GESSO: AN AESTHETIC PERCEPTUAL LOSS FOR PIXEL-SPACE TEXT-TO-IMAGE PRETRAINING

Anonymous authors
Paper under double-blind review



Figure 1: **Samples from PixelGen-SigLIP + GESSO at H/32**. Aesthetic signal is supplied via a dense, per-patch perceptual term, at no additional inference cost.

ABSTRACT

Aesthetic quality is conventionally added through data filtering, curation, preference finetuning, or reward optimization. Here, we ask whether a perceptual loss can supply aesthetic signal during pretraining. We study two design choices not previously examined in text-to-image pretraining. First, a system-level study of ten ViT-B encoders shows that language-supervised encoders perform best, while the self-supervised default, DINOv2, is last or tied for last on aesthetics, prompt alignment and fidelity. Swapping in SigLIP gives a stronger perceptual recipe. Second, how an aesthetic scorer enters the loss matters: with the scorer fixed, distilling it into a reference-anchored, per-patch distance improves held-out aesthetic metrics, whereas maximizing its score does not. We introduce **GESSO**, an aesthetic perceptual loss for pretraining: a lightweight projector over frozen SigLIP features maps a reference image and the model output to a distance that tracks the difference in their aesthetic scores, so aesthetic signal reaches every pretraining image rather than only a small curated set afterwards, while adding only 0.8% to step time. Added to the SigLIP recipe, GESSO improves every human-preference metric and is preferred by every VLM judge across three model scales. At B/32 the full recipe improves HPSv2.1 by 10.2% and reduces FID by 12.4% relative to the PixelGen recipe; GESSO alone contributes 4.3% and 6.0%. Prompt alignment is similar, and the gains persist after supervised finetuning.

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.

1 INTRODUCTION

Text-to-image models improve in aesthetic quality through data filtering (Schuhmann et al., 2022) or, after pretraining, through finetuning on a curated set (Dai et al., 2023; Chen et al., 2025a), preference finetuning (Wallace et al., 2024), or reward optimization (Black et al., 2024; Clark et al., 2024). An aesthetic term in the pretraining loss would complement these routes, yet none of them adds one: filtering changes only the data, and the post-training routes rely on three resources scarce at pretraining scale. Preference labels come from human annotators, so preference datasets stay orders of magnitude smaller (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023) than pretraining corpora (Schuhmann et al., 2022). Reward optimization starts from a pretrained model, which does not yet exist during pretraining, and needs in-loop sampling (Black et al., 2024) and a KL anchor (Fan et al., 2023) to keep the model near the data its scorer was fitted to (Gao et al., 2023). Finally, the strongest aesthetic scorers are multi-billion-parameter vision-language models (Wu et al., 2024b; Ma et al., 2025), too costly to backpropagate through at every pretraining step (Clark et al., 2024).

Yet perceptual losses already supply such a signal at pretraining scale: they need no labels, no pretrained model and no costly scorer. In pixel space, x-prediction makes the model's output an image at every step (Li & He, 2026), and PixelGen (Ma et al., 2026) applies two perceptual losses to that predicted image: LPIPS over VGG features for texture (Zhang et al., 2018) and a patch-level cosine over DINOv2 features for semantics (Oquab et al., 2024). Distances in such feature spaces track human judgments of perceptual similarity (Zhang et al., 2018; Stein et al., 2023). Aesthetic quality, a judgment of how detail is organized into content (Leder et al., 2004; Palmer et al., 2013), is diffuse in those feature spaces: in a PCA projection, high- and low-scoring images are largely mixed (App. A.2.1), and the models that capture aesthetic quality (Wu et al., 2024b; Ma et al., 2025) output one scalar per image that is anchored to no reference, and the models are too costly to run at every step. Pretraining lacks a perceptual loss whose error tracks aesthetic quality.

In this work, we show that aesthetics can be formulated as a dense perceptual loss for text-to-image pixel diffusion pretraining, and introduce **GESSO**, which has three components: an *aesthetic teacher*, a large preference model that scores images; a frozen SigLIP *encoder* that extracts patch features; and a lightweight *aesthetic projector* that maps those features to a space where distance tracks the teacher's aesthetic judgment. The projector is trained once, before pretraining. For each training image, a pretrained generator produces a *reconstruction*, a one-step prediction of the image from a noised copy, and the projector learns to make the per-patch distance between the image and its reconstruction track the gap between the teacher's scores for the two. During pretraining, GESSO applies this distance to the generator's prediction and its training image, so it estimates the aesthetic cost of the prediction's errors. The projector is frozen and reusable, and its features also help as a representation-alignment target (Sec. 4.9). Notably, it adds 0.8% to step time and is dropped at inference. The name reflects the role: gesso is the ground a painter lays before any paint, whereas post-training alignment is a final varnish. The following two findings shape this design.

Language-supervised encoders are better perceptual targets. Prior work inherited DINOv2 from class-conditional generation (Ma et al., 2026; Yu et al., 2026), where encoder comparisons rank DINOv2 first (Yu et al., 2025; Singh et al., 2026). A system-level comparison of ten ViT-B encoders from five training objectives shows that language-supervised encoders (SigLIP, SigLIP2, CLIP and the Perception Encoder, PE) lead on aesthetics and fidelity, while the self-supervised DINOv2 is last or tied for last on aesthetics, prompt alignment and fidelity; ImageNet linear-probe accuracy does not predict this ranking (cf. Singh et al., 2026). We hypothesize that this is because aesthetic judgment is expressed in language: captions name an image's light, palette and style, so an encoder trained to match captions keeps those attributes, whereas self-supervised objectives carry a weaker signal of them: their augmentations encourage invariance to colour, tone and sharpness, and DINOv2's training data was curated to resemble recognition benchmarks (Oquab et al., 2024). We refer to the PixelGen recipe with SigLIP in place of DINOv2 as PixelGen-SigLIP.

An anchored per-patch distance outperforms direct score maximization during pretraining. With the scorer (MUSIQ) held fixed, a reference-anchored, per-patch distance outperforms both maximizing the score as a reward and matching the prediction's score to the reference's. The distance is zero only when the prediction's features match its reference on every patch. In contrast, a maximized score has no reference, so nothing ties its optimum to a real image, and a matched score is satisfied by any image that scores like the reference, however different it looks.

Together, the two findings let GESSO apply an aesthetic signal to every pretraining image (Fig. 1). We train three model sizes on CC12M at 512px, B/32, L/32 and H/32 (260M to 1.88B parameters, patch size 32). At B/32, relative to PixelGen, PixelGen-SigLIP with GESSO improves HPSv2.1 by 10.2% and LAION-aesthetic by 7.1% and reduces FID by 12.4%. GESSO alone contributes about half of each (4.3%, 3.8% and 6.0%) with similar prompt alignment. The gains hold on PartiPrompts and persist through supervised finetuning. In pairwise comparisons, three vision-language model (VLM) judges not used in training prefer the GESSO arm to every baseline at every model size.

Contributions.

1. **Which encoder should supply perceptual supervision?** To our knowledge, this is the first study of perceptual-loss encoders for pixel-space text-to-image generation, over ten ViT-B encoders. DINOv2, the default inherited from class-conditional alignment studies, is last or tied for last on every column; SigLIP is first or tied for first and gives PixelGen-SigLIP (Sec. 4.2).
2. **How should an aesthetic scorer become a loss?** With the scorer fixed, maximizing it as a reward fails to raise held-out aesthetic metrics, whereas distilling it into a reference-anchored, per-patch distance raises every metric above the base (Sec. 4.3).
3. **GESSO, an aesthetic perceptual loss.** A lightweight projector over frozen SigLIP whose distance tracks the teacher's score gap between a reference and its reconstruction. Its loss is zero at the reference, so its zero-loss target is anchored to the reference image, not a scorer's maximum. GESSO adds 0.8% step time and nothing at inference (Sec. 3).
4. **Aesthetics as a perceptual loss.** Adding GESSO to PixelGen-SigLIP improves every human-preference metric and FID at three model scales with similar prompt alignment, and its lead persists through supervised finetuning (Secs. 4.4–4.7).

2 Related Work

Aesthetic alignment. Pretraining corpora are filtered by an aesthetic predictor (Schuhmann et al., 2022; Schuhmann, 2022). After pretraining, a finetuning set is curated or distilled (Dai et al., 2023; Chen et al., 2025a;b), or a preference or reward model is optimized (Wallace et al., 2024; Black et al., 2024; Fan et al., 2023; Clark et al., 2024). Both act on a distribution already fit. Dense, per-region feedback also improves generators after training but needs region labels (Liang et al., 2024). Aesthetic scorers (Talebi & Milanfar, 2018a; Ma et al., 2025) are maximized directly as rewards (Talebi & Milanfar, 2018b; Clark et al., 2024), which chases an absolute score and risks style collapse (Fan et al., 2023; Gao et al., 2023).

Perceptual losses. Pixel diffusion (Ho et al., 2020; Dhariwal & Nichol, 2021), which forgoes the latent autoencoder of most text-to-image systems (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024), became competitive with latent diffusion through the x-prediction transformer of Li & He (2026). Applying a perceptual loss to that *predicted clean sample* is recent: Lin & Yang (2024) use the model's own features, LPL (Berrada et al., 2025) compares clean *latents*, and Ma et al. (2026) bring the loss to pixel space, pairing LPIPS for texture with a patch-level DINOv2 cosine for semantics. These losses measure how similar the prediction is to its reference, not how aesthetic it is. Learned metrics such as DreamSim (Fu et al., 2023) extend similarity beyond low-level texture but still do not measure aesthetic quality.

Choosing the encoder. GAN discriminators have long used frozen encoders (Sauer et al., 2021; Kumari et al., 2022; Sauer et al., 2023; 2024). For diffusion, which encoder to use in either role has been asked, though not in text-to-image generation: REPA (Yu et al., 2025), representation alignment, ranks nine encoders as alignment targets and puts DINOv2 first, iREPA extends it to 27 (Singh et al., 2026), and Ghildyal et al. (2025) and Hernández-Cámara et al. (2025) compare backbones as perceptual *distances*; none measures aesthetic signal, and both alignment studies select on class-conditional ImageNet. Ma et al. (2026) and Yu et al. (2026) carried DINOv2 into text-to-image, for the perceptual term and for alignment, without comparing encoders there. One contrary datapoint exists: REPA with a CLIP target beats DINOv2 on CC12M (Li et al., 2026a).

Measuring aesthetic quality. Aesthetic quality is measured with predictors that score an image alone (Schuhmann, 2022), preference models that score it against its prompt, and VLM judges. Preference models rank text-to-image models in close agreement with humans (Kirstain et al., 2023; Wu et al., 2023; Xu et al., 2023; Ma et al., 2025), and VLM judges approach inter-annotator agreement on generated images (Wu et al., 2024a; Huang et al., 2024; Ku et al., 2024; Zhang et al., 2023; Li

et al., 2026b); both are standard instruments of reward finetuning (Black et al., 2024; Clark et al., 2024; Yuan et al., 2024; Li et al., 2025).

3 METHOD

3.1 PRELIMINARIES

Flow matching (Lipman et al., 2023) interpolates $x_t = (1-t)\,x_0 + t\,s\varepsilon$ between a clean image x_0 ($t = 0$) and noise $s\varepsilon$ ($t = 1$, scale $s = 2$) and regresses the constant velocity $s\varepsilon - x_0$. JiT (Li & He, 2026) has the network f_θ predict the clean image $f_\theta(x_t)$ (text conditioning omitted) and reads the velocity off that prediction:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\,x_0,\,\varepsilon}\left\| \tfrac{x_t - f_\theta(x_t)}{t} \;-\; (s\varepsilon - x_0)\right\|^2. \tag{1}$$

This x-prediction makes large patch sizes viable in pixel space and gives a clean-image estimate at every noise level, so an image-space loss applies to it directly, without latent decoding.

A perceptual loss minimizes the error between the prediction and its reference x_0 in the feature space of a frozen pretrained encoder (Johnson et al., 2016; Zhang et al., 2018), and can be added to the flow-matching objective (Lin & Yang, 2024; Berrada et al., 2025; Ma et al., 2026):

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{FM}} + \sum_k \lambda_k\, \mathcal{L}_{E_k}(f_\theta(x_t),\, x_0) \tag{2}$$

The encoder E_k of each term decides which errors are penalized. We study two choices that Eq. 2 leaves open: the encoder and how to build an aesthetic term of this form.

3.2 THE AESTHETIC PROJECTOR

The aesthetic projector turns a frozen backbone into a distance d_{aes} that tracks how far apart a frozen *aesthetic teacher* S, a preference model such as HPSv3, scores two images.

Architecture. The projector reads a frozen SigLIP ViT-B/16 backbone at its final block, one token per image patch, $\phi(x) \in \mathbb{R}^{|\mathcal{P}|\times 768}$ over the patch grid $\mathcal{P}$, and maps each token through the three-layer MLP projector h_ψ (Yu et al., 2025) to a per-patch embedding $z^p(x) \in \mathbb{R}^{768}$. The aesthetic distance between two images x and y is the mean cosine distance of their ℓ_2-normalized embeddings $\hat{z}$:

$$d_{\text{aes}}(x, y) \;=\; \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\big(1 - \langle \hat{z}^{\,p}(x),\, \hat{z}^{\,p}(y)\rangle\big) \tag{3}$$

The backbone tokens have already passed through global self-attention, so each token is a spatially indexed but globally contextualized descriptor. Applied per patch, the projector reshapes the distances in the backbone's feature space without mixing tokens.

Training data. The projector is trained on *reconstruction sets*: a reference image x_0 and three reconstructions $\tilde{x}$, each one pass of a pretrained generator on x_0 at a different noise level. A reconstruction set spans mild to severe degradation of one image, so the gap regression sees the full range of aesthetic cost for each image. The teacher S scores every image once. The training target is the *aesthetic gap* $g = |S(x_0) - S(\tilde{x})|$, unsigned because a cosine distance is non-negative. We rescale g to $\bar{g} \in [0, 1]$ using its 5th and 95th percentiles on the training split (App. B.1). Where LPIPS is fitted to human judgments of how similar two images look (Zhang et al., 2018), the projector is fitted to how far apart the teacher scores the two images. We deploy HPSv3 (Ma et al., 2025) as the teacher, and ablate two other teachers.

Objective. We train the projector and a score head with four terms:

$$\mathcal{L}_{\text{proj}} = \underbrace{\text{Huber}\big(d_{\text{aes}}(x_0, \tilde{x}),\, \bar{g}\big)}_{\mathcal{L}_{\text{gap}}\text{: magnitude}} + \lambda_{\text{rank}}\underbrace{\mathcal{L}_{\text{rank}}}_{\text{ranking}} + \lambda_{\text{score}}\underbrace{\mathcal{L}_{\text{score}}}_{\text{readout}} + \lambda_{\text{anc}}\underbrace{\mathcal{L}_{\text{anc}}}_{\text{geometry}} \tag{4}$$

The first term regresses the aesthetic distance between a reference and its reconstruction onto the normalized gap, so the distance becomes an estimate of the normalized gap, which is linear in the teacher's score gap between its 5th and 95th percentiles. $\mathcal{L}_{\text{rank}}$ orders d_{aes} across different images

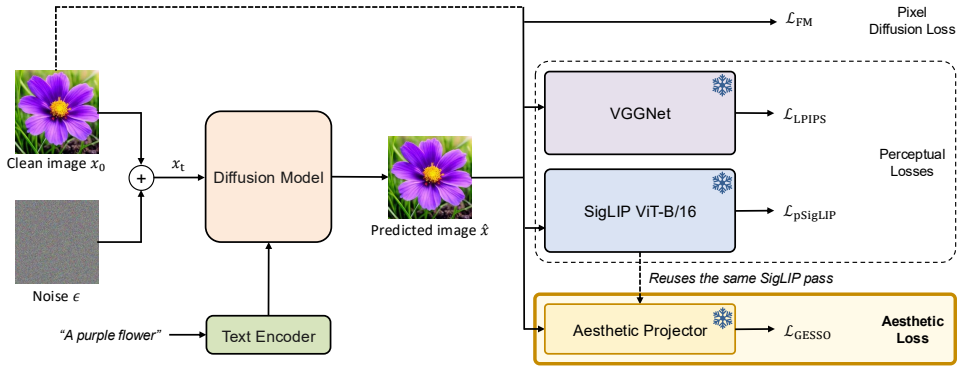

Figure 2: **GESSO in pixel-space diffusion pretraining.** The generator's predicted clean image passes through two frozen perceptual encoders, VGG for $\mathcal{L}_{\text{LPIPS}}$ and SigLIP for $\mathcal{L}_{\text{pSigLIP}}$ to be compared against its reference image. GESSO adds the trained frozen aesthetic projector on SigLIP features, so the aesthetic term reuses the pass that $\mathcal{L}_{\text{pSigLIP}}$ already computes.

reconstructed by the same generator at the same noise level by their gaps g, so the ordering must come from the teacher rather than from the generator or the noise level. $\mathcal{L}_{\text{score}}$ regresses the normalized teacher score from the mean-pooled embedding through a SiLU and linear head, an auxiliary training head discarded afterwards. $\mathcal{L}_{\text{anc}}$ anchors the projector's patch-to-patch cosine geometry to the raw backbone tokens, preserving SigLIP's spatial structure. Although the teacher labels whole images, each patch contributes its own term to the pooled distance.

3.3 GESSO: An Aesthetic Perceptual Loss for Pretraining

GESSO applies the projector's distance to the generator's prediction like any perceptual loss, with the projector and backbone frozen: $\mathcal{L}_{\text{GESSO}}(f_\theta(x_t), x_0) = d_{\text{aes}}(f_\theta(x_t), x_0)$ (Fig. 2). The full pretraining objective adds GESSO to two fixed perceptual terms:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \underbrace{\lambda_1 \mathcal{L}_{\text{LPIPS}}(f_\theta(x_t),\, x_0)}_{\text{local texture}} + \underbrace{\lambda_2 \mathcal{L}_{\text{pSigLIP}}(f_\theta(x_t),\, x_0)}_{\text{global semantics}} + \underbrace{\lambda_3 \mathcal{L}_{\text{GESSO}}(f_\theta(x_t),\, x_0)}_{\text{aesthetics}} \tag{5}$$

$\mathcal{L}_{\text{LPIPS}}$ compares multi-level VGG activations for local texture, and $\mathcal{L}_{\text{pSigLIP}}$, a patch-level cosine on last-layer SigLIP features, carries global semantics. Without GESSO, Eq. 5 is the recipe we call PixelGen-SigLIP; the PixelGen arms replace $\mathcal{L}_{\text{pSigLIP}}$ with $\mathcal{L}_{\text{pDINO}}$ at the same weight. The arm trained with the full objective is written + GESSO in tables and called the GESSO arm in the text.

We set $\lambda_3 = 0.04$ (swept in Table 13) and keep λ_1 and λ_2 at the PixelGen values ($\lambda_1 = 0.1$, $\lambda_2 = 0.01$). The GESSO arm trains as PixelGen-SigLIP for its first 100k steps and adds GESSO for the remaining 150k, so every arm trains 250k steps in total. Perceptual terms apply only at noise levels $t \leq 0.7$ ($t = 0$ is clean). GESSO adds 0.8% to step time and no measurable increase in peak GPU memory, reusing the SigLIP pass of $\mathcal{L}_{\text{pSigLIP}}$.

3.4 Implementation Details

Arms. We call each trained configuration an *arm*. Every generator is MiniT2I (Wang et al., 2026), the multimodal implementation of JiT: an x-prediction transformer at patch size 32, trained with flow matching, conditioned on frozen FLAN-T5, with no time conditioning (Sun et al., 2025), pretrained for 250k steps on CC12M (Changpinyo et al., 2021) with LLaVA-NeXT recaptions at 512px. B/32, L/32 and H/32 have 260M, 915M and 1.88B parameters, plus a frozen 341M FLAN-T5-Large encoder (Table 15). *MiniT2I* alone denotes flow matching with no perceptual term. PixelGen denotes our reimplementation of the recipe of Ma et al. (2026): the baseline objective with LPIPS at $\lambda_1 = 0.1$ and $\mathcal{L}_{\text{pDINO}}$, a patch-level cosine over DINOv2 (Oquab et al., 2024), at $\lambda_2 = 0.01$. Our reimplementation omits PixelGen's representation-alignment term (Yu et al., 2025), so that every difference between arms comes from the pixel-space perceptual losses.

The aesthetic projector. The reconstruction sets are built from CC12M (Changpinyo et al., 2021) images, $\sim$427.5k for training and $\sim$22.5k held out, with reconstructions from MiniT2I, PixelGen and PixelGen-SigLIP at the B/32, L/32 and H/32 scales at their final checkpoints. Images are centre-cropped, loaded at 512px and downsampled to 256px, giving $|\mathcal{P}| = 16 \times 16 = 256$ patch tokens. The projector has $\sim$7.3M parameters.

Inference. We adopt MiniT2I's sampler unchanged (Wang et al., 2026): we generate 512px samples over 100 uniform Euler steps from Gaussian noise at the training noise scale, converting the predicted clean image to a velocity at each step and clamping it to the pixel range. Classifier-free guidance masks the text tokens, and we report the performance with the EMA weights.

4 RESULTS

4.1 SETUP

Metrics. We measure aesthetic quality with four human-preference models, each fitted to human judgments collected independently of this work: HPSv2.1 (Wu et al., 2023), LAION-aesthetic (Schuhmann, 2022), PickScore (PS; Kirstain et al., 2023) and ImageReward (IR; Xu et al., 2023), the standard outcome measures of preference finetuning (Wallace et al., 2024; Clark et al., 2024), on the HPS v2 prompts and on PartiPrompts (Yu et al., 2022). Three VLM judges also compare samples pairwise, a protocol validated against human raters (Wu et al., 2024a;b; Huang et al., 2024; Ku et al., 2024; Zhang et al., 2023; Peng et al., 2025; Chen et al., 2025d). The teacher, HPSv3, is not a metric. Prompt alignment and fidelity follow MiniT2I's protocol (Wang et al., 2026): GenEval (Ghosh et al., 2023), DPGBench (Hu et al., 2024) and MSCOCO-30k FID (Heusel et al., 2017).

Protocol. GenEval, DPGBench and FID are read at each arm's own guidance optimum, and preference numbers at one common guidance scale, $w = 3$, which three of the four arms select in independent per-arm tournaments at H/32 after supervised finetuning (App. B.4). Means and *seed deviations* are over three generation seeds of one trained model; bold marks a block's best cell and every cell whose mean is within its own seed deviation of the best; a deviation below the last reported digit is shown as one unit of it. At B/32, three independent PixelGen-SigLIP pretrains have across-run deviations of 0.12 HPSv2.1, 0.005 LAION-aesthetic, 0.019 GenEval, and 0.02 FID; matched GESSO branches are analyzed in App. A.1.4.

4.2 FINDING 1: LANGUAGE-SUPERVISED ENCODERS ARE BETTER PERCEPTUAL TARGETS

A perceptual loss minimizes error in a frozen encoder's feature space, so the generator is asked to get right only what that encoder distinguishes. Holding LPIPS and the recipe fixed, we evaluate ten pretrained ViT-B encoders as the perceptual-loss encoder, each read from its last layer at a matched budget of 196 patch tokens (Table 1; App. B.5).

Table 1: **Pretrained encoders for the perceptual loss.** All ViT-B, last layer, 196 tokens, DINOv2-tuned loss weight (Ma et al., 2026).

	Aesthetics		Prompt align.	Fidelity
Encoder	HPSv2.1↑	LAION↑	GenEval↑	FID↓
Self-distillation				
DINOv2	20.04 ±0.06	5.210 ±0.006	0.581 ±0.015	10.81 ±0.04
DINOv3	20.73 ±0.02	5.374 ±0.013	0.595 ±0.012	10.16 ±0.03
Language supervision				
PE	20.76 ±0.01	5.387 ±0.011	0.599 ±0.008	**9.89** ±0.07
CLIP	20.85 ±0.08	5.385 ±0.008	**0.606** ±0.007	10.10 ±0.01
SigLIP	**21.03** ±0.02	**5.427** ±0.007	**0.607** ±0.009	**9.87** ±0.01
SigLIP2	**21.00** ±0.05	5.414 ±0.011	**0.597** ±0.012	10.11 ±0.05
Reconstruction				
Pixio	20.62 ±0.06	5.365 ±0.008	**0.609** ±0.002	10.23 ±0.03
MAE	20.43 ±0.03	5.318 ±0.004	0.585 ±0.007	10.44 ±0.06
Latent prediction				
LeJEPA	20.68 ±0.03	5.350 ±0.017	0.578 ±0.020	10.50 ±0.07
Multi-teacher distillation				
EUPE	20.78 ±0.02	5.346 ±0.004	**0.609** ±0.010	10.10 ±0.04

Language-supervised encoders lead as a group. SigLIP, SigLIP2, CLIP and PE hold the top four places on LAION-aesthetic, and together with EUPE the top five on HPSv2.1 and FID. SigLIP is first or tied for first on all four columns. DINOv2, the default in prior work (Yu et al., 2025; Ma et al., 2026), is last or within seed deviation of last on all four. Its gap is also the clearest in the table: 0.39 HPSv2.1 below the next encoder (about three training-run deviations) whereas most neighbouring encoders lie within one training-run deviation of one another, so we read the ranking by group rather than by position. The loss weight was tuned for DINOv2 and not retuned per encoder. We hypothesize that captions name colour, contrast and style, which language supervision keeps and self-distillation's augmentations discourage.

4.3 FINDING 2: AN ANCHORED PER-PATCH DISTANCE BEATS A MAXIMIZED REWARD

To see how a learned scorer should enter the loss, we hold the scorer, MUSIQ (Ke et al., 2021), fixed and vary how its term is built: whether the term is anchored to a reference image, and whether it is resolved per patch. Each variant is added to, and compared against, PixelGen-SigLIP (Table 2).

Scalar reward maximizes the prediction's score, the reward-finetuning form (Black et al., 2024; Clark et al., 2024) without its reference model and anchor, so nothing ties its optimum to a real image.

Neither metric held out from training rises (HPSv2.1 -0.16, LAION-aesthetic -0.035, one training run each), and the maximized score rises slightly ($4.250 \rightarrow 4.271$).

Scalar gap instead minimizes the squared score difference between prediction and target, so the reference zeroes the loss, but so does every image the scorer rates the same: one number covers the whole image, so an oversaturated sky, for example, can pay for a blurry face. MUSIQ rises ($4.250 \rightarrow 4.329$), but the held-out metrics do not.

Table 2: **Turning an aesthetic score into a training signal.** Every arm below *none* adds a MUSIQ term to the same PixelGen-SigLIP base; the per-patch arm distils MUSIQ into a distance over frozen SigLIP.

	held out		optimized
MUSIQ used as	HPSv2.1↑	LAION↑	MUSIQ↑
none	21.12 ±0.05	5.407 ±0.008	4.250 ±0.010
scalar reward	20.96 ±0.04	5.372 ±0.011	4.271 ±0.004
scalar gap	21.04 ±0.06	5.377 ±0.011	4.329 ±0.004
per-patch	**21.49** ±0.09	**5.494** ±0.012	**4.345** ±0.006

Per-patch distance keeps the reference and distils the scorer into a distance over patch embeddings, which is zero only when the prediction matches the reference at every patch in the projector's feature space: each patch adds a non-negative term, so a match in one region cannot offset an error in another. All three metrics improve by many seed deviations, the optimized column included, where the distilled form outscores direct maximization (4.345 against 4.271). This arm is GESSO's construction with MUSIQ as the teacher, at $\lambda_3 = 0.01$. The gain is not simply more SigLIP weight: an untrained projector, the same per-patch distance over SigLIP with nothing learned from a scorer, reaches 21.30 HPSv2.1, and distilling MUSIQ into it raises this to 21.49 (App. A.1.5).

4.4 Aesthetic quality

Table 3 places the GESSO arm on a ladder: PixelGen adds two perceptual terms to MiniT2I, PixelGen-SigLIP swaps DINOv2 for SigLIP, and the GESSO arm adds GESSO. End to end, the full recipe improves HPSv2.1 by 10.2% and LAION-aesthetic by 7.1% over PixelGen and reduces FID by 12.4%; against MiniT2I, by 13.8%, 9.4% and 25.2%.

Table 3: **The baseline ladder at B/32, before and after supervised finetuning**, on an aesthetic (HPS v2) and a neutral (PartiPrompts) prompt set. FID after pretraining only.

	Human-preference scorers								Prompt align.		Fidelity
	HPS v2 benchmark				PartiPrompts						
Arm	HPSv2.1↑	LAION↑	PS↑	IR↑	HPSv2.1↑	LAION↑	PS↑	IR↑	GenEval↑	DPGBench↑	FID↓
After pretraining											
MiniT2I†	19.36 ±0.07	5.131 ±0.007	19.40 ±0.01	−0.219 ±0.015	21.20 ±0.07	4.982 ±0.019	20.31 ±0.01	−0.120 ±0.031	0.502 ±0.012	78.22 ±0.08	12.88 ±0.03
PixelGen	20.00 ±0.02	5.241 ±0.008	19.61 ±0.01	−0.033 ±0.003	21.99 ±0.02	5.125 ±0.005	20.58 ±0.02	0.147 ±0.018	0.600 ±0.007	81.80 ±0.31	10.99 ±0.12
PixelGen-SigLIP	21.12 ±0.05	5.407 ±0.008	19.89 ±0.01	0.173 ±0.009	22.84 ±0.06	5.236 ±0.012	20.80 ±0.01	0.347 ±0.015	**0.607** ±0.011	**82.33** ±0.65	10.25 ±0.05
+ GESSO	**22.03** ±0.06	**5.615** ±0.012	**20.14** ±0.01	**0.325** ±0.019	**23.61** ±0.06	**5.427** ±0.020	**20.98** ±0.02	**0.460** ±0.004	**0.608** ±0.011	**82.72** ±0.17	**9.63** ±0.06
After supervised finetuning											
MiniT2I†	25.95 ±0.03	6.395 ±0.010	20.97 ±0.01	0.885 ±0.006	26.62 ±0.02	6.105 ±0.006	21.68 ±0.01	0.905 ±0.007	0.825 ±0.003	83.17 ±0.10	—
PixelGen	26.58 ±0.02	6.479 ±0.002	21.22 ±0.01	0.969 ±0.003	27.40 ±0.01	6.193 ±0.009	21.94 ±0.01	1.012 ±0.004	0.846 ±0.001	85.15 ±0.06	—
PixelGen-SigLIP	27.41 ±0.04	6.590 ±0.016	21.43 ±0.01	1.043 ±0.003	28.05 ±0.02	6.246 ±0.008	22.12 ±0.01	1.091 ±0.002	**0.856** ±0.003	**85.54** ±0.09	—
+ GESSO	**27.97** ±0.01	**6.635** ±0.005	**21.60** ±0.01	**1.085** ±0.006	**28.56** ±0.04	**6.299** ±0.013	**22.26** ±0.01	**1.107** ±0.004	0.852 ±0.002	85.35 ±0.16	—

† Our reproduced MiniT2I baseline matches or exceeds the performance reported in Wang et al. (2026).

The two steps contribute in roughly equal measure. The encoder comparison matches every encoder at 196 patch tokens; here SigLIP is read at its native 256. At that budget, replacing DINOv2 (PixelGen) with SigLIP (PixelGen-SigLIP) still improves HPSv2.1 by 5.6% and LAION-aesthetic by 3.2% and lowers FID by 6.7%; GESSO then adds a further 4.3%, 3.8% and 6.0% on top of that stronger baseline, and improves all four human-preference columns on both prompt sets. The gains hold on PartiPrompts, a set not selected for aesthetic content, and cost nothing on the other axes: GenEval and DPGBench are unchanged within seed deviation, and FID improves. The gain replicates across three independent pretrains ($+0.91$ to $+1.09$ HPSv2.1, FID -0.62 to -0.89; App. A.1.4).

Table 4: **The ladder at every scale.** Preference columns and DPGBench in Table 8. PG-SigLIP: PixelGen-SigLIP. †After finetuning.

Scale	Arm	HPSv2.1↑	LAION↑	GenEval† ↑	FID↓
B/32	MiniT2I	19.36 ±0.07	5.131 ±0.007	0.825 ±0.003	12.88 ±0.03
	PixelGen	20.00 ±0.02	5.241 ±0.008	0.846 ±0.001	10.99 ±0.12
	PG-SigLIP	21.12 ±0.05	5.407 ±0.008	**0.856** ±0.003	10.25 ±0.05
	+ GESSO	**22.03** ±0.06	**5.615** ±0.012	0.852 ±0.002	**9.63** ±0.06
L/32	MiniT2I	20.40 ±0.02	5.318 ±0.013	0.858 ±0.001	12.47 ±0.04
	PixelGen	21.81 ±0.05	5.560 ±0.001	**0.867** ±0.002	10.08 ±0.04
	PG-SigLIP	22.32 ±0.02	5.603 ±0.011	**0.866** ±0.005	9.57 ±0.01
	+ GESSO	**23.15** ±0.01	**5.774** ±0.001	**0.865** ±0.005	**9.29** ±0.04
H/32	MiniT2I	22.10 ±0.04	5.632 ±0.012	**0.879** ±0.007	10.50 ±0.03
	PixelGen	21.68 ±0.04	5.543 ±0.003	0.858 ±0.001	10.30 ±0.06
	PG-SigLIP	23.17 ±0.01	5.742 ±0.010	0.880 ±0.002	9.42 ±0.04
	+ GESSO	**23.84** ±0.01	**5.862** ±0.009	**0.884** ±0.001	**9.06** ±0.10

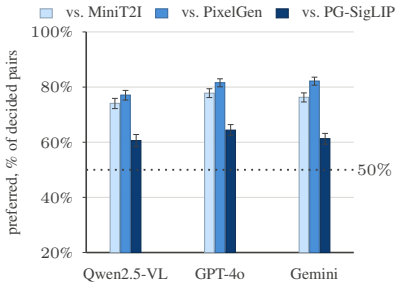

Figure 3: **The GESSO arm is preferred, H/32 after pretraining.** Whiskers are 95% Wilson intervals.

Training cost. The gains cost little (Table 5). The perceptual losses the recipe inherits are its main overhead: they take MiniT2I from 427 ms/step and 49 GiB to the recipe's 605 and 67 (+42% time, +37% memory). The encoder swap adds 0.2% of step time and GESSO a further 0.8%, with no measurable increase in peak GPU memory, because neither runs a new backbone: the swap substitutes one frozen ViT-B for another, and GESSO reuses the SigLIP pass that $\mathcal{L}_{\text{pSigLIP}}$ already computes.

Table 5: **Training cost and quality at B/32 measured on 8× H200 GPUs**, cells: value over change vs. the grey PixelGen row.

	Time	Memory	Quality	
Arm	ms/step↓	GiB/GPU↓	HPSv2.1↑	FID↓
PixelGen	605	67	20.00	10.99
PixelGen-SigLIP	606 (+0.2%)	67 (+0.0%)	21.12 (+5.6%)	10.25 (−6.7%)
+ GESSO	611 (+1.0%)	67 (+0.0%)	22.03 (+10.2%)	9.63 (−12.4%)

4.5 Composition with supervised finetuning

A perceptual loss should complement supervised finetuning (SFT) rather than replace it: SFT changes the corpus the model trains on, while the loss anchors the model to whichever corpus that is. To test this, every arm in Table 3 continues for 40k further steps on the same 119k-image synthetic set, MiniT2I's released finetuning mix of GPT-4o and DALL·E 3 images (Chen et al., 2025a;b; OpenDatasets, 2024). Each arm keeps its pretraining losses (App. B.3), so the lower block of the table compares complete pipelines. Following Wang et al. (2026), we do not report FID after SFT, as the synthetic corpus moves every arm away from the MSCOCO distribution. SFT lifts every arm, but the GESSO arm still leads PixelGen by 5.2% HPSv2.1 and 2.4% LAION-aesthetic, and PixelGen-SigLIP by 2.0% and 0.7%, leading both on all eight human-preference columns. Prompt alignment changes little: the GESSO arm trails PixelGen-SigLIP by 0.004 GenEval and 0.19 DPGBench after SFT, just beyond one seed deviation, and leads PixelGen by 0.006 and 0.20.

4.6 Scaling with model size

The GESSO arm scales with model size across aesthetics and fidelity (Table 4). Against PixelGen the margin stays wide at every scale (HPSv2.1 +1.34 to +2.16, LAION-aesthetic +0.214 to +0.374, FID −0.79 to −1.36), and against PixelGen-SigLIP, which isolates GESSO, HPSv2.1 gains +0.91, +0.83 and +0.67, with the B/32 gain replicated across three pretrains ($p < 0.05$). GenEval is reported after finetuning, following Wang et al. (2026). The GESSO arm leads or ties both prompt-alignment benchmarks at L/32 and H/32, improves from B/32 to H/32 on every column, and at H/32 wins or ties every column it is measured on. In contrast, PixelGen-SigLIP stays above MiniT2I at every scale, while PixelGen drops below MiniT2I on every aesthetic column at H/32 (App. A.1.1).

4.7 Pairwise preference

The pointwise scorers read one image at a time, so the second instrument puts the pretrained H/32 GESSO arm before three VLM judges choosing between paired samples on the 3200 HPS v2 prompts, both orders shown and counted only on agreement (App. B.6). The GESSO arm is preferred in 74–78% of decided pairs against MiniT2I, 77–82% against PixelGen, and 61–64% against PixelGen-SigLIP, the comparison that isolates GESSO, with every interval excluding 50% (Fig. 3). All 27 judge–scale–baseline comparisons favour the GESSO arm (App. A.1.2).

4.8 PLACEMENT AGAINST PUBLISHED SYSTEMS

Every comparison so far holds model family and recipe fixed. Table 6 places the GESSO arm against the PRISM-Bench leaderboard (Fang et al., 2026), grouped by generation space, and against the published MiniT2I models at patch 16 (Wang et al., 2026), which share our architecture but use $4\times$ more image tokens. Among the pixel-space systems listed, our H/32 arm is the strongest. B/32 exceeds MiniT2I-B/16 on all three headline columns, L/32 leads MiniT2I-L/16 on alignment while trailing it by 0.6 points in average score and 1.4 points in aesthetics, and H/32 leads it on all three. The latent systems other than SD1.5 generate at 1024px or more, so we report where they sit rather than compete with them; SD1.5, at our 512px, trails all three of our arms; Janus-Pro-7B, autoregressive at 384px, leads our best arm on aesthetics by 0.2.

Table 6: **Comparison on PRISM-Bench (Fang et al., 2026).** AR: autoregressive. Ours after SFT; judge Qwen2.5-VL-72B; per-track: Table 9.

Space	Model	Res	Avg	Align	Aes
AR	GPT-Image-1 [High] (OpenAI, 2025b)	1024	80.7	82.7	78.7
	Janus-Pro-7B (Chen et al., 2025c)	384	62.1	64.9	59.4
Latent	Qwen-Image (Qwen Team, 2025)	1328	74.1	80.0	68.3
	FLUX.1-dev (Black Forest Labs, 2024)	1024	68.5	72.1	64.9
	SD3-Medium (Esser et al., 2024)	1024	66.1	68.0	64.2
	SDXL (Podell et al., 2024)	1024	57.0	60.1	54.0
	SD1.5 (Rombach et al., 2022)	512	46.0	48.8	43.3
Pixel	PixelGen-SigLIP + GESSO (H/32)	512	63.4	67.7	59.2
	MiniT2I-L/16 (Wang et al., 2026)	512	62.4	65.9	58.9
	PixelGen-SigLIP + GESSO (L/32)	512	61.8	66.1	57.5
	PixelGen-SigLIP + GESSO (B/32)	512	57.6	62.0	53.3
	MiniT2I-B/16 (Wang et al., 2026)	512	55.8	61.3	50.2

4.9 ANALYSIS OF THE PROJECTOR

What GESSO optimizes. GESSO pulls each prediction toward its own training image, not toward the teacher's highest score. The distance is zero at the reference, and its target is the size of the score gap regardless of direction, so no image scores better than the reference itself. What GESSO adds is an aesthetic weighting: errors that cost more aesthetically are penalized more. The generator therefore keeps more of its data's aesthetic quality, and when SFT moves training to a more aesthetic corpus, GESSO continues to anchor the model to that corpus (Sec. 4.5).

What makes a good teacher. App. A.1.5 ablates the teacher, the training step when the GESSO loss starts being applied, and λ_3. The noise level already tells the projector how degraded a reconstruction is, so a good teacher must rank aesthetic quality beyond degradation. At $\lambda_3 = 0.01$, a projector trained on noise order alone, which also drops the other projector terms, lands 0.54 HPSv2.1 below the HPSv3 projector, and an untrained projector recovers only about a third of its gain, so the gain is not merely more weight on SigLIP features. Among teachers that rank aesthetic quality, the choice is second-order for HPSv2.1: HPSv3, Q-Align and MUSIQ land within about one training-run deviation, with HPSv3 giving the best FID.

The projector as an alignment target. A learned loss can raise metrics by letting the generator exploit the network it trains through. To rule this out, we drop the loss and use the projector only as a REPA (Yu et al., 2025) target: the generator's intermediate features are regressed onto the projector's features of the clean training image, so its outputs never pass through the projector and the objective contains no aesthetic term (Table 7). The projector is the only target that improves all four metrics, raising HPSv2.1 from 19.36 to 19.55 and lowering FID from 12.88 to 12.40. Raw SigLIP, the same backbone without the projector, leaves both aesthetic scores unchanged, so the aesthetic gain comes from training the projector. DINOv2, the inherited target, lowers FID as in prior work (Yu et al., 2025; Singh et al., 2026) but costs 0.38 HPSv2.1, consistent with Finding 1, and the projector beats it even on FID.

Table 7: **The aesthetic projector as a REPA target.** B/32, flow matching plus alignment only; the target varies.

Target	HPSv2.1↑	LAION↑	GenEval↑	FID↓
none	19.36 ± 0.07	5.131 ± 0.007	0.502 ± 0.012	12.88 ± 0.03
DINOv2	18.98 ± 0.07	5.048 ± 0.015	0.486 ± 0.007	12.53 ± 0.07
SigLIP	19.36 ± 0.05	5.133 ± 0.025	0.524 ± 0.011	12.69 ± 0.02
Aes. proj.	**19.55** ± 0.06	**5.207** ± 0.022	**0.543** ± 0.014	**12.40** ± 0.01

5 CONCLUSION

We show that a perceptual loss can improve aesthetic quality during pretraining, and that two choices determine how much it helps. First, language-supervised encoders are better perceptual targets than the default, DINOv2. Second, a scorer helps when distilled into a reference-anchored, per-patch distance, not when its score is maximized. GESSO combines the two at 0.8% extra step time and no inference cost, and improves every human-preference scorer at three model scales and keeps its lead even after SFT on a curated aesthetic dataset.

AI USE STATEMENT

Generative models as experimental apparatus. Large multimodal models are part of the measurement apparatus of this paper, and the role of each model is described where the model is used. Three vision-language models, Qwen2.5-VL-72B, GPT-4o and Gemini 3.1 Pro, judge the aggregate pairwise comparisons (Sec. 4.7, App. A.1.2); the guidance tournament uses Qwen2.5-VL-72B alone. All judges work under the prompts reproduced verbatim in Apps. B.4 and B.6. The same protocol selects the guidance scale at which every preference number is read (App. B.4). The aesthetic projector is trained against a frozen vision-language preference model, HPSv3, as its teacher (Sec. 3.2). These models are part of the method and the evaluation, and every number the models produce is reported with its protocol.

Model-generated data. Three of the datasets this paper trains on contain model-generated content. The pretraining corpus pairs CC12M images with captions generated by LLaVA-NeXT, inherited unchanged from the released MiniT2I setup (Wang et al., 2026). The aesthetic projector's reconstruction sets pair real CC12M images with reconstructions generated by our own pretrained diffusion arms (Sec. 3.2), labelled by the frozen teacher above. The supervised-finetuning corpus consists entirely of images generated by GPT-4o and DALL·E 3, MiniT2I's released finetuning mix (Chen et al., 2025a;b; OpenDatasets, 2024) (Sec. 4.5).

Generative AI in the research. In addition to the model-generated data above, we used generative AI tools to assist with writing and debugging code for training and evaluation, to provide feedback on research methodology and experiments, to help interpret results, and to edit manuscript text for clarity and presentation. We did not use generative AI tools to develop the method's conceptual framework, propose or refine hypotheses, design the experiments, clean or reformat datasets, or for translation. The paper contains no mathematical claims or proofs, and no qualitative or thematic data analysis, so those tasks are not applicable.

Responsibility. We have reviewed all AI-assisted work. AI-written code was reviewed and tested by the authors; every reported number was checked by the authors against the run logs and outputs; and AI-suggested interpretations and wording were checked against the results. We take responsibility for the final content of this paper, including its text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human-subjects research. All training data are public: CC12M images with LLaVA-NeXT recaptions (Changpinyo et al., 2021; Wang et al., 2026) and MiniT2I's released finetuning mix of GPT-4o and DALL·E 3 images (Chen et al., 2025a;b; OpenDatasets, 2024). Web-scraped corpora such as CC12M can contain biased, copyrighted or objectionable content, which models trained on them may reproduce. Learned aesthetic preference models reflect the preferences of the raters whose judgments they were fitted to, and a loss distilled from them inherits those preferences. As with any image generator, the models could be misused to produce misleading or harmful imagery; they are research-scale models trained only on public data.

REPRODUCIBILITY STATEMENT

Sec. 3.4 and App. B.3 give the generator architecture, data and training schedule; Apps. B.1 and B.2 describe how the aesthetic projector's reconstruction sets are built and how the projector is trained, with all loss weights and schedules; and App. B.7 gives pseudocode for training the projector and for pretraining with GESSO (Algs. 1 and 2). The evaluation protocol, including metrics, prompt sets, seeds and how deviations are computed, is in Sec. 4.1; guidance-scale selection and the verbatim VLM-judge prompts are in Apps. B.4 and B.6, the encoder comparison protocol in App. B.5, and the pairwise-preference protocol in App. B.6. Training-run variance across three independent pretrains is reported in App. A.1.4. All datasets and baseline recipes we build on are public (Changpinyo et al., 2021; Wang et al., 2026; Ma et al., 2026).

REFERENCES

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Tariq Berrada, Pietro Astolfi, Melissa Hall, Marton Havasi, Yohann Benchetrit, Adriana Romero-Soriano, Karteek Alahari, Michal Drozdzal, and Jakob Verbeek. Boosting latent diffusion with perceptual objectives. In *International Conference on Learning Representations (ICLR)*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Black Forest Labs. FLUX.1. `https://github.com/black-forest-labs/flux`, 2024.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. 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.

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.

Zhaorun Chen, Zichen Wen, Yichao Du, Yiyang Zhou, Chenhang Cui, Siwei Han, Jen Weng, Chaoqi Wang, Zhengwei Tong, Leria Huang, Canyu Chen, Haoqin Tu, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-Bench: Is your multimodal reward model really a good judge for text-to-image generation? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025d. arXiv:2407.04842.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-Reason-6M & PRISM-Bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.09680.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.09344.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

Abhijay Ghildyal, Nabajeet Barman, and Saman Zadtootaghaj. Foundation models boost low-level perceptual similarity metrics. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Pablo Hernández-Cámara, Jorge Vila-Tomás, Valero Laparra, and Jesús Malo. Dissecting the effectiveness of deep features as metric of perceptual image quality. *Neural Networks*, 185:107189, 2025.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. AesBench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024.

David R. Hunter. MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics*, 32 (1):384–406, 2004.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for GAN training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95(4):489–508, 2004.

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

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2511.13720.

Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, and Emad Barsoum. ReNeg: Learning negative embedding with reward guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhikai Li, Jiatong Li, Xuewen Liu, Wangbo Zhao, Pan Du, Kaicheng Zhou, Qingyi Gu, Yang You, Zhen Dong, and Kurt Keutzer. K-Sort Eval: Efficient preference evaluation for visual generation via corrected VLM-as-a-Judge. In *International Conference on Learning Representations (ICLR)*, 2026b.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J. Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shanchuan Lin and Xiao Yang. Diffusion model with perceptual loss. *arXiv preprint arXiv:2401.00110*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2508.03789.

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

OpenAI. Addendum to GPT-4o system card: Native image generation. `https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf`, 2025a.

OpenAI. Introducing our latest image generation model in the API. `https://openai.com/index/image-generation-api/`, 2025b.

OpenDatasets. DALL·E 3 dataset. `https://huggingface.co/datasets/OpenDatasets/dalle-3-dataset`, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Stephen E. Palmer, Karen B. Schloss, and Jonathan Sammartino. Visual aesthetics and human preference. *Annual Review of Psychology*, 64:77–107, 2013.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. DreamBench++: A human-aligned benchmark for personalized image generation. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (ICLR)*, 2024.

Qwen Team. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. In *International Conference on Machine Learning (ICML)*, 2023.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, 2024.

Christoph Schuhmann. LAION-Aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research (TMLR)*, 2026.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *International Conference on Learning Representations (ICLR)*, 2026.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? In *International Conference on Machine Learning (ICML)*, 2025.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018a.

Hossein Talebi and Peyman Milanfar. Learned perceptual image enhancement. In *IEEE International Conference on Computational Photography (ICCP)*, 2018b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. MiniT2I: A minimalist baseline for text-to-image generation. `https://peppaking8.github.io/#/post/minit2i`, 2026.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations (ICLR)*, 2024a.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning (ICML)*, 2024b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

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

Lihe Yang, Shang-Wen Li, Yang Li, Xinjie Lei, Dong Wang, Abdelrahman Mohamed, Saining Xie, Hengshuang Zhao, Kaiming He, and Hu Xu. In pursuit of pixel supervision for visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research (TMLR)*, 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. In *International Conference on Learning Representations (ICLR)*, 2025.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. GPT-4V(ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

Chenchen Zhu, Saksham Suri, Cijo Jose, Maxime Oquab, Marc Szafraniec, Wei Wen, Yunyang Xiong, Patrick Labatut, Piotr Bojanowski, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficient universal perception encoder. *arXiv preprint arXiv:2603.22387*, 2026.

A Additional Results

A.1 Quantitative Results

A.1.1 The Full Ladder at Every Scale

Table 8: **The full ladder at every scale, all metrics.** All four arms at B/32, L/32 and H/32, patch size 32, data and schedule fixed within each scale. †Measured after supervised finetuning.

		Human-preference scorers				Prompt align.		Fidelity
Scale	Arm	HPSv2.1↑	LAION↑	PS↑	IR↑	GenEval† ↑	DPGBench† ↑	FID↓
B/32	MiniT2I	19.36 ±0.07	5.131 ±0.007	19.40 ±0.01	−0.219 ±0.015	0.825 ±0.003	83.17 ±0.10	12.88 ±0.03
	PixelGen	20.00 ±0.02	5.241 ±0.008	19.61 ±0.01	−0.033 ±0.003	0.846 ±0.001	85.15 ±0.06	10.99 ±0.12
	PixelGen-SigLIP	21.12 ±0.05	5.407 ±0.008	19.89 ±0.01	0.173 ±0.009	**0.856** ±0.003	**85.54** ±0.09	10.25 ±0.05
	+ GESSO	**22.03** ±0.06	**5.615** ±0.012	**20.14** ±0.01	**0.325** ±0.019	0.852 ±0.002	85.35 ±0.16	**9.63** ±0.06
L/32	MiniT2I	20.40 ±0.02	5.318 ±0.013	19.63 ±0.01	−0.065 ±0.006	0.858 ±0.001	84.31 ±0.15	12.47 ±0.04
	PixelGen	21.81 ±0.05	5.560 ±0.001	20.04 ±0.02	0.247 ±0.011	**0.867** ±0.002	86.39 ±0.16	10.08 ±0.04
	PixelGen-SigLIP	22.32 ±0.02	5.603 ±0.011	20.24 ±0.01	0.346 ±0.003	**0.866** ±0.005	86.96 ±0.16	9.57 ±0.01
	+ GESSO	**23.15** ±0.01	**5.774** ±0.001	**20.45** ±0.01	**0.456** ±0.016	**0.865** ±0.005	**87.15** ±0.02	**9.29** ±0.04
H/32	MiniT2I	22.10 ±0.04	5.632 ±0.012	20.08 ±0.01	0.255 ±0.004	**0.879** ±0.007	86.13 ±0.08	10.50 ±0.03
	PixelGen	21.68 ±0.04	5.543 ±0.003	19.97 ±0.01	0.162 ±0.017	0.858 ±0.001	86.38 ±0.10	10.30 ±0.06
	PixelGen-SigLIP	23.17 ±0.01	5.742 ±0.010	20.44 ±0.01	0.453 ±0.008	0.880 ±0.002	**87.19** ±0.30	9.42 ±0.04
	+ GESSO	**23.84** ±0.01	**5.862** ±0.009	**20.63** ±0.01	**0.537** ±0.007	**0.884** ±0.001	**87.35** ±0.27	**9.06** ±0.10

Table 8 restores the columns Table 4 omits, and the full grid supports two readings.

First, the ranking of the ladder is stable across scale. The GESSO arm leads every human-preference column and FID at all three sizes, by margins many seed deviations wide, and at H/32 the GESSO arm leads or ties both prompt-alignment benchmarks as well: the largest model we train is the one where the full recipe wins or ties every column it is measured on. Second, the encoder swap keeps the recipe improving with scale: PixelGen-SigLIP stays above MiniT2I at every scale, while PixelGen, with DINOv2, falls below the MiniT2I baseline on every aesthetic column and on GenEval at H/32 after beating it at both smaller scales.

A.1.2 Pairwise Preference at All Scales

Fig. 4 repeats Fig. 3 on the pretrained B/32 and L/32 models. Over three judges, three scales and three baselines the comparison runs to 27 head-to-heads, and the GESSO arm is preferred in every one, with a 95% Wilson interval that excludes the null throughout. Two regularities hold. First, the margin against PixelGen-SigLIP, the comparison that isolates GESSO, is the narrowest of the three in all 9 judge–scale cells, while the margins against MiniT2I and against PixelGen swap order between scales: at H/32 the margin against PixelGen is the wider. Second, Qwen2.5-VL-72B, the judge sharing a model family with GESSO's teacher, is never the most favourable of the three in any of the 9 scale–baseline comparisons.

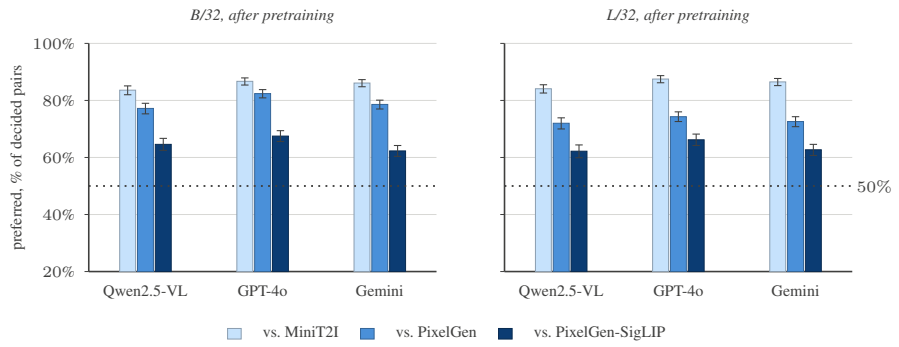

Figure 4: **Pairwise preference after pretraining at B/32 and L/32, three judges.** The counterpart of Fig. 3 at the two smaller scales: share of *decided* pairs in which the GESSO arm is preferred, with 95% Wilson intervals. As at H/32, the margin against PixelGen-SigLIP is the narrowest for every judge, and Qwen2.5-VL is never the most favourable judge.

Table 9: **PRISM-Bench per-track composites** for the models of Table 6 (Fang et al., 2026). Grouped by generation space (AR: autoregressive), ordered by average within each group.

			Overall			Per-track composite						
Space	Model	Res	Avg	Align	Aes	Imag.	Entity	Text	Style	Aff.	Comp.	Long
AR	GPT-Image-1 [High] (OpenAI, 2025b)	1024	80.7	82.7	78.7	66.6	84.1	76.8	87.5	84.0	88.5	77.4
	Janus-Pro-7B (Chen et al., 2025c)	384	62.1	64.9	59.4	51.9	66.0	36.7	72.9	74.3	76.7	56.4
Latent	Qwen-Image (Qwen Team, 2025)	1328	74.1	80.0	68.3	56.5	72.0	64.5	85.5	80.1	86.7	73.8
	FLUX.1-dev (Black Forest Labs, 2024)	1024	68.5	72.1	64.9	54.2	66.2	62.6	73.4	79.4	81.6	62.1
	SD3-Medium (Esser et al., 2024)	1024	66.1	68.0	64.2	51.0	66.3	50.9	77.0	78.0	82.5	56.8
	SDXL (Podell et al., 2024)	1024	57.0	60.1	54.0	44.3	68.0	27.9	72.1	72.6	70.0	44.3
	SD1.5 (Rombach et al., 2022)	512	46.0	48.8	43.3	32.2	56.9	17.8	59.1	63.8	55.4	37.0
Pixel	PixelGen-SigLIP + GESSO (H/32)	512	63.4	67.7	59.2	55.5	62.2	36.9	78.2	77.0	81.4	52.5
	MiniT2I-L/16 (Wang et al., 2026)	512	62.4	65.9	58.9	57.9	60.3	30.6	79.9	77.9	78.4	52.0
	PixelGen-SigLIP + GESSO (L/32)	512	61.8	66.1	57.5	58.7	60.8	29.6	78.8	75.1	77.8	52.1
	PixelGen-SigLIP + GESSO (B/32)	512	57.6	62.0	53.3	53.0	54.3	20.8	73.6	72.9	77.4	51.6
	MiniT2I-B/16 (Wang et al., 2026)	512	55.8	61.3	50.2	56.1	47.1	22.4	73.4	72.1	72.3	47.2

A.1.3 PRISM-BENCH: PER-TRACK SCORES

The published rows of Tables 6 and 9 are taken from the Qwen2.5-VL-72B-judged leaderboard of Fang et al. (2026). Our three arms are scored with the benchmark's released evaluation code under the same judge, and the two MiniT2I rows are taken from its release post (Wang et al., 2026), which reports the same track. GPT-Image-1 is listed as autoregressive following OpenAI's description of 4o image generation (OpenAI, 2025a), the model that GPT-Image-1 serves in the API (OpenAI, 2025b).

A.1.4 TRAINING-RUN VARIANCE AT B/32

Every scalar-metric mean in this paper is over generation seeds of one trained model per arm. This section varies the training run instead. The PixelGen-SigLIP B/32 arm was pretrained three times. A GESSO arm under the deployed recipe ($\lambda_3 = 0.04$, switched on at 100k steps) was branched from each of these three training runs (*trunks*), giving three matched pairs (Table 10).

The gain is invariant to the training run. GESSO adds 0.91, 1.09 and 0.95 HPSv2.1 on the three runs, and LAION-aesthetic, PickScore and ImageReward show the same pattern. GenEval moves by $+0.001$, -0.016 and $+0.018$, mixed in sign, so prompt alignment is unaffected, while FID falls by 0.62, 0.89 and 0.68. The absolute level, by contrast, belongs to the run: HPSv2.1 spans 0.24 (PixelGen-SigLIP) to 0.29 (the GESSO arm) across the three trunks, several times the generation-seed deviation, with the same ranking on every seed. The paper therefore reports the stable quantity, the gain.

Table 10: **Training-run variance at B/32.** Three independent PixelGen-SigLIP pretrains and a matched GESSO arm branched from each at 100k under the deployed recipe. PS and IR are PickScore and ImageReward. The *all* block is the mean ± training-run deviation across the three runs, and Δ there is taken over the three paired per-run changes rather than from the two means above it. ImageReward is a signed score near zero, so its relative change measures the denominator more than the gain. **Bold** marks a change that differs from zero at $p < 0.05$ under a two-tailed paired t-test over the three runs ($n = 3$).

		Aesthetics				Prompt align.		Fidelity
Run	Arm	HPSv2.1↑	LAION↑	PS↑	IR↑	GenEval↑	DPGBench↑	FID↓
1	PixelGen-SigLIP	21.12	5.407	19.89	0.173	0.607	82.33	10.25
	+ GESSO	22.03	5.615	20.14	0.325	0.608	82.72	9.63
2	PixelGen-SigLIP	21.03	5.405	19.90	0.173	0.625	83.11	10.27
	+ GESSO	22.12	5.640	20.13	0.326	0.609	82.90	9.38
3	PixelGen-SigLIP	20.88	5.398	19.86	0.138	0.588	82.62	10.29
	+ GESSO	21.83	5.586	20.07	0.280	0.606	83.06	9.61
all	PixelGen-SigLIP	21.01 ±0.12	5.403 ±0.005	19.88 ±0.02	0.161 ±0.020	0.607 ±0.019	82.69 ±0.39	10.27 ±0.02
	+ GESSO	21.99 ±0.15	5.614 ±0.027	20.11 ±0.04	0.310 ±0.026	0.608 ±0.001	82.90 ±0.17	9.54 ±0.14
	Δ	**+4.7%** ±0.5	**+3.9%** ±0.4	**+1.2%** ±0.1	**+93.1%** ±8.5	+0.3% ±2.9	+0.3% ±0.4	**−7.1%** ±1.4

A.1.5 ABLATIONS

Tables 11–13 vary the decisions that the construction of Secs. 3.2 and 3.3 leaves open, one at a time. Every row is a B/32 PixelGen-SigLIP arm that switches GESSO on from the same trunk and pretrains to 250k steps, read under the protocol of Sec. 4.1 with three generation seeds. The teacher and switch-on ablations are run at $\lambda_3 = 0.01$, a quarter of the deployed weight, so their gains understate the deployed effect; Table 13 sweeps λ_3 itself. Rows within a table share a parent, so differences within a table are directly comparable. The deployed teacher, switch-on step and weight were selected on the benchmarks of Sec. 4.1.

Teacher. The teacher only supplies the score gaps the projector is fitted to, so this ablation asks how much the choice of scorer matters once it is distilled. HPSv3, a prompt-conditioned VLM preference model, Q-Align, a prompt-free VLM rater, and MUSIQ, a small quality model trained without language, differ in architecture, size and training data, yet land within 0.14 HPSv2.1 (21.49 to 21.63), about one training-run deviation. On HPSv2.1 the teacher is therefore second-order: any scorer that ranks aesthetic quality supplies a usable target. The teacher matters more for fidelity. LAION-aesthetic spans 0.033, about one training-run deviation of the GESSO arm (0.027; Table 10), GenEval moves within one training-run deviation, and FID spans 0.47, from 9.73 with HPSv3 to 10.20 with Q-Align, several seed deviations apart. HPSv3 gives the best HPSv2.1 and the best FID, which is why it is deployed.

Table 11: **Ablation: the teacher supplying the gap.** Switch-on at 100k, $\lambda_3 = 0.01$. Green marks the configuration carried to the next ablation.

Teacher	HPSv2.1↑	LAION↑	GenEval↑	FID↓
HPSv3	**21.63** ±0.04	5.512 ±0.005	0.605 ±0.008	**9.73** ±0.04
Q-Align	21.58 ±0.04	**5.527** ±0.007	0.608 ±0.007	10.20 ±0.04
MUSIQ	21.49 ±0.09	5.494 ±0.012	**0.616** ±0.001	10.04 ±0.06
degradation-only	21.09 ±0.02	5.438 ±0.010	0.599 ±0.003	10.79 ±0.06
untrained proj.	21.30 ±0.05	5.443 ±0.006	0.607 ±0.003	10.16 ±0.06

Control: degradation-only. A control tests what degradation ordering alone provides. The *degradation-only* control keeps the backbone, the projector and the schedule, but trains the projector only to order the three reconstructions of an image by noise level. It sees no aesthetic label, so it measures what sensitivity to degradation alone contributes. It lands 0.54 HPSv2.1 and 0.074 LAION-aesthetic below the HPSv3 projector and gives up 1.06 FID, the worst FID in the table: ordering degradations alone does not supply the signal the loss needs. Both rows use $\lambda_3 = 0.01$.

Control: untrained projector. GESSO adds a second distance in SigLIP's feature space, so part of its gain could come from more SigLIP-space perceptual pressure rather than from what the projector learned. The *untrained-projector* control keeps the backbone, the distance and $\lambda_3 = 0.01$ but leaves the projector at its random initialization, adding that pressure at the same weight with nothing learned from the teacher. Against the shared parent (21.12 HPSv2.1, 5.407 LAION-aesthetic, 0.607 GenEval, 10.25 FID; Table 10) the untrained projector adds 0.18 HPSv2.1 and 0.036 LAION-aesthetic, leaves GenEval unchanged and lowers FID by 0.09; the HPSv3 projector adds 0.51 and 0.105 and lowers FID by 0.52. Extra SigLIP-space pressure thus accounts for about a third of the HPSv2.1 and LAION-aesthetic gains at this weight and a sixth of the FID gain. The rest of the gain comes substantially from the trained aesthetic projector.

Switch-on step. GESSO resumes from a PixelGen-SigLIP checkpoint rather than training from scratch, and this ablation asks whether the switch-on step matters. Switching on at 50k, 100k or 150k of the 250k steps leaves the arms within 0.15 HPSv2.1 and 0.028 LAION-aesthetic of one another, with FID tied across all three. Starting earlier, and so applying GESSO for more steps, does not help; starting later, with only 100k steps of GESSO, still recovers the gain, and gives the best GenEval of the three. The result is robust to this choice, and 100k is kept as the deployed switch-on step.

Table 12: **Ablation: when GESSO switches on.** HPSv3 teacher, $\lambda_3 = 0.01$.

Switch-on	HPSv2.1↑	LAION↑	GenEval↑	FID↓
50k	21.48 ±0.04	5.497 ±0.005	0.611 ±0.008	**9.75** ±0.03
100k	**21.63** ±0.04	5.512 ±0.005	0.605 ±0.008	**9.73** ±0.04
150k	21.52 ±0.05	**5.525** ±0.003	**0.631** ±0.007	**9.80** ±0.11

Weight λ_3. This ablation sets how strongly GESSO pulls against the other terms of Eq. 5. Both aesthetic scores rise monotonically with λ_3, HPSv2.1 from 21.37 at 0.005 to 22.17 at 0.1 and LAION-aesthetic from 5.459 to 5.673. The two guardrails behave differently: GenEval stays flat up to 0.04 (0.605–0.608) and FID improves up to 0.04 (9.87 to 9.63), but at 0.1 GenEval falls to 0.588 and FID rises to 10.00. Beyond 0.04 the extra aesthetic gain is bought with prompt alignment and fidelity, so among the four tested weights 0.04 is the largest at which GESSO improves aesthetics without a trade-off. It is the deployed weight, used in every experiment except the teacher and switch-on ablations and the per-patch arm of Table 2, which use 0.01.

Table 13: **Ablation: the weight λ_3.** HPSv3 teacher, switch-on at 100k. Green is the deployed weight.

λ_3	HPSv2.1↑	LAION↑	GenEval↑	FID↓
0.005	21.37 ±0.06	5.459 ±0.008	**0.606** ±0.004	9.87 ±0.08
0.01	21.63 ±0.04	5.512 ±0.005	**0.605** ±0.008	9.73 ±0.04
0.04	22.03 ±0.06	5.615 ±0.012	**0.608** ±0.011	**9.63** ±0.06
0.1	**22.17** ±0.03	**5.673** ±0.006	0.588 ±0.005	10.00 ±0.03

A.2 Qualitative Results

A.2.1 Qualitative Structure of the Learned Space

Fig. 5 projects four embedding spaces to two dimensions over the same images and colours each point by its teacher score. In the three existing spaces, high and low scores are largely mixed across the leading two components, with at most a weak gradient. In the aesthetic-projector panel, colour varies more clearly along one axis, from low scores at one edge to high scores at the other.

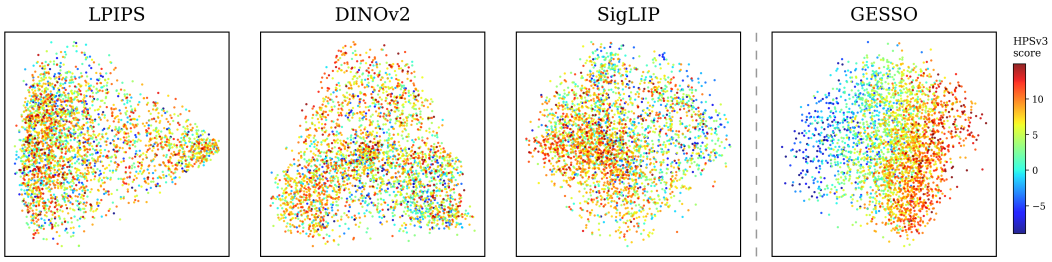

Figure 5: **Aesthetic signal is present in existing spaces, but diffuse.** Two-dimensional PCA of four embedding spaces over the same held-out images, coloured by HPSv3 score.

Each scatter belongs to an image that was held as the validation split that the projector was not trained on. The panel therefore indicates that the frozen SigLIP features can be reorganized so that aesthetic quality becomes a direction in the space rather than something diffused across it.

A.2.2 What Each Component Contributes

Fig. 6 carries two prompts through the four arms of the ladder at H/32 after supervised finetuning, with the initial noise held fixed, so each row differs only in the training objective and each column adds one change to the loss. The examples are selected for illustration; Tables 4 and 8 give the measurements, with aesthetics read after pretraining.

MiniT2I	+ LPIPS + pDINO	+ LPIPS + pSigLIP	+ GESSO
flow matching only	prior work's recipe	encoder swap	*ours*
no aesthetic term	+ perceptual losses	+ better target	+ aesthetic term

A ballerina dancing on an empty theater stage, spotlight, graceful motion, cinematic atmosphere.



A close-up of a delicate saffron flower with a vibrant purple hue and a yellow center, surrounded by green grass and soil.



Figure 6: **What each component contributes.** The four arms of the ladder (Table 4) at H/32 after supervised finetuning. Within a row all four arms decode the *same* initial noise.

B Implementation Details

B.1 Building the Aesthetic Projector's Training Set

The aesthetic projector is trained on *reconstruction sets* built from CC12M (Changpinyo et al., 2021) images, $\sim$427.5k for training and $\sim$22.5k held out. A reconstruction set is one reference image with three reconstructions of that reference. Each reconstruction is made the way the generator was trained: the reference is mixed with Gaussian noise, and the generator predicts the clean image back from that mixture in a single forward pass. The three members differ in how much noise went in, so the projector sees the artifacts these models produce when denoising, rather than blur or compression applied by hand. The reconstructions are drawn from all three arms of the ladder, MiniT2I, PixelGen and PixelGen-SigLIP, at each of the B/32, L/32 and H/32 scales, so the artifacts span recipes and model capacities. Each reference is reconstructed by one of these nine generators, chosen at random and read at its final 250k-step checkpoint, at $t = 0.1$, 0.4 and 0.7 ($t = 0$ is clean, so $t = 0.7$ is the most degraded member). Each reference therefore contributes exactly one reconstruction set, and the $\sim$450k total counts references and sets alike. The reference images come from the same CC12M corpus the generator pretrains on. The reported gains are therefore measured on benchmarks other than CC12M: HPS v2 prompts, PartiPrompts, GenEval, DPGBench and MSCOCO.

Every image is centre-cropped, loaded at 512px and bicubic-downsampled to 256px under the SigLIP normalization, and one horizontal flip is drawn per reconstruction set so that the members stay in correspondence. A frozen teacher, HPSv3 (Ma et al., 2025), then scores every image in the dataset; HPSv3 is prompt-conditioned, and each image is scored against its reference's CC12M recaption.

For each reconstruction the raw target is the aesthetic gap $g = |S(x_0) - S(\tilde{x})|$, mapped through the 5th and 95th percentiles of g on the training split alone and clipped to $[0, 1]$. The gap is the absolute score difference rather than a signed difference clamped at zero: for a small fraction of pairs the teacher scores the reconstruction above its reference, and the absolute difference treats deviations in either direction from the reference symmetrically.

B.2 Training the Aesthetic Projector

The projector is fitted to predict that gap through $\mathcal{L}$gap. The ranking term $\mathcal{L}$rank compares reconstructions of different images made by the same generator at the same noise level, and orders their distances by the teacher's gap with a logistic loss at temperature 0.25. Since generator and noise level are shared within each comparison, they cannot by themselves determine the order, although image content still varies. Pairs whose raw teacher gaps differ by at most 0.05 are skipped, so reconstructions the teacher scores alike are never forced apart. Table 14 is the deployed recipe, transcribed from the checkpoint's own stored configuration.

Table 14: **Hyperparameters to train the aesthetic projector.** The backbone is frozen; the projector and score head are trained. The corpus and the label being fitted are described in App. B.1.

Backbone and head	
Backbone	SigLIP ViT-B/16 at 256px, frozen, read at the final block after the output LayerNorm
Projector	$768 \to 2048 \to 2048 \to 768$ with SiLU, 7,344,896 trained parameters
Score head	$\mathrm{SiLU} \to \mathrm{Linear}(768 \to 1)$ on the mean-pooled embedding
Objective	
$\mathcal{L}_{\text{gap}}$	Huber, $\delta = 1$, neither side batch-standardized; weight 1
Gap normalization	5th–95th percentile of the training split, clipped to $[0, 1]$
$\mathcal{L}_{\text{rank}}$	logistic, temp. 0.25, per generator and noise level; ties ≤ 0.05 skipped; $\lambda_{\text{rank}} = 0.1$
$\mathcal{L}_{\text{anc}}$	Gram MSE to the detached backbone, ≤ 256 tokens; $\lambda_{\text{anc}} = 5$
$\mathcal{L}_{\text{score}}$	MSE to the normalized teacher score; $\lambda_{\text{score}} = 0.01$
Optimization	
Optimizer	AdamW, weight decay 0.1, gradient clipping 1.0
Schedule	cosine, 30 epochs, 2 warmup
Batch	$8 \times 128 = 1024$ reconstruction sets (4096 images) per step
Learning rate	$2.5 \times 10^{-4} \times (\text{effective batch})/256 = 10^{-3}$
Model selection	checkpoint with the best validation gap correlation

The degradation-only control. The control of Table 11 keeps the backbone, the projector and the schedule of the deployed recipe, with a rank term over the members of a reconstruction set, a hinge with margin 0.1 that asks only for the distance to grow as the reconstruction is made from more noise. The hinge orders the noise levels only so no aesthetic label reaches the objective at any point.

B.3 Generator Training

The scalar arms of Table 2. Every arm of Table 2 adds a MUSIQ term to the same PixelGen-SigLIP base, using the MUSIQ checkpoint trained on the AVA aesthetic dataset. The two scalar arms read the frozen MUSIQ model S_{M} itself, backpropagating through MUSIQ: the scalar-reward arm maximizes the score of the prediction and the scalar-gap arm minimizes $\big(S_{\mathrm{M}}(f_\theta(x_t)) - S_{\mathrm{M}}(x_0)\big)^2$, with the score normalized by the AVA scale so that both objectives are read on one range. The per-patch arm instead reads the distillation of Sec. 3.2, a projector over frozen SigLIP fitted with MUSIQ as its teacher. The scorer is therefore held fixed across the three arms. The arms differ in how its signal enters the loss, and the per-patch arm also differentiates through a different network: a trained projector over SigLIP rather than MUSIQ itself. The untrained-projector control of App. A.1.5 uses that same network with nothing learned from a scorer, which separates MUSIQ's contribution from the network's (Sec. 4.3). All three branch from the same PixelGen-SigLIP checkpoint at 100k.

The weights of the two scalar terms are set so that the two terms exert comparable force, rather than to a common nominal value. Measured against the predicted pixels, the reward term at $\lambda_3 = 0.01$ contributes gradient of 2.5×10^{-7} and 1.99% of the total loss, against 1.9×10^{-7} for the $\mathcal{L}_{\text{pSigLIP}}$ term beside it and 7.6×10^{-7} for LPIPS. The gap term at that same nominal weight contributes 7.3×10^{-9} and 0.011%, so the gap term is inert at that weight and is carried at $\lambda_3 = 1.0$ instead: $\big(S_{\mathrm{M}}(f_\theta(x_t)) - S_{\mathrm{M}}(x_0)\big)^2$ shrinks as the prediction approaches its reference, where $1 - S_{\mathrm{M}}(f_\theta(x_t))$ does not. Both scalar arms are therefore read at a weight at which the term measurably acts on the model. Every arm pretrains for 250k steps under the public MiniT2I recipe with only the loss changed, and the finetuned arms continue for 40k further steps (to 290k) on the shared 119k-image synthetic set of Sec. 4.5, each keeping its own perceptual and aesthetic terms. In the deployed arms GESSO is switched on at 100k steps by resuming from the PixelGen-SigLIP checkpoint. All samples use 100 Euler steps during inference.

Table 15 lists every generator setting; Table 16 lists what distinguishes the arms. Both are transcribed from the run configurations, and the only settings that change between arms are the loss terms. The projector's own settings are in App. B.2.

Table 15: **Generator training.** The three scales differ only in the generator's dimensions; every other setting is shared, and every arm of every table in the paper uses them unchanged. Loss weights are in Table 16.

	B/32	L/32	H/32
Generator			
Parameters (generator)	260M	915M	1.88B
Hidden size	768	1248	1536
Double-stream blocks	17	23	32
Attention heads	12	24	24
Head dimension	64	52	64
MLP ratio	2.667	2.705	2.667
Text-preamble blocks		2	
Patch size		32 (at 512px, so 16×16 tokens)	
Patch-embed bottleneck		128 channels	
Diffusion			
Parameterization		x-prediction, flow-matching loss in velocity space	
t sampling		logit-normal, $\mu = 0.8$, $\sigma = 0.8$	
Forward noise scale		2 (PixelGen uses 1)	
Caption drop rate		0.1	
Optimization			
Optimizer		AdamW, $\beta_2 = 0.95$, weight decay 0	
Learning rate		4×10^{-4}, 5k warmup steps	
Batch size		1024 (128 per device)	
EMA decay		0.99995	
Precision		bf16	
Pretraining steps		250k	
Finetuning steps		40k further (290k total)	
Text and data			
Text encoder		FLAN-T5-Large, frozen; 256 prompt tokens	
Pretraining corpus		CC12M, LLaVA-NeXT recaptions	
Finetuning corpus		BLIP3-o-60k, DALL·E 3, ShareGPT-4o	

Table 16: **Loss terms by arm.** PixelGen adds LPIPS and pDINO to MiniT2I; PixelGen-SigLIP replaces pDINO with pSigLIP; the GESSO arm adds its aesthetic term; "—" means the term is absent. Weights are those of Eqs. 2 and 5, held fixed at all three scales.

	MiniT2I	PixelGen	PixelGen-SigLIP	+ GESSO
$\mathcal{L}_{\text{FM}}$	✓	✓	✓	✓
λ_1, LPIPS (VGG, 256px)	—	0.1	0.1	0.1
λ_2, pDINO (DINOv2 ViT-B/14, block 11, 224px)	—	0.01	—	—
λ_2, pSigLIP (SigLIP ViT-B/16, 256px)	—	—	0.01	0.01
λ_3, $\mathcal{L}_{\text{GESSO}}$ (SigLIP ViT-B/16, 256px)	—	—	—	0.04
Training step when GESSO is switched on	—	—	—	100k
Noise gate threshold for perceptual losses	—		$t \leq 0.7$	

B.4 Choosing the Guidance Scale

Aesthetic scores rise with classifier-free guidance, so the aesthetic and preference numbers depend on where on that curve the arms are read. This section gives the tournament that fixed one scale for those numbers and the separate sweep that reads prompt alignment and fidelity.

Selecting the scale. Each arm is ranked *independently* over a grid of six guidance values, so no arm is placed at a scale chosen by another. For one arm, every pair of scales is compared on 150 prompts in a complete round-robin, 15 pairs by 150 prompts, giving 2250 judged pairs. Qwen2.5-VL-72B makes a forced choice with no tie option at temperature 0, every pair is asked in both image orders, and a pair is a win only when the two orders agree. A pair whose two orders disagree, 44–48% of pairs depending on the arm, is recorded as a tie and credited half a win to each scale, so position bias cancels rather than favouring either scale. In the prompt below {`caption`} is filled with the prompt

the pair was generated from. Ratings come from a Bradley–Terry fit (Bradley & Terry, 1952) by the MM algorithm (Hunter, 2004), reported on the Elo scale and mean-anchored to 1000.

The criterion has to name the cost of guidance. Selecting the scale needs a different criterion from the ranked one used to compare arms. The criterion we use names the failure modes that guidance causes (e.g. oversaturated colour, crushed shadows, blown highlights, over-sharpened edges, waxy skin and an artificial appearance) and asks which image the judge would rather ship.

```
Caption: "{caption}"

You are shown two images generated from that caption. These two images come from the
SAME model and the SAME caption, differing only in classifier-free guidance strength.

Raising guidance makes an image follow the caption more closely, but past a point it also
causes over-saturated colour, crushed shadows, blown highlights, over-sharpened edges,
waxy skin, and a generally artificial "AI look".

Pick the image you would rather ship: the one that follows the caption well WITHOUT those
over-guidance artefacts. Do not simply pick whichever follows the caption more closely if
it looks over-processed; equally, do not pick a clean image that has lost the caption.

Respond with a single JSON object on one line and nothing else:
{"criterion": "<adherence|artefacts>", "reason": "<one short sentence>", "winner": <1 or 2>}
```

Table 17: **Guidance tournament, H/32 after supervised finetuning.** Bradley–Terry ratings on the Elo scale, mean-anchored to 1000, from a complete round-robin over six guidance values judged by Qwen2.5-VL-72B under the *guidance* criterion. Each arm is ranked independently, so the column an arm peaks in is chosen by that arm alone. **Bold** is each arm's own optimum; ‡PixelGen-SigLIP ties 2.5 with 3.0 to 13 decimal places.

	Guidance scale						
Arm	1.5	2.5	3.0	5.0	7.0	9.0	Optimum
MiniT2I	994	1033	**1061**	1017	962	933	3.0
PixelGen	974	1017	**1034**	1025	1002	947	3.0
PixelGen-SigLIP	1016	**1037**	**1037**	1027	972	910	$2.5 = 3.0^{\ddagger}$
+ GESSO	976	1039	**1057**	1019	985	924	3.0

What the tournament returns. Table 17 gives the four H/32 arms after supervised finetuning. All four curves are inverted U shapes and all four rank $w = 9$ worst, so the grid edge is rejected by the models rather than by us. Three of the four select $w = 3$ outright and the fourth, PixelGen-SigLIP, ties 2.5 with 3.0. Since the baselines select or tie at $w = 3$ independently of our arm, the operating point is not set by our method. We apply it unchanged to every arm, scale and stage.

Prompt alignment and fidelity are swept, not fixed. The common scale governs the aesthetic scorers and the VLM preference judgments only. GenEval, DPGBench and FID are each read at the arm's own guidance optimum, found by a sweep run per arm and per metric.

B.5 Choosing the Perceptual Encoder

The ten encoders, grouped by pretraining objective as in Table 1, are: *self-distillation*, DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2026); *language supervision*, PE (Bolya et al., 2025), CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023) and SigLIP2 (Tschannen et al., 2025); *reconstruction*, Pixio (Yang et al., 2026) and MAE (He et al., 2022); *latent prediction*, LeJEPA (Balestriero & LeCun, 2025); and *multi-teacher distillation*, EUPE (Zhu et al., 2026).

The comparison holds the recipe fixed. All encoders are ViT-B, so encoder identity is not confounded with parameter count, and base is the largest size the comparison can afford. All spatial patches are read from the last layer, and all are given a matched patch-token budget despite differing native resolutions and patch sizes, so no encoder receives more signal than another; we match tokens rather than input resolution because the perceptual loss is computed per token. Each encoder runs at the input resolution that gives a 14×14 patch grid: 224px for patch-16 encoders and 196px for DINOv2, whose patch size is 14.

Shared loss settings. Two settings are inherited unchanged from the recipe of Ma et al. (2026), which tuned them for DINOv2: the perceptual weight $\lambda_2 = 0.01$ and the last-layer readout. All ten encoders share this configuration, as in the encoder comparisons of Yu et al. (2025) and Singh et al. (2026), which hold the loss weight fixed across encoders.

B.6 Pairwise Preference: Protocol

With the scale fixed, arms are compared against one another on the same prompts, and that comparison uses a ranked criterion, given in the prompt below. Both orders of every pair are shown, and a pair counts only when the two orders agree. Unlike the guidance tournament, undecided pairs are excluded, and win rates are over decided pairs. All judgments use one generation seed.

```
Caption: "{caption}"

You are shown two images generated from that caption. Judge them in this priority order:

1. CAPTION ADHERENCE - does the image actually contain what the caption describes?
2. REALISM - plausible detail and natural lighting, without over-saturation,
   over-sharpening, or an artificial "AI look".
3. AESTHETIC QUALITY - composition, colour, lighting, overall visual appeal.


Respond with a single JSON object on one line and nothing else:
{"criterion": "<adherence|aesthetic|realism>", "reason": "<one short sentence>", "winner":
↪  <1 or 2>}
```

B.7 Pseudocode

The two stages are given separately because they are run separately and only the first produces a learned component. Algorithm 1 fits the projector: a frozen backbone reads each reconstruction set, teacher scores computed once offline give the gap, the per-patch distance is regressed onto the normalized aesthetic gap, and the gradient reaches the projector and its score head alone. Algorithm 2 then holds that projector fixed and shows where its distance enters generator training, as one term on the predicted clean image beside flow matching and the two perceptual losses. Read together they give the whole method: everything learned happens in the first, and the second adds no trainable parameters and nothing at inference.

Algorithm 1 Training the aesthetic projector

Require: reconstruction sets $\{(x_0, \tilde{x}_1, \tilde{x}_2, \tilde{x}_3)\}$, degradation increasing with the index; teacher scores $S(\cdot)$ of every image, precomputed once offline; frozen SigLIP ViT-B/16 features $\phi(\cdot)$ (256px, final block); projector h_ψ ($768 \to 2048 \to 2048 \to 768$) and score head r; $\lambda_{\text{rank}} = 0.1$, $\lambda_{\text{score}} = 0.01$, $\lambda_{\text{anc}} = 5$

lo, hi $\leftarrow$ 5th, 95th percentile of $\{|S(x_0) - S(\tilde{x}_\ell)|\}$ over the *training* split ▷ stored
for epoch $= 1 \dots 30$ **do**
  for batch of reconstruction sets **do**
    $x \leftarrow$ centre-crop, 512px $\to$ 256px, random horizontal flip shared across the set
    $\phi(x) \leftarrow \text{LN}_{\text{post}}(\text{backbone}(x))$ ▷ no grad; backbone frozen
    $z \leftarrow h_\psi(\phi(x)), \quad \hat{z} \leftarrow z/\|z\|_2$ ▷ per patch
    for $\ell \in \{1, 2, 3\}$ **do**
      $d_\ell \leftarrow d_{\text{aes}}(x_0, \tilde{x}_\ell) = \frac{1}{|\mathcal{P}|} \sum_p \big(1 - \langle \hat{z}^p(x_0), \hat{z}^p(\tilde{x}_\ell) \rangle\big)$
      $g_\ell \leftarrow |S(x_0) - S(\tilde{x}_\ell)|$ ▷ raw gap, from the stored scores
      $\bar{g}_\ell \leftarrow \text{clip}\big((g_\ell - \text{lo})/(\text{hi} - \text{lo}),\, 0,\, 1\big)$
    end for
    $\mathcal{L}_{\text{gap}} \leftarrow \text{mean}_\ell \, \text{Huber}_{\delta=1}(d_\ell, \bar{g}_\ell)$
    $\mathcal{L}_{\text{rank}} \leftarrow \text{mean}_{(\text{generator},\, \ell)} \, \text{mean}_{i,j:\, g_\ell^{(i)} > g_\ell^{(j)} + 0.05} \, \text{softplus}\big(-(d_\ell^{(i)} - d_\ell^{(j)})/0.25\big)$ ▷ images i, j
    $\mathcal{L}_{\text{anc}} \leftarrow \text{MSE}\big(\hat{z}\hat{z}^\top,\, \hat{\phi}\hat{\phi}^\top\big)$ ▷ $\hat{\phi}$: ϕ with ℓ_2-normalized rows, detached; ≤ 256 tokens
    $\mathcal{L}_{\text{score}} \leftarrow \text{MSE}\big(r(\text{mean}_p\, z),\, \bar{S}\big)$ ▷ $\bar{S} \in [0, 1]$; head discarded later
    $\mathcal{L}_{\text{proj}} \leftarrow \mathcal{L}_{\text{gap}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}} + \lambda_{\text{score}} \mathcal{L}_{\text{score}} + \lambda_{\text{anc}} \mathcal{L}_{\text{anc}}$ ▷ Eq. 4
    AdamW step on ψ and r only ▷ cosine, wd 0.1, clip 1.0
  end for
  validate: $\rho \leftarrow \text{mean}_\ell \, \text{Spearman}(d_\ell, \bar{g}_\ell)$; keep the best
end for

Algorithm 2 Pretraining with GESSO

Require: generator f_θ; frozen ϕ, h_ψ from Alg. 1; $\lambda_1 = 0.1$, $\lambda_2 = 0.01$, $\lambda_3 = 0.04$; gate $t_{\max} = 0.7$; switch-on step $K = 100$k

for step $= 1 \dots 250$k **do**
  $(x_0, c) \sim$ data; drop c with the caption-dropout probability ▷ for classifier-free guidance
  $t \sim$ schedule, $\varepsilon \sim \mathcal{N}(0, I)$; $\quad x_t \leftarrow (1 - t)\, x_0 + t\, s\varepsilon$, $s = 2$
  $\hat{x}_0 \leftarrow f_\theta(x_t, c)$ ▷ $f_\theta(x_t)$ of Eq. 1, caption shown
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{FM}}(\hat{x}_0)$ ▷ Eq. 1
  $m_t \leftarrow \mathbf{1}[t \le t_{\max}]$, renormalized to $|B|/\sum_B m_t$ over the batch B ▷ gates all perceptual terms
  $\mathcal{L} \leftarrow \mathcal{L} + \lambda_1\, m_t\, \mathcal{L}_{\text{LPIPS}}(\text{resize}(\hat{x}_0, 256), \text{resize}(x_0, 256))$
  $\phi_{\text{pred}} \leftarrow \phi(\text{resize}(\hat{x}_0, 256))$
  $\phi_{\text{tgt}} \leftarrow \phi(\text{resize}(x_0, 256))$ ▷ no grad
  $\mathcal{L} \leftarrow \mathcal{L} + \lambda_2\, m_t\, \mathcal{L}_{\text{pSigLIP}}(\phi_{\text{pred}}, \phi_{\text{tgt}})$ ▷ same forward pass reused below
  if step $> K$ **then**
    $z_\bullet \leftarrow h_\psi(\phi_\bullet)$, $\hat{z}_\bullet \leftarrow z_\bullet/\|z_\bullet\|_2$ per patch, for $\bullet \in \{\text{pred}, \text{tgt}\}$
    $\mathcal{L}_{\text{GESSO}} \leftarrow d_{\text{aes}}(\hat{x}_0, x_0) = \frac{1}{|\mathcal{P}|} \sum_p \big(1 - \langle \hat{z}^p_{\text{pred}}, \hat{z}^p_{\text{tgt}} \rangle\big)$ ▷ Eq. 3
    $m_c \leftarrow \mathbf{1}[c \text{ retained}]$, renormalized
    $\mathcal{L} \leftarrow \mathcal{L} + \lambda_3\, m_t\, m_c\, \mathcal{L}_{\text{GESSO}}$
  end if
  optimizer step on θ
end for

B.8 Samples Across Model Scale

Figs. 7–9 show the same twenty HPS v2 benchmark prompts generated by the GESSO arm at each of the three scales of Table 4, with prompt, seed and sampler held fixed across the three grids so that a cell can be read down the sequence.

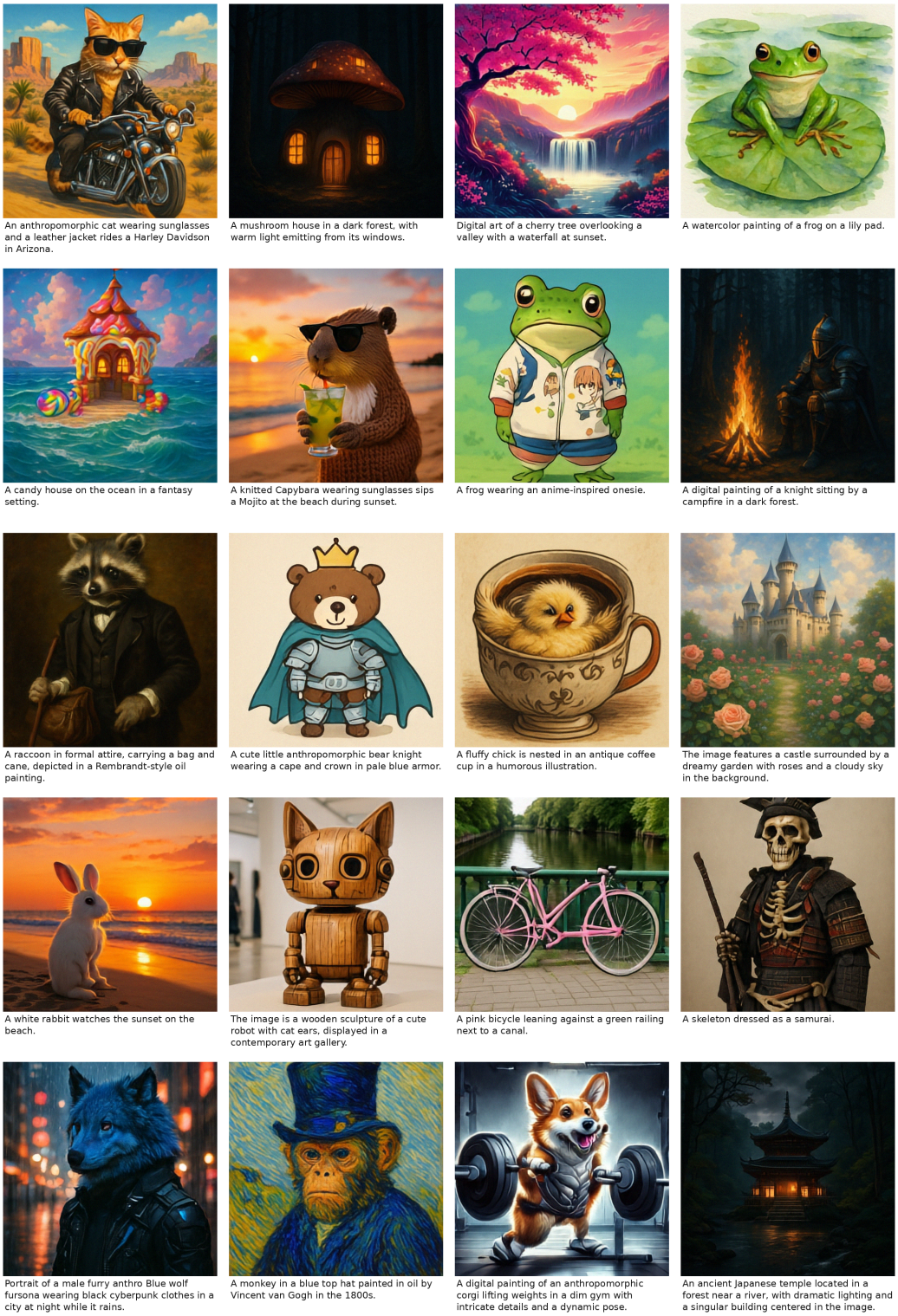

Figure 7: **B/32.** Twenty HPS v2 benchmark prompts, GESSO arm, smallest scale.

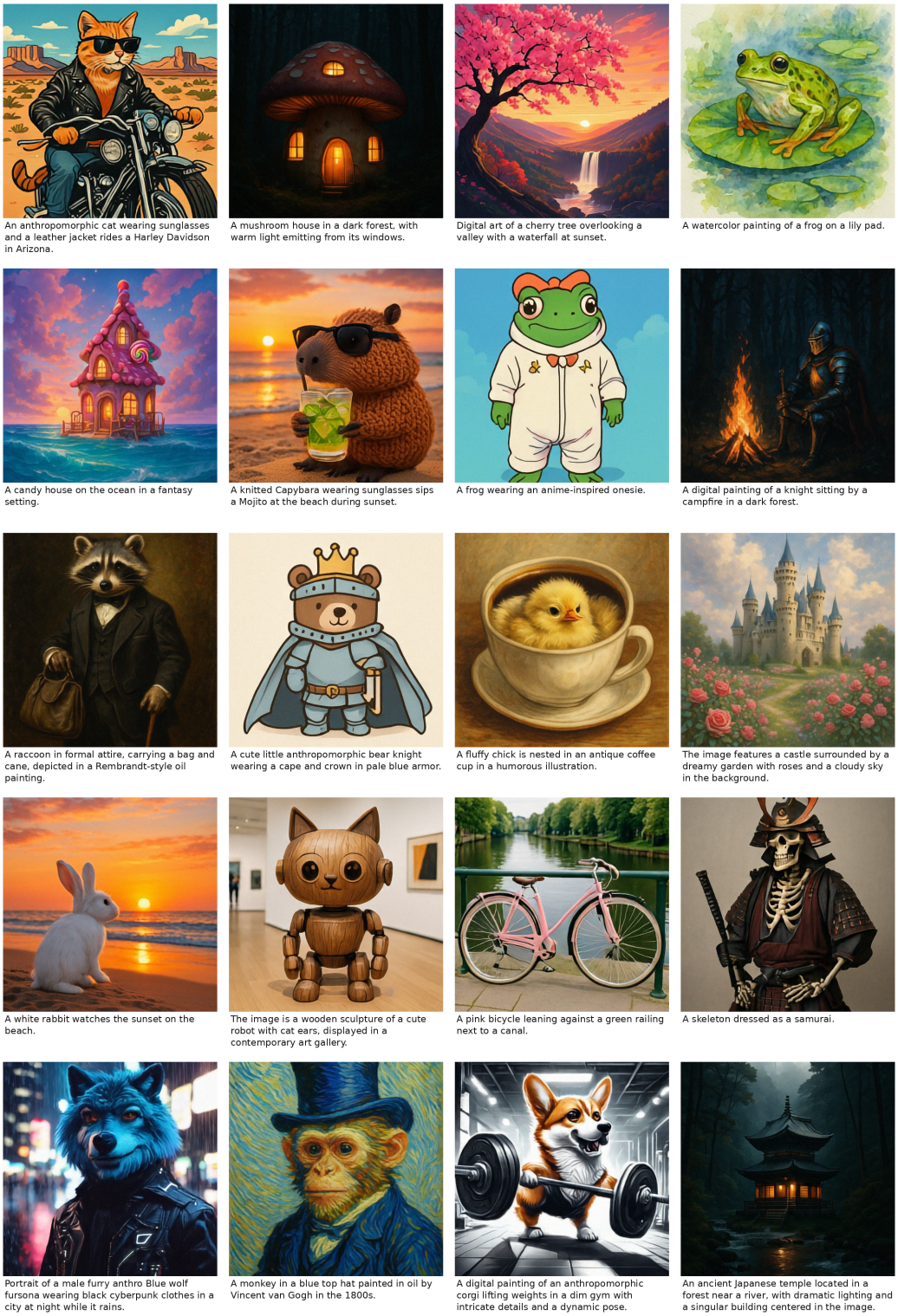

Figure 8: **L/32.** The same twenty prompts and seeds as Fig. 7.

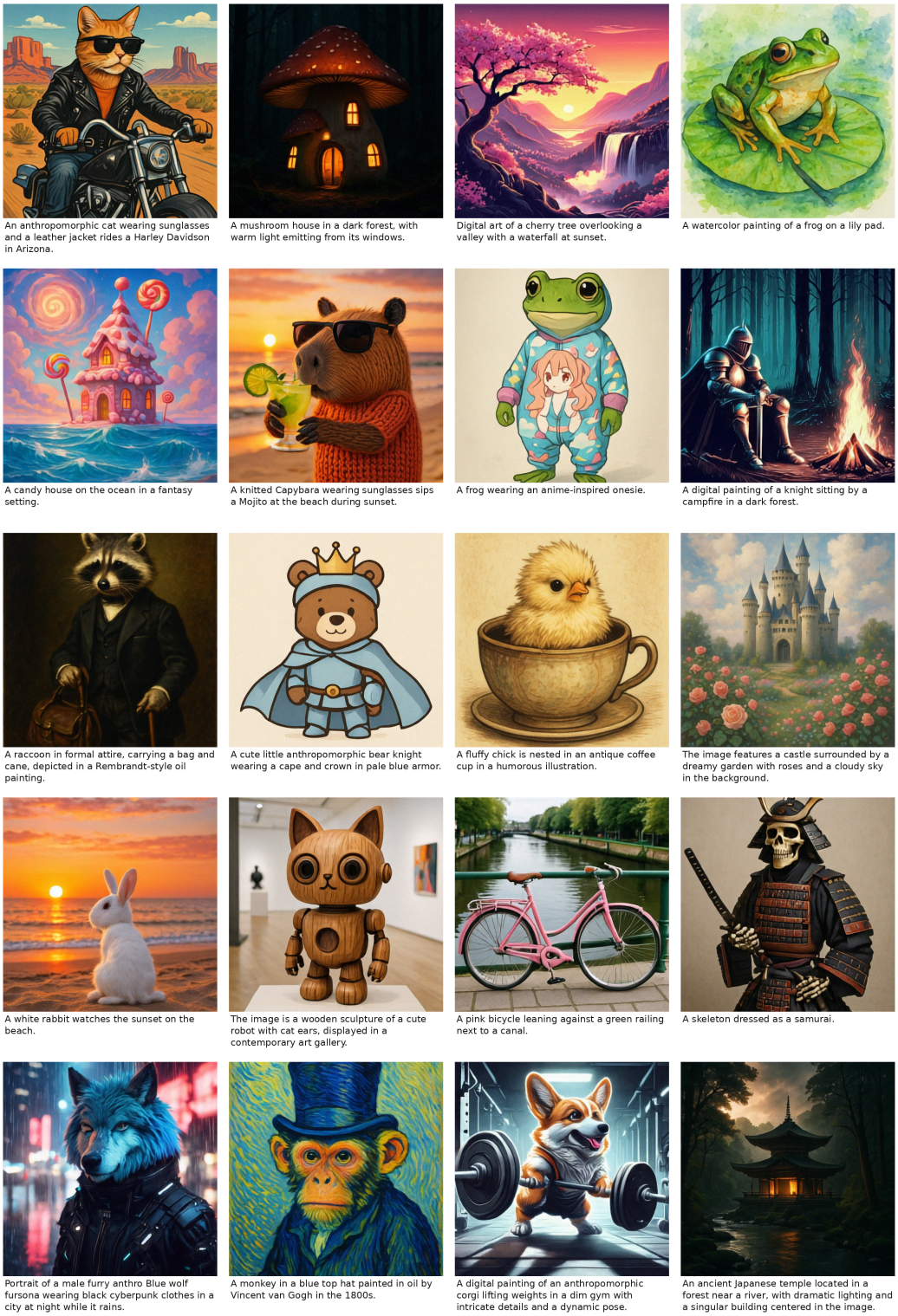

Figure 9: **H/32.** The same twenty prompts and seeds as Fig. 7.