / 17.0	.61 / .39	–
Shared (CoRe)	54.94 / 55.47 / 63.97	67.2 / 49.0	26.6 / 15.5	.56 / .34	0.0713
Separate	54.79 / 55.51 / 63.43	67.3 / 48.8	28.1 / 21.9	.48 / .28	0.0809

separate arm's target $\mathrm{LN}(P_v^{\mathrm{gen}} h_v)$), but 99% of the separate target's variance lies in 39 components against 91 for the shared one. "Lower rank is easier to denoise" therefore does not simply explain the gap; a rank-matched separate space, or one with its own alignment loss, is the clean test. The price of sharing is linear dense accessibility: the shared representation's effective rank falls below the initialization's (21.8 vs. 28.4; separate 30.0) and its linear-SGD ADE20K mIoU with it (15.4 vs. 21.9), a gap that whitening narrows to 2.5 points and an MLP nearly closes (26.0 vs. 26.6)—the same readout effect as in Section 5. The tap study places the cost at the representation: from the initialization, the linear representation tap falls from 17.0 to 15.5 in the shared arm and rises to 21.9 in the separate one, while the backbone tap rises in both (26.0 to 26.6 and 28.1). The pair is a controlled 16K-step continuation with one seed, not a full-length run.

D.4 What does fine-tuning change?

Table 16 compares the pretrained model with three fine-tunes on MiniT2I's 120K synthetic mixture (Appendix A): all parameters for MiniT2I-B/16's sample budget; the same with both encoders frozen, so that only the denoisers and their null codes train; and the final recipe of the main tables, which trains only the image denoiser after a 512^2 continuation on CC12M-R.

Table 16: **Fine-tuning on the 120K synthetic mixture.** FID: development-selected recipe; CAPTURE: DetailCaps. $^{=}$Encoders identical to the pretrained model. $^{\sharp}$Understanding and captioning read at the model's native 512^2 input; the second stage leaves every parameter they read untouched. $^{\natural}$Same 512^2 first stage, then all parameters for 8K steps with half of each batch from CC12M-R; read at 512^2. –: not measured at 512^2.

	Understanding				Image → text		Text → image			
Model	ZS	mR	LP h	SC	SPICE	CAPTURE	GenEval	DPG	COCO FID↓	CC12M FID↓
CoRe + Gram/iBOT, pretrained	50.90	70.88	73.62	80.0	13.50	44.15	.448	73.49	10.20	4.37
all parameters, 5K steps, 256^2	33.02	54.52	67.68	64.2	9.14	34.52	.732	79.25	28.33	13.28
denoisers only, 5K steps, 256^2	50.90$^{=}$	70.88$^{=}$	73.62$^{=}$	80.0$^{=}$	10.86	34.06	.696	78.10	29.40	13.24
final: image denoiser only, $512^{2\sharp}$	51.49	72.02	–	80.6	13.91	–	.729	77.79	–	13.84
joint: all parameters, 50% CC12M-R replay, $512^{2\natural}$	45.60	67.99	–	–	13.03	–	.748	78.26	–	5.22

Full fine-tuning buys prompt following (GenEval .448 → .732, DPG 73.49 → 79.25) at the cost of every other use: zero-shot accuracy falls to 33.0%, retrieval to 54.5, SPICE to 9.14, and DetailCaps CAPTURE to 34.52, while COCO-30K and CC12M-50K FIDs rise to 28.3 and 13.3 as generations adopt the mixture's style. Most of the understanding loss is over-fitting to the small mixture rather than a conflict between the objectives: an all-parameter fine-tune at 512^2 that replays CC12M-R for half of each batch loses 5.9 zero-shot points (51.49 → 45.60) instead of 17.9. Freezing both encoders keeps understanding exactly and still reaches 87% of the GenEval gain and 80% of the DPG gain, so prompt following does not require moving the shared representation; but captions still drift (SPICE

10.86), because the text denoiser now trains on the mixture's prompts alone, and both FIDs rise as much as under full fine-tuning, so these changes belong to the denoisers. Its samples for the prompts of Figure 2 are in Figure 4. The final recipe therefore trains the image denoiser only: its understanding and captioning are exactly those of its 512^2 first stage, and it keeps the prompt-following gain (.729). Freezing the encoders keeps the model unified: the image denoiser still generates, and understanding still reads, one representation, bit-identical to that of the first stage. The highest prompt following we measured still comes from joint training: with CC12M-R replay, all parameters reach GenEval .748 and CC12M-50K FID 5.22, but zero-shot accuracy falls to 45.60 and retrieval to 67.99. *Overlap with the benchmarks.* Of the mixture's 119,239 captions, 5 GenEval prompts appear verbatim (all single-object, all from the BLIP3o part), and 22 GenEval prompts (4.0%; 11 if word order must match) have a training caption with the same objects, counts, colors, and relation; no DPG-Bench prompt appears (largest token Jaccard 0.56). The mixture is, however, close to GenEval in format: 3,256 captions (2.7%) are complete GenEval-template captions over GenEval's vocabulary (7,227 with any object words), and 35.5% have at most 12 tokens, GenEval's length range. GenEval gains after fine-tuning therefore partly reflect in-format training; DPG gains (73.49 $\rightarrow$ 77.79 for the final model) are the cleaner evidence, and MiniT2I fine-tunes on the same mixture.

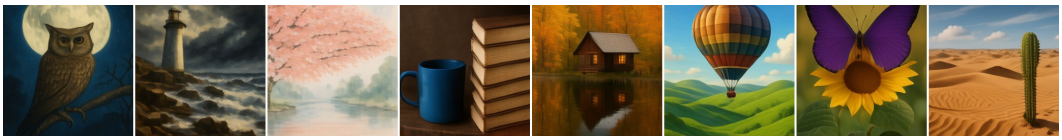

Figure 4: **The all-parameter 5K fine-tune on the prompts of Figure 2.** 256^2, 64 steps with CFG rising linearly from 1 to 5, churn 1; one of four seeds chosen by the model's own image–text score. This model reaches GenEval .732 but loses most of its understanding (Table 16).

E Extended capability results

This appendix expands the three main tables for the models and controls they report. Every row uses the protocols of Appendix B; the frozen-contrastive control shares the contrastive-only encoders and is listed only where it has its own generator.

E.1 Global and dense probes

Table 17: **ImageNet-1k linear probes and full-resolution ADE20K linear segmentation.** Linear probes: full-train LARS, last-epoch top-1 at three taps; the 95% interval resamples validation images for h. ADE20K: mIoU at the primary learning rate 10^{-4} and at 10^{-3} (sensitivity).

	ImageNet-1k linear probe				ADE20K mIoU			
					10^{-4}		10^{-3}	
Model	h	95% interval	Z	g	h	Z	h	Z
Contrastive only	72.51	[72.11, 72.88]	61.78	72.34	36.12	29.04	37.41	30.13
Contrastive only + Gram/iBOT	74.06	[73.67, 74.42]	63.91	73.98	38.74	31.99	40.07	33.16
No contrastive loss	71.72	[71.33, 72.11]	63.39	70.22	33.89	22.28	35.80	23.88
CoRe	72.82	[72.42, 73.20]	63.70	73.48	34.53	23.42	36.43	24.63
CoRe + Gram/iBOT	73.62	[73.24, 74.02]	64.78	74.31	36.63	24.05	38.26	25.44

Linear probes on h and g of all aligned models lie within 2 points of one another, whereas the dense representation separates them: joint learning lowers ADE20K mIoU on Z by 5.6 points (23.42 vs. 29.04), and the order is unchanged at the higher learning rate. The no-contrastive-loss control keeps 71.72% ImageNet accuracy on h with near-chance zero-shot transfer (Table 1): its information is retained but not organized for language.

Table 18: **COCO 5K retrieval, SugarCrepe, and ARO.** Retrieval R@1/5/10 in each direction. SugarCrepe (%): R/S/A are replace/swap/add. ARO (%): VG-Attribution and VG-Relation, chance 50; VG-Relation is at chance for every model and is shown only for completeness.

	COCO retrieval		SugarCrepe								ARO	
Model	I→T	T→I	R-obj	R-att	R-rel	S-obj	S-att	A-obj	A-att	Mean	Attr.	Rel.
Contrastive only	60.16/83.10/89.52	41.75/67.21/76.78	92.1	83.6	70.2	69.0	80.2	81.3	79.5	79.4	66.4	48.7
Contrastive only + Gram/iBOT	61.90/83.46/90.12	43.65/68.35/77.71	93.2	83.2	71.6	69.8	77.2	82.6	79.6	79.6	66.5	49.3
No contrastive loss	0.40/2.06/3.72	0.19/0.95/1.73	57.8	56.0	52.9	50.2	47.9	58.0	54.0	53.8	50.6	49.2
CoRe	60.78/83.00/89.64	43.12/68.19/77.32	93.2	85.5	71.6	70.2	81.2	82.7	79.5	80.6	63.6	47.8
CoRe + Gram/iBOT	61.08/83.80/89.64	44.06/68.91/77.80	92.9	84.5	72.3	69.0	79.0	83.0	79.5	80.0	65.1	49.1

E.2 Cross-modal and compositional matching

CoRe matches the contrastive-only control on retrieval and matches or exceeds it on every SugarCrepe split. ARO is less favorable: attribute accuracy is 2.8 points lower for CoRe (63.6 vs. 66.4), and relation accuracy stays at chance for every model, including the contrastive-only control.

E.3 Prompt-following profiles

Table 19: **GenEval and DPG-Bench profiles of our models** (fixed 128-step recipe; four images per prompt). GenEval tasks in percent, overall as a fraction. External profiles are in Table 11.

	GenEval							DPG-Bench					
Model	Single	Two	Count	Color	Position	Attr.	Overall	Global	Entity	Attribute	Relation	Other	Overall
No contrastive loss	62.5	6.3	11.6	32.7	0.5	2.0	.193	72.95	60.37	69.70	79.30	37.2	45.31
Frozen contrastive → generation	83.8	47.7	28.8	49.2	7.8	22.8	.400	77.20	79.14	81.65	89.21	63.2	70.12
CoRe	88.4	54.5	32.8	67.0	8.3	25.8	.461	79.94	82.91	82.85	90.29	68.4	74.43
CoRe + Gram/iBOT	91.6	58.3	30.0	60.9	8.0	20.3	.448	79.64	82.00	81.85	90.37	68.4	73.49
all parameters fine-tuned, 5K steps, 256^2	99.1	87.1	58.1	89.9	41.5	63.3	.732	80.85	86.87	85.44	91.53	74.8	79.25
final fine-tune, 512^2	99.1	86.6	60.3	87.8	40.0	63.5	.729	81.16	85.93	85.04	91.14	68.8	77.79

Without fine-tuning, prompt following fails mostly on relative position and attribute binding; fine-tuning on the synthetic mixture raises these most (final fine-tune: position 8.0 → 40.0, attributes 20.3 → 63.5), while single objects and colors approach their ceiling. DPG moves less, mainly on entity and attribute questions.

E.4 Reconstruction

Table 20: **Low-noise reconstruction on 5K ImageNet validation images** (full-condition route; VAEs decode their posterior mode). Two independent draws of real images give rFID 3.058 at 50K (7.526 at 20K); reconstructions are paired with their own originals and can fall below this reference, and the 5K values are dominated by finite-sample bias.

Model	PSNR	SSIM	LPIPS↓	rFID 5K↓	rFID 50K↓
SD-VAE kl-f8 ft-EMA (Rombach et al., 2022)	25.04	0.711	0.141	3.05	–
SDXL-VAE (Podell et al., 2024)	26.10	0.749	0.128	2.95	–
No contrastive loss	26.91	0.784	0.141	4.76	–
Frozen contrastive → generation	19.26	0.492	0.306	11.58	–
CoRe	24.09	0.669	0.206	6.43	2.31
CoRe + Gram/iBOT	24.03	0.665	0.203	6.24	2.20

The encoded representation reconstructs within 1–2 dB of the two VAEs, which are deterministic and trained for reconstruction alone; at 50K images CoRe's rFID falls to 2.31, so most of its 5K value (6.43) is finite-sample bias. Alignment costs fine detail (no contrastive loss: 26.91 dB), and a representation never shaped by generation reconstructs worst (19.26 dB).

What the representation retains as noise rises. Figure 5 decodes six images from their own contextual representation corrupted at increasing noise, with one noise draw per image and learned

null conditions, so that the noised representation is the only image-side input. Global layout and color survive to $\gamma \approx 3$ (under 5% signal variance), legible text to $\gamma \approx 2$, and fine lettering degrades already at the lowest noise level, consistent with the 24 dB reconstruction level of Table 20. Averaged over the six images, CoRe reaches 24.8/24.2/22.5/20.6/17.4/13.9/12.4 dB at $\gamma = -3.34/0/2/3/4/5/6.5$; the no-contrastive-loss control is sharper at low noise (28.0 dB at $\gamma = -3.34$) but indistinguishable from $\gamma = 3$ on (20.9, 17.1, 14.3, 12.8 dB): alignment costs fine detail, not coarse structure.

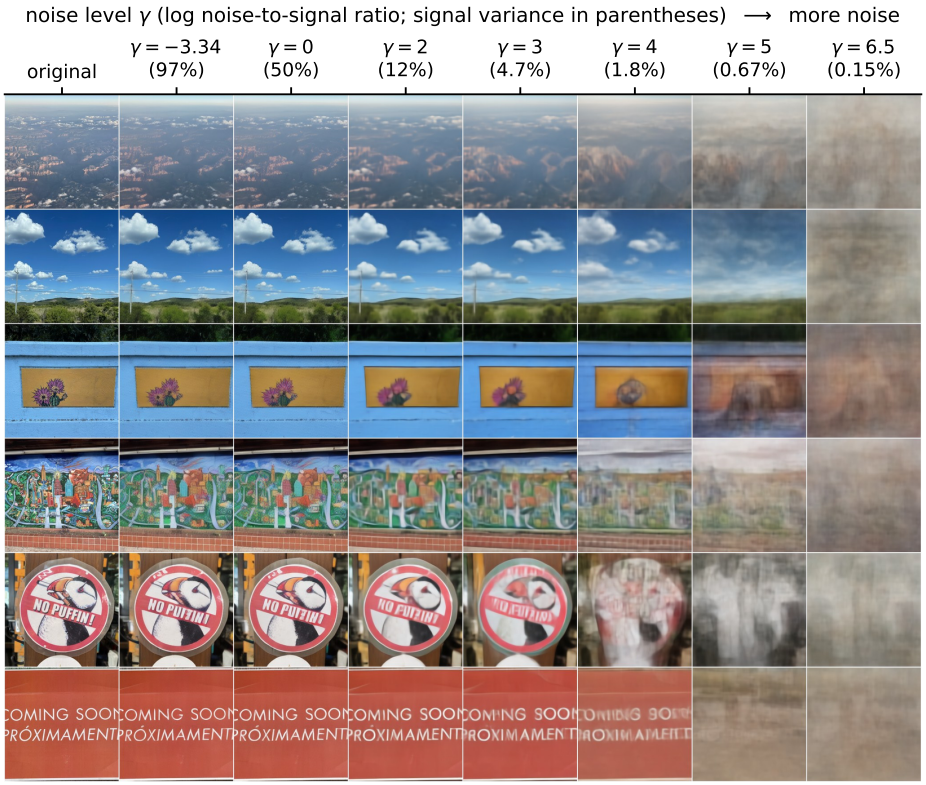

Figure 5: **What the contextual representation retains as noise rises.** CoRe; each row is one person-free DOCCI image (CC BY 4.0): landscapes, a painting, a mural, and signs with benign text. Columns: the original, then one null-route denoiser call from the image's own representation corrupted at noise level γ (top axis; signal variance $\mathrm{sigmoid}(-\gamma)$ in parentheses).

E.5 COCO CAPTION QUALITY AND HALLUCINATION

CIDEr-D penalizes our long captions (43 words against about 10 for the references); scoring only the first sentence raises it to 14.80 for CoRe at nearly unchanged SPICE (13.44). The own-score selection rule was fixed in advance. At matched length it raises SPICE by 0.31 (paired 95% interval $[0.20, 0.42]$), CIDEr-D by 0.40 ($[0.18, 0.64]$), and RefCLIP-S by .0024 ($[.0016, .0032]$) over the single sample, and beats random-of-four on SPICE by 0.29 ($[0.08, 0.49]$); CHAIR_i does not change significantly. Unconstrained selection prefers longer captions (+8.9 words), which raises SPICE by 1.06 but lowers CIDEr-D, so the length control is necessary. The rule costs four samples plus four caption encodings and one image encoding (about 40 ms on one H100, one percent of the sampling time).

E.6 DETAILED CAPTIONING

On all three benchmarks our models lie within 2 points of one another. They trail captioners built on 7B–17B pretrained language models by 5–16 CAPTURE points on DetailCaps, and GPT-4o and

Table 21: **COCO Karpathy 5K captioning.** One DDPM-64 sample per image unless marked; CHAIR in percent. The human row scores one reference against the other four. Bottom: CoRe + Gram/iBOT choosing among four samples per image (main-protocol seeds) with its own image–text score; the length-matched rule picks the highest-scoring caption within two words of the seed-0 length when one exists.

Model	NFE	CIDEr-D	SPICE	$\text{CHAIR}_s\downarrow$	$\text{CHAIR}_i\downarrow$	RefCLIP-S	Words
Human reference	–	89.28	21.19	3.5	2.5	.8175	10.4
BLIP-base (Li et al., 2022a)	–	94.39	17.53	3.8	3.2	.8081	6.2
No contrastive loss	128	8.66	13.86	18.8	10.1	.7414	43.1
Frozen contrastive → generation	128	3.63	10.06	29.8	16.5	.7358	54.2
CoRe	128	9.10	13.53	21.6	11.2	.7398	42.6
CoRe + Gram/iBOT	128	9.13	13.50	20.9	11.2	.7399	43.8
final fine-tune (512^2 representation)	128	9.45	13.91	19.9	10.4	.7434	43.4
Selection among four samples by the model's own readout (CoRe + Gram/iBOT)							
Random of 4	512	9.70	13.52	–	10.9	.7396	44.8
Own score, length-matched	512	9.54	**13.81**	–	11.4	.7423	43.8
Own score, best of 4	512	7.43	14.56	–	10.9	.7458	52.7

Table 22: **Detailed-caption benchmarks.** DetailCaps-4870: CAPTURE and its object/attribute/relation F1 (three references). CompreCap: object coverage (%), attribute and relation scores (0–5), and the unified score, with a fixed single-BOS Llama-3-8B-Instruct judge. DOCCI-5K: CAPTURE transferred to one human description. External rows: released captions rescored with the same evaluators (Dong et al., 2024; Lu et al., 2025).

	DetailCaps				CompreCap				DOCCI
Model	CAPTURE	Object	Attribute	Relation	Object %	Attribute	Relation	Unified	CAPTURE
LLaVA-1.5-7B (Liu et al., 2024a)	49.44								
ShareCaptioner-7B (Chen et al., 2024)	59.80								
CogVLM (Wang et al., 2024a)	60.06								
GPT-4o (Hurst et al., 2024)								60.07	
LLaVA-NeXT-34B (Liu et al., 2024b)								58.64	
No contrastive loss	44.43	48.83	41.79	40.06	46.27	1.559	2.510	42.56	45.84
Frozen contrastive → generation	43.73	48.07	41.71	37.95	45.43	1.514	2.394	41.11	44.46
CoRe	43.62	48.01	41.02	39.17	47.37	1.594	2.480	42.84	45.66
CoRe + Gram/iBOT	44.15	48.56	41.44	39.88	45.69	1.537	2.442	41.72	45.52

LLaVA-NeXT-34B by 16–19 points on CompreCap. Parser truncation affects at most 2.1% of our DetailCaps captions (5.5% for the frozen-contrastive control, whose captions are longest).

E.7 QUALITATIVE SAMPLES

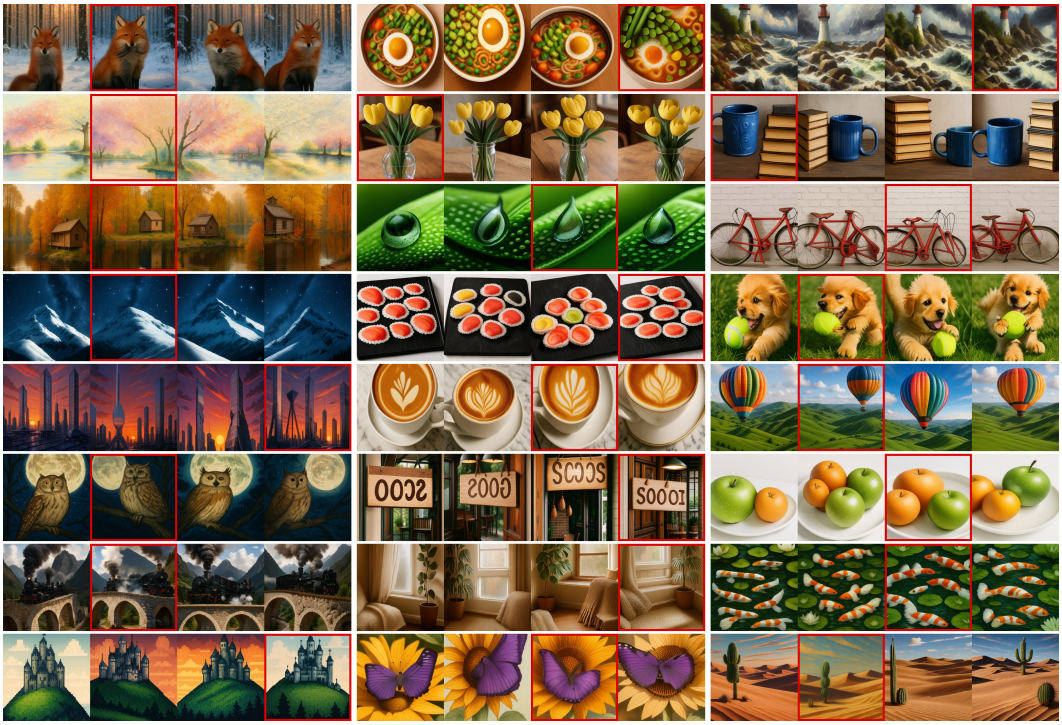

Figure 6: **All samples, unselected.** Four seeds for each of 24 person-free prompts from the model and sampler of Figure 2; the seed chosen by the model's own image–text score is framed in red. Read left to right, top to bottom, three prompts per row: (1) A red fox sitting in a snowy forest at dawn; (2) A bowl of ramen with a soft-boiled egg and green onions, viewed from above; (3) An oil painting of a lighthouse on a rocky coast during a storm; (4) A watercolor painting of cherry blossoms along a quiet river; (5) Three yellow tulips in a clear glass vase on a wooden table; (6) A blue ceramic mug to the left of a stack of old books; (7) A small wooden cabin beside a lake surrounded by autumn trees, reflected in the water; (8) A macro photograph of a dewdrop on a green leaf; (9) A vintage red bicycle leaning against a white brick wall; (10) A snowy mountain peak under a starry night sky; (11) A plate of sushi arranged on a black slate board; (12) A golden retriever puppy playing with a tennis ball on the grass; (13) A futuristic city skyline at sunset with glass towers; (14) A cup of cappuccino with latte art on a marble table; (15) A hot air balloon floating over rolling green hills; (16) An owl perched on a branch at night under a full moon; (17) A wooden sign that says "OPEN" hanging on a cafe door; (18) A green apple and two oranges on a white plate; (19) A steam locomotive crossing a stone bridge in the mountains; (20) A cozy reading nook with a window, a potted plant, and a knitted blanket; (21) A koi pond with orange and white fish among lily pads; (22) A pixel-art style castle on a hill; (23) A purple butterfly resting on a sunflower; (24) A desert landscape with sand dunes and a single cactus. Rendered text (prompt 17) fails, counts are not respected (prompt 5 shows more than three tulips in every seed), and the chosen desert sample adds a second small cactus (prompt 24).

Image	Verbatim caption from CoRe (DDPM-64, first sample)
	A leatherwater sea turtle with its textured pattern and prominent sharp eyes rests in a blue ocean environment, surrounded by small fish.
	A serene scene of a clear blue sky with fluffy white clouds, a lush green hillside, evoking tranquility and natural beauty beauty.
	A tall, spiky ta stands tall in a vast, sandy landscape, its needle-like leaves contrasting with the light and flat, rocky sand under a clear blue sky.
	A wooden sign with a stylized graphic of a gull wearing a helmet with the text “no smoking” above it. the sign is mounted on a wooden pole and there are two circular stickers with the text “nonos” in bold, capital letters. the background is blurred, but it appears to be a store setting with various light, possibly a window and a dining area. the sign is is likely sale, as indicated by the text anding on the,. it is not legd.



Figure 7: **Both directions from one pretrained model.** CoRe after CC12M pretraining only. Top: verbatim captions of person-free DOCCI test images (the main-table DDPM-64 recipe, per-image seeds, first sample; defects such as repeated words are left in). Bottom: one of four seeds per prompt, chosen by the model’s own image–text score, with the development-selected 64-step sampler; all four seeds are in the supplementary material, where web-watermark artifacts of the pretraining data appear in some samples; fine-tuning on the synthetic mixture removes them. Prompts, seeds, picks, and checkpoint are listed in the supplementary material.

F FEW-STEP TRANSPORT AND OBSERVABLE READOUT

A continuous contextual representation admits a variable inference budget. We distill the CoRe + Gram/iBOT denoisers with the encoders fixed. Each student copies its modality's denoiser and adds a step-size embedding. Shortcut distillation trains dyadic jumps with a teacher-anchored finest level and EMA bootstrap targets; flow-map matching trains arbitrary start/end queries (Frans et al., 2025; Boffi et al., 2024). All students distill the guided, churn-free teacher ODE (guidance 7 for text, 3 for images); at inference the guidance is inside the student and each jump costs one denoiser evaluation.

Table 23: **Adjusting inference computation.** Distilled students (own readout) versus the teacher (CoRe + Gram/iBOT) at matched NFE, which counts every denoiser evaluation including both guidance branches and the observable readout. FD-DINOv2 and CLIP-S on a 1K COCO draw (two seeds; a disjoint draw reproduces the students within 6.1 FD); GenEval on all 553 prompts $\times$ 4; SPICE on the first sentence of each caption (length-controlled; 1K images). Teacher rows use the best recipe found at each budget; the 64-step caption teacher is deterministic (with churn it reaches CLIP-S .791 and SPICE_S 14.28); Wall-clock cost on one H100 is in Table 25. FD-DINOv2 favors models trained with a DINOv2 loss, as the teacher and students all are, so it ranks rows within this study only.

Text $\to$ image					Image $\to$ text			
Sampler	NFE	$\text{FD}_{\text{DINOv2}}\downarrow$	CLIP-S	GenEval	Sampler	NFE	CLIP-S	SPICE_S
Teacher, 2 steps	5	913.4	.752	.309	Teacher, 2 steps	5	.760	10.88
Teacher, 4 steps	9	762.4	.771	.392	Teacher, 4 steps	9	.778	12.37
Teacher, 64 steps	129	679.6	.774	.448	Teacher, 64 steps	129	.782	13.08
Student, 1 jump	2	931.6	.745	.317	**Student, 1 jump**	2	.763	11.51
Student, 4 jumps	5	**775.1**	.764	**.379**	**Student, 8 jumps**	9	**.781**	**13.43**
Student, 8 jumps	9	**730.6**	.773	**.428**	w/o KL anchor	9	.749	–

Transport and output anchoring. For source and destination log-noise levels $\gamma_a > \gamma_b$, a predicted clean state defines the jump

$$J_{a\to b}(u, \widehat{z}_0) = \alpha_b \widehat{z}_0 + \frac{\sigma_b}{\sigma_a}(u - \alpha_a \widehat{z}_0). \tag{8}$$

The student approximates teacher transport with a pseudo-Huber loss on clean states. Optional renoising replaces the implied noise by $\sqrt{1-r^2}\,\widehat{\varepsilon} + r\varepsilon'$ after non-final jumps; we use $r = 1$ for multi-jump students. Because the observable head shares the denoiser trunk, its behavior can drift while the transport is fitted. We therefore add a readout anchor,

$$\mathcal{L}_{\text{read}} = \mathbb{E}_{(u,C)} \sum_j \text{KL}\left(p_{\text{teacher},j}(\cdot \mid u, C) \,\|\, p_{\text{student},j}(\cdot \mid u, C)\right) \tag{9}$$

for text (weight 0.2), and the analogous pixel-space anchor for images (weight 1). Both heads are queried at the terminal noise level; half the anchored representations come from corrupted encoded observations and half from a detached one-jump student rollout. Text students use batch 2048, learning rate 10^{-4}, and 20K training steps; image students use effective batch 2048 and 20K training steps. Tables report the final live weights at 20K training steps (the 14K checkpoint is 5–15 FD-DINOv2 better, but we keep the pre-chosen final checkpoint).

Evaluation. A fixed 1000-image draw from the COCO Karpathy test set with all five references, a disjoint 1000-image real control (FD-DINOv2 534.7, FID 58.1 against the draw), and a second disjoint data draw for replication. T2I prompts are the first reference captions. Student NFE is $k + 1$ for k jumps, including the observable readout; guided teacher NFE is $2s + 1$ for s steps. Teacher recipes (steps $\times$ guidance $\times$ churn) are swept separately at each budget. These 1K-image distances are dominated by finite-sample bias and are comparable only within this study.

At 5 and 9 NFE the students beat the best teacher recipe found at the same budget (Table 23): at 9 NFE the eight-jump image student reaches FD-DINOv2 730.6 against 762.4, and the caption student reaches CLIPScore .781 and first-sentence SPICE 13.43 against .778 and 12.37; the 64-step stochastic teacher remains the quality ceiling. Decoding the same student latent with the teacher's readout is 8–12 FD-DINOv2 better, so the student's own readout is not the main limit. The readout anchor

is what keeps the students' observable heads usable: without it the eight-jump caption student's own-head CLIPScore falls from .781 to .749, and the pixel anchor improves FD-DINOv2 by 18–26 at every jump count (anchored and unanchored image students at 10K training steps, two seeds).

Table 24: **Matched-budget image generation on the main-table protocols.** Zero-shot COCO-30K (FID / FD-DINOv2 on the frozen test list), training-distribution CC12M-50K FID, GenEval (553 × 4) and DPG-Bench (1065 × 4); one sample per prompt. Teacher rows at 5, 9 and 17 NFE use the best recipe found at that budget on the 1K draw; the 128- and 257-NFE rows are the main-table recipes. Students: four jumps with the teacher's learned stops, one and eight jumps with linear stops (one jump is the same under both), own readout. *CFG-7 target*: the same student recipe distilled from the CFG-7 teacher trajectory (10K training steps), linear stops at four and eight jumps. Bold: best at each budget. The 5–17-NFE teacher recipes and the students' stop schedules were selected on the 1K Karpathy-test draw, 752 of whose 1000 images are also in the COCO-30K test list (741 expected by chance).

			COCO-30K		CC12M-50K		
Model	Jumps / steps	NFE	FID↓	FD-DINOv2↓	FID↓	GenEval	DPG
Teacher	2 steps, CFG 5	5	22.15	480.5	17.91	.309	64.58
Teacher	4 steps, CFG 7, churn 1	9	**11.95**	310.3	7.18	.392	70.37
Teacher	8 steps, CFG 7, churn 1	17	10.61	255.8	–	–	–
Teacher	64 steps, dev-selected	128	10.20	221.5	4.37	.433	73.20
Teacher	128 steps, CFG 3, churn 1	257	13.66	–	–	.448	73.49
Student	1 jump	2	16.06	495.7	9.53	.317	67.38
Student	4 jumps	5	**11.79**	317.3	**6.14**	.379	70.57
Student	8 jumps	9	12.45	279.1	**6.07**	.428	72.09
Student, CFG-7 target	1 jump	2	15.38	463.0	8.89	.329	67.53
Student, CFG-7 target	4 jumps	5	13.58	**275.5**	6.68	**.457**	**72.16**
Student, CFG-7 target	8 jumps	9	15.04	**259.2**	7.21	**.485**	**73.52**

At the benchmark scale of the main table, the four-jump student beats the teacher's best 5-NFE sampler on every metric (COCO-30K FID 11.79 vs. 22.15; CC12M-50K 6.14 vs. 17.91; DPG 70.57 vs. 64.58). At 9 NFE the eight-jump student is better on FD-DINOv2, CC12M-50K, GenEval, and DPG, while the teacher's four-step sampler has the lower COCO-30K Inception FID (11.95 vs. 12.45). One jump (2 NFE) beats the 5-NFE teacher on every metric except COCO-30K FD-DINOv2. Distilling the CFG-7 instead of the CFG-3 teacher trajectory trades Inception FID for alignment. At 9 NFE the CFG-7 student reaches GenEval .485, above every teacher sampler we scored (the 257-NFE teacher: .448; paired prompt bootstrap +.036, 95% CI [.016, .057]), and DPG 73.52, level with the 257-NFE teacher (73.49). Guidance alone does not explain the GenEval gain: the teacher itself at CFG 7 and 64 steps scores .436 without and .424 with churn. The CFG-7 student also has the lowest FD-DINOv2 of any sampler at 9 NFE or fewer on both sets (COCO-30K 259.2, CC12M-50K 106.3 vs. 128.4 for the CFG-3 student), but a higher Inception FID at four and eight jumps (COCO-30K 13.58 and 15.04 vs. 11.79 and 12.45), the usual effect of stronger guidance.

On first sentences (equal length across rows; 1K images, paired bootstrap), the anchored caption student beats the teacher at equal NFE by 1.05 SPICE at 9 NFE, 2.48 at 5, and 3.11 at 2–3 NFE (all intervals exclude zero), and trails the 64-step stochastic teacher by 0.85. On the main-table I2T protocols (Table 2), the eight-jump student at 9 NFE reaches SPICE 13.66, CHAIR_i 11.9%, TrueMatch 58.9/50.0/48.4, CAPTURE 44.86, CompreCap 42.19, and DOCCI 46.54, against 13.50 / 11.2% / 60.6/51.9/47.9 / 44.15 / 41.72 / 45.52 for the teacher's DDPM-64 recipe at 128 NFE; one jump (2 NFE) keeps most of it. Students add only the step embedding to the 383M base. On fixed hardware (Table 25) the eight-jump caption student (79 ms at batch 1) is faster than BLIP-base (107 ms), Florence-2-base (136 ms short, 394 ms detailed), and CoCa-L/14 (352 ms), whose autoregressive decoders emit one token per step; at batch 64 the one-jump student captions 1848 images/s against 73 for BLIP-base.

Table 25: **Fixed-hardware cost.** One dedicated H100 80GB, fp32 weights with TF32 matrix multiplies (the precision of all our evaluations), real COCO test images and captions, CUDA-synchronized wall-clock time, median of five (batch 1) or two (batch 64) calls after two warm-up calls. Latency at batch 1 includes guidance branches and the observable readout; encoding is listed separately because one encoded representation serves every task. External captioners use the decoding settings of Table 2.

Path	NFE	Latency, batch 1 (ms)	Throughput, batch 64 (/s)	Peak memory (GB)
CoRe: encode image	–	7.3	2284	3.1
CoRe: encode caption	–	8.2	4196	2.9
I2T teacher, DDPM-64, CFG 5	128	1079	18.1	9.5
I2T student, 8 jumps	9	78.7	403	6.6
I2T student, 1 jump	2	16.9	1848	6.5
T2I teacher, fixed 128 steps, CFG 3	257	4751	4.1	6.2
T2I teacher, dev-selected 64 steps	128	2342	8.3	6.2
T2I student, 8 jumps	9	172.5	117	6.2
T2I student, 4 jumps	5	92.8	210	6.2
T2I student, 1 jump	2	37.7	528	6.1
BLIP-base caption (beam 3)	–	107	72.8	9.7
Florence-2-base, <CAPTION> / <DETAILED_CAPTION>	–	136 / 394	26.2 / 12.3	14.9
CoCa-L/14 caption	–	352	23.7	7.2

G ITERATIVE CONTEXTUAL REFINEMENT

The representation head produces a clean-state estimate at every step, which can be fed back as an additional input to the next denoiser evaluation through a zero-initialized projection (self-conditioning; Chen et al., 2022). We compare three 15K-step continuations of CoRe + Gram/iBOT on CC12M with *frozen encoders*: a matched control without self-conditioning; previous-step conditioning on the clean-state estimate $\widehat{Z}_0$; and previous-step conditioning on a *re-encoded token posterior*, in which the predicted token probabilities form soft input embeddings that are passed through the frozen contextual text encoder. Self-conditioning is trained with probability 0.5 in both modalities. This input is separate from the self training route of Appendix D. Captions are scored on COCO test images with the native sampler (CFG 5), CLIPScore, SPICE, and the non-word rate d, the fraction of generated words with zero English word frequency (dropped-subword artifacts such as "windowboard").

Table 26: **Previous-step conditioning at equal NFE.** 500 COCO images, 32 steps (64 denoiser NFE), one caption per image, same seeds; SPICE from the COCO scorer on the full caption. Re-encoding adds one encoder pass per step.

Conditioning (frozen encoders)	CLIP-S	$d\downarrow$	Words	SPICE
Matched control	.7708	.0694	46.5	12.65
Previous clean state $\widehat{Z}_0$	.7751	.0694	47.0	12.92
Re-encoded token posterior	**.7846**	**.0631**	49.1	**13.60**

Re-encoding raises CLIPScore by .0138 and lowers the non-word rate by 9% over the matched control at no extra denoiser cost; feeding back $\widehat{Z}_0$ directly gains only .0043. Captions are 2.6 words longer, but the gain survives length control: scoring only the first sentence of each caption (equal length between the arms), re-encoding raises SPICE by 1.10 and CIDEr-D by 4.6 over the matched control (paired bootstrap 95% intervals exclude zero; full captions: SPICE +0.95). The contrast between the two feedback paths is the interesting part: routing the model's prediction through its observable and back through its own contextual encoder is more useful than recycling the latent estimate, consistent with the observable being the stable coordinate system of an evolving representation (Section 3). The gain is specific to language: for images, previous-step conditioning raises FD-DINOv2 on 1K COCO prompts (680.5 $\rightarrow$ 777.6 for the re-encoded arm with 32 steps and churn 1).

G.1 Loop depth versus sampling steps, with a compute control

The re-encoded-posterior checkpoint above also supports an *inner* loop: at each of the 32 outer diffusion steps, the same noise level is revisited K times before advancing, each pass re-encoding the previous pass's token posterior. $K = 1$ is the single-pass decode used everywhere else in this paper. Table 27 asks whether this inner loop is worth its extra denoiser calls by comparing it against a control that spends the identical NFE as ordinary outer steps instead: at 128 total denoiser calls, 32 steps with $K = 2$ scores .6550 against .6556 for 64 plain steps with no loop; at 256 calls, 32 steps with $K = 4$ scores .6598 against .6592 for 128 plain steps. Every paired difference is within its bootstrap interval of zero. At matched compute the loop is interchangeable with ordinary steps: it is only occasionally, and never significantly, better.

Table 27: **Inner loops versus ordinary steps at equal compute.** Image-to-text, native sampler, self-conditioning carry on, 500 COCO images; score = CLIPScore minus the per-caption rates of repeated words, dropped subwords, and invented names. Rows pair the inner-loop route against a plain-step control at the same total denoiser NFE.

NFE	32 steps $\times$ loop K	plain steps, no loop
64	32×1: .6493	–
128	32×2: .6550	64×1: .6556
256	32×4: .6598	128×1: .6592

Qualitatively, though, the two routes are not indistinguishable: since K trades against ordinary steps only in aggregate score, a fixed 32-step budget with more inner loops still visibly changes individual captions, because the extra passes let a poor early posterior get re-encoded and corrected. Figure 8 shows this on three person-free COCO validation images (chosen from six person-free examples because each shows a clear change with K): $K = 1$ can degenerate into repetition (a full paragraph of stutters on the street scene) or state a plausible but wrong detail ("glass of water" for a vase, an undercounted "two sheep"); by $K = 2$ or $K = 4$ the caption stabilizes and is usually more accurate. This is the same phenomenon Table 27 scores in aggregate: more denoiser calls help, whether they arrive as loop depth or as steps, and the loop's only advantage is not needing a second solver step to spend them.

Image	K	Caption (re-encoded-posterior checkpoint, native sampler, 32 steps, CFG 5, carry on)
	1	Two sheep grazing in a green field.
	2	Three sheep graze in a lush green field, surrounded by trees and clear sky.
	4, 8	Three sheep graze in a lush green field, surrounded by trees and a clear blue sky.
	1	A large, clear, cylindrical glass of water sits on a balcony, overlooking a city street scene with residential buildings and a clear blue sky.
	2, 4, 8	A large, clear, cylindrical, clear vase sits on a balcony, overlooking a city street scene with residential buildings and a clear blue sky.
	1	A vibrant street scene on a city. the foreground is dominated by a white car, its its bluring of the right. the car's head ising. the road, a the,. thea white hues out on the right side of the road, marked with treesa the lights, adding a sense of the the., the right side of the scene. the is a traffic of a light, moving through the same direction. the trees are . the the side, their trees, adding a sense of order-madeity to the scene. the street is lined with several cars, their their their motion, rusting against the in the sky. the traffic sign, the's a green limit of visible, indicating thatas a the urban area. the sky overhead is a clear blue, suggesting a bright and sunny day and. below. overall, image encapsulates a typical of urban life, life where thea of of the life..
	2	A street lined with cars, trees, and a green with a blurred red trafficwalk signal in the distance.
	4, 8	A street lined with cars, trees, and street lights, with a blurred blue car driving motion in the distance.

Figure 8: **Loop depth on three examples.** Ground truth (not shown): "A herd of sheep are grazing in a green field"; "A tall glass vase on a balcony"; "A smaller car is stopped at the red light, as traffic drives on." Captions are verbatim, including their defects. Checkpoint: the frozen-encoder, re-encoded-token-posterior self-conditioning continuation of Appendix G. Quantitative compute control: Table 27 and Appendix G.1.

H Representation measurements

H.1 Readout ladder on fixed features

For a declared target, retained information concerns the risk attainable from the representation by an unrestricted predictor; a fitted probe measures one practical readout of it. Table 28 fits six readouts of increasing flexibility to the frozen contextual representation of every arm on the reduced-resolution ADE20K diagnostic (twofold-downsampled package, 8K training images, 512-pixel input, 40 epochs), all from one run per cell.

Table 28: **Readout ladder on the frozen contextual representation.** ADE20K mIoU, reduced-resolution diagnostic, one run per cell. The frozen-contrastive control shares the contrastive-only encoders; its row is an independent probe run on the same features, so the two rows differ only by probe-training noise.

	Linear SGD			Linear AdamW		MLP
Model	raw	std.	ZCA	raw	ZCA	AdamW
Contrastive only	24.51	25.24	27.82	27.30	27.91	27.21
Frozen contrastive $\rightarrow$ generation	24.50	25.26	27.79	27.30	27.90	27.26
No contrastive loss	9.66	15.49	21.51	18.26	22.58	23.48
CoRe ($\lambda_{\text{align}} = 1$)	15.83	17.80	23.04	21.02	23.59	26.22
$\lambda_{\text{align}} = 0.25$	12.21	16.08	20.91	19.43	22.02	26.49
Self off, proportional	15.64	17.31	22.74	20.77	23.46	26.78
Self off, to null	16.10	18.00	23.17	20.84	23.73	26.62
Contrastive only + Gram/iBOT	27.54	28.26	31.00	31.10	31.26	29.76
CoRe + Gram/iBOT	16.53	18.59	24.53	22.69	25.11	28.51

Every generative arm gains more from a flexible readout than the contrastive controls do: the gap to the control shrinks from 8.67 (linear SGD) to 0.99 (MLP) for CoRe, and from 11.0 to 1.25 between

the two Gram/iBOT models. Whitening is an invertible affine preprocessing, so its gain reflects regularized finite fitting rather than new information; the MLP additionally changes capacity. A matched study across feature taps (backbone, projection before LayerNorm, and final representation) gives best-readout gaps to the control of 1.55, 1.42, and 1.47 mIoU, while the best purely linear gap grows from 1.55 at the backbone to 4.32 at the representation: the narrow contextual representation is where linear accessibility is lost.

H.2 Neighborhoods and connectivity

Fixed-rank enrichment. Normalize each nonzero contextual token to unit length. For a token i with label s_i (its source image, class, word, or word family), let $\mathcal{N}_k(i)$ be its k nearest tokens within an eligible set $\mathcal{E}(i)$, and define

$$E_s(k) = \frac{1}{T}\sum_i \frac{|\{j \in \mathcal{N}_k(i) : s_j = s_i\}|/k}{b_s(i)}, \qquad b_s(i) = \frac{|\{j \in \mathcal{E}(i) : s_j = s_i\}|}{|\mathcal{E}(i)|}, \tag{10}$$

so that 1 is chance and $1/b_s$ the ceiling. The eligible sets exclude the trivial matches: all other tokens for the same image, tokens of other images for the class, tokens of other captions for the word, and tokens of other captions and other words for the family. Every model contributes exactly k neighbors per token, so peaks at different radii or support cannot confound the comparison. The vision sample takes 24 ImageNet classes, 10 images per class, and 25 random patches per image (6000 tokens; each token has 24 same-image and 225 same-class, other-image companions), over three draws. The text sample is frequency-matched: at most 60 occurrences of each of 51 single-token words in seven families, one draw of 2711 occurrences shared by every model. As a scale check, the mean cosine between two patches of the same image is .274 for contrastive only and .025 for CoRe (.055 and .002 for different classes): CoRe's patch representations are also more spread overall, which rank-based enrichment is insensitive to.

Sampled connectivity and relative detour. Construct a union-symmetrized k-nearest-neighbor graph on normalized tokens with Euclidean edge weights. M is the largest-component fraction. From 48 sampled source vertices, compute shortest-path lengths h_{ij}; for reachable pairs with $d_{ij} > 10^{-6}$, set $\rho_{ij} = h_{ij}/d_{ij}$ and

$$D = \frac{\text{median}_{s_i \neq s_j,\,\text{reachable}}\ \rho_{ij}}{\text{median}_{s_i = s_j,\,\text{reachable}}\ \rho_{ij}}. \tag{11}$$

U is the unreachable between-label fraction over the same source/target pairs. The graphs use the frequency-matched text occurrences (1500 and 2500 occurrences $\times$ three seeds per k). Visual graphs are fully connected ($M = 1$) for both models. A graph vertex is a token occurrence whereas a contextual representation is a whole sequence, so these statistics describe the sampled token cloud.

Table 29: **Geometry at fixed rank.** Left: vision enrichment over chance among each patch's k nearest patches (same image) and among its k nearest patches of other images (same class); mean over three draws, s.d. $\leq$1.2 and $\leq$0.31. Middle: text enrichment among the ten nearest tokens of other captions (same word; ceiling 52.3) and of other captions and words (same family; ceiling 8.0). Right: text token graphs, mean $\pm$ s.d. over six draws; M largest-component fraction, U unreachable between-family fraction, D relative between-family detour.

	Vision patches				Text tokens		Text token graph				
	Same image		Same class		Word	Family	$k = 12$			$D, k = 6$	$D, k = 24$
Model	$k = 10$	$k = 50$	$k = 10$	$k = 50$	$k = 10$	$k = 10$	$M \uparrow$	$U \downarrow$	$D \downarrow$		
Contrastive only	83.3	27.9	9.01	6.69	51.4	5.88	.68±.02	.53±.04	1.86±.12	2.77±.29	1.49±.06
+ Gram/iBOT	86.8	29.5	9.33	6.93	51.4	5.92	.71±.10	.51±.16	1.73±.10	2.56±.49	1.40±.05
CoRe	6.5	3.0	2.01	1.48	50.8	5.80	.97±.03	.08±.05	1.24±.07	1.38±.04	1.25±.04
+ Gram/iBOT	6.7	3.1	2.15	1.54	50.5	5.81	.98±.01	.05±.05	1.13±.05	1.18±.05	1.21±.05

Under equal support the vision and text pictures differ (Table 29). Contrastive patch representations cluster by their source image, with a third of each patch's ten nearest patches from the same image, and by class; CoRe's patches are far less tied to either, yet remain above chance, and Gram/iBOT, which raises backbone ADE20K mIoU by about two points (Table 1), leaves both unchanged. Text tokens

keep same-word and same-family neighbors at the control's rate in every model, while both joint models form a nearly connected cloud with shorter between-family detours at every k; Gram/iBOT moves neither family appreciably, so connectivity is a property of the generative objective, not of the structural regularizer.

H.3 CONTEXTUAL SENSE SEPARATION

Table 30: **Contextual sense separation.** Leave-one-out 1-NN accuracy (cosine) on COCO captions for six polysemous words (senses balanced; sense chance 47%). *Word*: 1-NN word identity among all target occurrences of 20 word–sense groups. *Domain*: the nearest occurrence of a *different* word shares the meaning domain (e.g. *plate*/home plate → *pitcher*/baseball; chance 11%). representation $= Z_t$; backbone $= h_t$ before projection.

	contextual representation Z_t			Backbone h_t		
Text tower	Sense	Word	Domain	Sense	Word	Domain
Lookup embedding (any model)	47.2	100.0	–	47.2	100.0	–
Contrastive only	97.9	99.9	43.2	97.9	100.0	49.2
No contrastive loss	73.9	52.9	9.5	79.2	87.9	9.2
CoRe	**98.8**	**99.9**	23.2	99.0	100.0	45.3
CoRe + Gram/iBOT	97.3	100.0	20.0	97.8	100.0	39.3

We mine the 202K COCO val2014 captions for polysemous single-token words whose senses are identified by unambiguous cue words (e.g. *trunk*: {elephant}, {tree, log, stump}, {car, luggage, suitcase}), keeping a caption for a sense only if it contains the target exactly once, cues of that sense, and no cue of another sense. Senses with at least 20 captions are kept (at most 100 per sense): *plate* (dish, home plate), *pitcher* (baseball, jug), *trunk* (elephant, tree, car), *glasses* (eyewear, drinkware), *apple* (fruit, computer), and *mouse* (computer, animal), plus single-sense groups for *keyboard*, *bank*, *palm*, *net*, *bat*, *fan*, and *bow* that serve as cross-word neighbors (20 groups; 1,520 occurrences in which the word is a single token). For every occurrence we read the lookup embedding, the text backbone output h_t, and the representation Z_t. Sense accuracy is balanced leave-one-out 1-NN accuracy within a word, averaged over the six words (chance 47.2%); a logistic probe trained on half the images and tested on the other half gives the same ordering. Word accuracy is 1-NN word identity over all occurrences. Domain accuracy asks whether the nearest occurrence of a *different* word shares the meaning domain (baseball, dining, computing, animal, vehicle, plant, wearable; chance 11.1%). Because senses are defined by cue words present in the caption, sense separation shows that the token representation absorbs its context; the comparison across objectives, not the absolute level, is the informative part. Raw results are released with the paper.