SCALING TRANSFORMER ENERGY-BASED MODELS FOR HIGH QUALITY IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy-Based Models (EBMs) are a flexible class of generative models that decouple probabilistic modeling from the generation process. This decoupling directly enables compositional generation and inference-time scaling, where sample quality can be refined simply by increasing sampling steps. However, severe challenges in high-dimensional sampling and training instabilities with deep architectures have restricted EBMs to low-dimensional regimes. To resolve these bottlenecks, we introduce the Transformer Energy-Based Model (TEM), a scalable transformer energy function operating entirely within a latent space. On top of this backbone, we extract energies at multiple layers and sum them, composing complementary constraints from local texture to global semantics. Furthermore, two components make this high-quality at scale: Positive Representation Alignment (PRA), which supervises TEM's intermediate representations with a vision foundation model, and advanced sampling strategies that improve sample quality at inference. TEM-XL achieves an FID-50K of 3.82 on class-conditional ImageNet 256×256 without classifier-free guidance, showing that a contrastive EBM can generate high-fidelity images. On ImageNet 512×512, TEM achieves an FID-50K of 8.24, demonstrating that the approach scales to higher resolution. We further validate the core promises of EBMs: compositional generation, zero-shot inpainting, and out-of-distribution detection.

1 INTRODUCTION

Energy-Based Models (EBMs) (LeCun et al., 2006; Du & Mordatch, 2019) parameterize probability distributions via an energy function that assigns lower energies to more probable samples and higher energies to less probable samples. EBMs provide a flexible and mathematically elegant framework that fundamentally separates explicit generation from probabilistic modeling. Unlike Diffusion (Ho et al., 2020; Sohl-Dickstein et al., 2015) and Flow (Lipman et al., 2022) models, which are mathematically constrained to pre-defined generative trajectories governed by carefully designed timestep schedules, EBMs define the data distribution solely through an energy function, entirely decoupling the model from the generative process. This decoupling of the model from the generative process yields profound advantages. First, EBMs decouple learning from inference, allowing the sampling procedure to be chosen independently of model training. This enables flexible test-time compute scaling by increasing the MCMC steps, progressively refining the samples through a deeper exploration of the learned energy landscape. Second, this flexibility enables the composition of different distributions together (Du & Mordatch, 2019; Du et al., 2023), which can enable zero-shot generation and generalization capabilities.

However, despite their theoretical advantages and compositional properties, EBMs have struggled with widespread adoption due to several critical unresolved challenges. First, EBMs face severe vulnerability to the curse of dimensionality: MCMC sampling suffers from poor mixing in high-dimensional spaces, leading to slow convergence and mode collapse. Second, EBMs scale poorly with model capacity. As network depth increases, conventional architectures exhibit severe training instability, making robust optimization exceedingly difficult (Du & Mordatch, 2019; Du et al., 2020). Consequently, existing EBM research has been largely limited to smaller-scale benchmarks, unable to demonstrate competitive performance on challenging, high-resolution datasets.

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



Figure 1: **High-fidelity ImageNet 256×256, 512×512 generation and compositional synthesis by TEM-XL/2.** Left: uncurated class-conditional samples from 80 Langevin steps. The two large images are 512×512 samples and the remaining tiles are 256×256. Right: compositional generation by summing the energy functions of two independent classes, producing a coherent scene that satisfies both concepts without retraining.

In this work, we address these fundamental limitations and demonstrate that EBMs trained contrastively, with negatives drawn from the model's own MCMC sampler, can generate high-quality 256×256 ImageNet (Deng et al., 2009) samples through persistent sampling initialized solely from Gaussian noise, achieving performance competitive with strong generative models (Table 1). We further validate the core theoretical promises of EBMs by demonstrating compositional generalization as well as flexible inference-time control. Our work addresses the long-standing scalability bottleneck of contrastive EBMs, reestablishing them as a viable contender for high-fidelity image generation.

Our key contributions are: **(1)** A scalable Vision Transformer-based EBM architecture with multiple energy heads that enables stable training at high resolution. **(2)** A comprehensive training framework encompassing latent space design, representation alignment, and advanced sampling strategies that together achieve state-of-the-art quality images for contrastive energy-based models. **(3)** Empirical validation of the theoretical advantages of energy-based modeling at high resolution: compositional generation, zero-shot inpainting, and out-of-distribution detection from a single scalar energy.

2 Background

2.1 Energy-Based Models

Energy-Based Models (EBMs) define a probability distribution implicitly over an input domain using a scalar energy function, expressed as $p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z(\theta)}$, where $E_\theta(\mathbf{x})$ is the parameterized energy function. In practice, $E_\theta(\mathbf{x})$ is parameterized by a neural network $E_\theta : \mathcal{X} \to \mathbb{R}$, yielding an unnormalized density model. Maximum likelihood training then shapes the energy landscape such that low-energy regions correspond to high-density regions of the data distribution $p_{\text{data}}(\mathbf{x})$, ensuring that $p_\theta(\mathbf{x})$ approximates $p_{\text{data}}(\mathbf{x})$. However, $Z(\theta) = \int_{\mathbf{x}} \exp(-E_\theta(\mathbf{x}))d\mathbf{x}$ is intractable, which makes exact maximum likelihood training of EBMs computationally infeasible. On the other hand, the gradient of log-likelihood remains tractable (Du & Mordatch, 2019; Hinton et al., 2006):

$$\nabla_\theta \mathbb{E}_{x\sim p_{\text{data}}(x)}\big[-\log p_\theta(x)\big] = \mathbb{E}_{x\sim p_{\text{data}}(x)}\big[\nabla_\theta E_\theta(x)\big] - \mathbb{E}_{x\sim p_\theta(x)}\big[\nabla_\theta E_\theta(x)\big] \tag{1}$$

The first term is estimated directly from samples from the data. The second requires expectations under $p_\theta(x)$, which is intractable to sample exactly. Running Markov Chain Monte Carlo (MCMC) to convergence is required, implying infinitely long chains with infinitesimal steps.

2.2 Short-Run Langevin MCMC

We therefore rely on short-run MCMC to obtain approximate samples. In particular, we start from a fixed initial distribution $p_0(\mathbf{z})$ and perform a finite number of T Langevin updates with step size η:

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \eta \nabla_{\mathbf{z}} \log \pi(\mathbf{z}_t) + \sqrt{2\eta}\,\boldsymbol{\epsilon}, \quad t = 0, \ldots, T-1, \tag{2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ (Du & Mordatch, 2019; Nijkamp et al., 2019). Let $\tilde{\pi}(\mathbf{z})$ denote the distribution of the final state $\mathbf{z}_T$. Since initialization p_0, step size η, and number of steps T are fixed, $\tilde{\pi}$ is

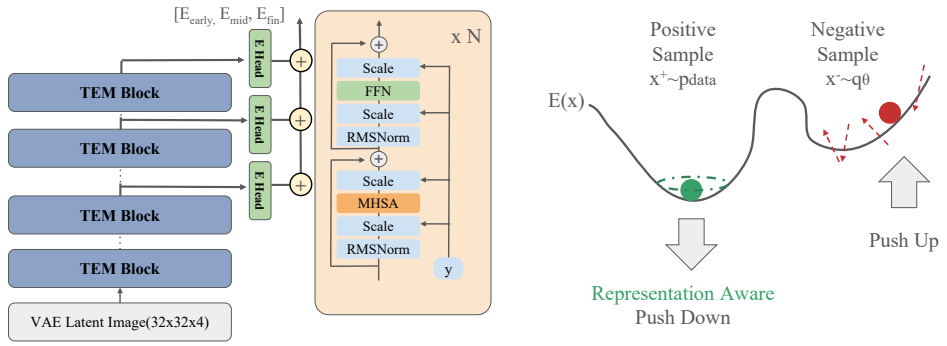


Figure 2: **Proposed Architecture and Training Strategy.** *Left*: the Transformer Energy-Based Model processes VAE-encoded latent tokens through a stack of TEM blocks. Energy heads at multiple intermediate layers produce scalar energies whose sum defines the total energy. *Right*: contrastive training pushes down the energy of data samples x^+ and pushes up the energy of MCMC negatives x^-, while Positive Representation Alignment (PRA) aligns the model's internal features with a frozen VFM.

well-defined. In practice, we use a finite number of steps and tune the step size and noise scale independently, yielding an efficient but biased sampler.

3 Scalable Transformer Energy-Based Models

3.1 Transformer Energy-Based Models Architecture

TEM Architecture. We parameterize the energy function $E_\theta(x, c)$ with a Vision Transformer (ViT) (Dosovitskiy et al., 2020) backbone adapted for large-scale energy-based modeling. Given an input x, we tokenize it into spatial patch embeddings, which are processed through a stack of transformer blocks. The final patch tokens are aggregated via global average pooling and passed through a small MLP head to produce a scalar energy $E_\theta(x, c)$.

Our architecture adopts several modern design choices for stability and expressivity. We encode spatial structure with rotary positional embeddings (Su et al., 2024), and apply query-key normalization (Henry et al., 2020) to stabilize attention logits at scale. Feed-forward layers use SwiGLU activations (Shazeer, 2020), and the conditioning signal is projected into feature-wise affine parameters in an AdaLN-style formulation (Peebles & Xie, 2023) built on the same RMSNorm variant.

Importantly, this design demonstrates stable and effective energy-based modeling at scale without requiring EBM specific components such as spectral normalization commonly used in prior energy-based models (Du & Mordatch, 2019).

Multi-layer Energy Extraction. Energy-based models are naturally suited to composition, since combining distributions amounts to addition in energy space. Du et al. (2020) demonstrate this by composing several separately trained models, each capturing a distinct concept. Composition of this kind has so far required multiple models.

We show that it can instead be obtained within one, extracting energies through multiple heads attached to representations at different depths of a single Transformer. Its layers encode progressively more abstract features, from low-level structure in early layers to high-level semantics in deeper ones (Raghu et al., 2021). For L depth layer with k heads, we attach them at the blocks nearest to $(L/k, 2L/k, \ldots, L)$. The resulting heads form a hierarchy of energy terms, each enforcing the structure visible at its depth. We train on the sum of these energies, allowing each layer to contribute at its own level of abstraction. This not only yields a better energy function but also decomposes the generative process across layers, opening the way to understanding each layer's role and further applications.

3.2 Training Transformer Energy-Based Models

Latent Space. Training E_θ requires sampling negatives by MCMC at every step, which is expensive in high dimensions and mixes poorly. Therefore, we operate entirely in the latent space induced by

the autoencoder of Stable Diffusion (Rombach et al., 2022), which is common in the literature (Peebles & Xie, 2023). Given an input image $x \in \mathbb{R}^{256\times256\times3}$, we encode to latent space $z \in \mathbb{R}^{32\times32\times4}$ where the encoder predicts the mean $\mu_{enc}(x)$ and variance $\sigma^2_{enc}(x)$. We sample latent variables as

$$z \sim \mathcal{N}(\mu_{enc}(x), \sigma^2_{enc}(x)) \tag{3}$$

which corresponds to drawing from the learned posterior distribution of the autoencoder. In this latent space, we define an energy-based model $E_\theta(z, c)$ and perform learning and sampling directly on z. The decoder $D(\cdot)$ maps latent samples back to the image space via $x \approx D(z)$.

Contrastive Energy Objective. We train the energy-based model using a contrastive objective that encourages lower energy for data samples than for model samples.

$$\mathcal{L}_{\text{contrast}} = E(z^+) - E(q_\theta(z_0)) + \lambda \left\{E(z^+)^2 + E(q_\theta(z_0))^2\right\} \tag{4}$$

Here, z^+ denotes an encoded positive sample drawn from the dataset, and $q_\theta(z_0)$ denotes a negative sample produced by sampler. z_0 is from pure noise or replay buffer. Since the likelihood gradient depends only on energy differences, the magnitudes of energies are relative and drift during training. We add an L2 regularization term penalizing the magnitude of the energy to resolve this gauge freedom and bound the energy magnitude following Du & Mordatch (2019).

Sampler Matching Loss. Because the sampler runs for a finite number of steps, its output distribution q_θ may not exactly match the model distribution p_θ from which the negative phase should draw. Following Du et al. (2020), we introduce a regularizer motivated by $\mathrm{KL}(q_\theta \,\|\, p_\theta)$:

$$\mathcal{L}_{\text{sampler-match}} = \underbrace{\mathbb{E}_{q_\theta(\hat{\mathbf{z}}^-)}\left[E_{\mathrm{sg}(\theta)}(\hat{\mathbf{z}}^-)\right]}_{\mathcal{L}_{\text{opt}}} + \underbrace{\mathbb{E}_{q_\theta(\hat{\mathbf{z}}^-)}\left[\log q_\theta(\hat{\mathbf{z}}^-)\right]}_{\mathcal{L}_{\text{ent}}}, \tag{5}$$

where $\hat{\mathbf{z}}^-$ is a sample generated by sampler and the stop-gradient $\mathrm{sg}(\cdot)$ freezes the energy parameters, so gradients reach θ only through the sampler. The first term drives the sampler toward low-energy regions. The second is an entropy term, estimated with a nearest-neighbor statistic (Beirlant et al., 1997), that keeps its outputs diverse. Together, these terms encourage low-energy, diverse sampler outputs. Appendix A, B gives the details and reports training settings.

Representation Supervision. Contrastive EBM training provides only a scalar supervisory signal per sample, one energy value, which makes representation learning challenging at high resolution. A natural remedy is to supply a dense signal to the representation itself, as in diffusion models, where aligning generative features with those of a pretrained vision foundation model (VFM) substantially improves sample quality (Yu et al., 2024; Wang & He, 2025). We study how such alignment should be formulated for energy-based models and introduce *Positive Representation Alignment* (PRA), an objective designed for the contrastive setting.

Let $\bar{h}_{\text{VFM}}(x^+) \in \mathbb{R}^{N\times D}$ denote the spatially normalized patch-wise output of the VFM, where N is the number of patches and D is the embedding dimension. Let $h_\theta(z^+)$ denote the internal representation of the EBM and $P_\phi(\cdot)$ a learnable projection network. PRA maximizes the patch-wise cosine similarity between the projected EBM features and the VFM targets:

$$\mathcal{L}_{\text{PRA}}(\theta, P_\phi) := -\frac{1}{N}\sum_{n=1}^{N} \mathrm{sim}\Big(\bar{h}_{\text{VFM}}(x^+)_{[n]},\, P_\phi\big(h_\theta(z^+)\big)_{[n]}\Big). \tag{6}$$

We apply this alignment loss only to positive (data) samples, and never to MCMC negatives. Two complementary intuitions motivate this choice. From a contrastive learning view, alignment acts as a shaping term on the low-energy side of the landscape, while the objective is simultaneously driving negatives toward high energy. Aligning negatives would therefore pull them back toward the data manifold in feature space, working directly against that push. From a manifold view, the VFM produces meaningful features only for inputs near the natural-data manifold. MCMC negatives are typically off-manifold, and this is most pronounced early in training and at intermediate sampling steps, so their embeddings provide no reliable semantic target to align against.

This positive-only design also clarifies how PRA is both similar and different to representation alignment in diffusion models, where the distinction lies in what each network receives as input. That line of work aligns the features of a noisy input x_t with the VFM features of its clean counterpart x_0; the clean target is dynamic and guides the denoising trajectory toward the data manifold. In our setting

Table 1: **Class-conditional generation on ImageNet-256×256 and 512×512 (FID-50K).** Params refer to the 256×256 configuration unless a resolution-specific model is used. Baseline values are taken as reported in the original papers; "–" indicates that the corresponding setting is or was not reported.

		ImageNet 256×256			ImageNet 512×512		
Method	Params	NFE	FID↓	FID↓ (CFG)	NFE	FID↓	FID↓ (CFG)
GAN							
BigGAN-deep	112M	1	6.95	–	1	8.43	–
GigaGAN	569M	1	3.45	–	–	–	–
Autoregressive / Masked							
MaskGIT	227M	8	6.18	–	12	7.32	–
VAR-d30	2B	10	–	1.92	–	–	–
Diffusion/Flow							
ADM	554M	250	10.94	–	250	23.24	–
LDM-4	400M	250	10.56	3.60	–	–	–
DiT-XL/2	675M	250	9.62	2.27	250	12.03	3.04
SiT-XL/2	675M	250	8.3	2.06	250	–	2.62
SiT-XL/2+REPA	675M	250	5.9	1.42	250	–	2.08
EBM							
TEM-L	415M	60	4.74	–	–	–	–
TEM-XL	609M	60	3.82	–	60	8.24	–

both the EBM and the VFM receive the same clean input x^+, so PRA does not guide any sampling trajectory. Instead it shapes the geometry of the low-energy region of the landscape directly, PRA encourages semantically structured representations of positive samples.

Overall Objective. Combining the contrastive energy objective (Eq. 4), the sampler matching term (Eq. 5), and the representation alignment loss (Eq. 6), the full training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrast}} + \lambda_1 \mathcal{L}_{\text{sampler-match}} + \lambda_2 \mathcal{L}_{\text{PRA}}, \tag{7}$$

where λ_1 and λ_2 are coefficients for each loss. Each term plays a distinct role: $\mathcal{L}_{\text{contrast}}$ shapes the energy landscape by pushing data samples to low energy and sampled negatives to high energy, $\mathcal{L}_{\text{PRA}}$ injects semantic structure into the model's internal representations, and $\mathcal{L}_{\text{sampler-match}}$ encourages low-energy, diverse sampler outputs. The full training algorithm is provided in Appendix D.

3.3 Inference Time Scaling

At inference, we adjust the sampling step size while keeping the injected noise scale fixed:

$$x_{t+1} = x_t - \eta \nabla_x E_\theta(x_t) + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{8}$$

η controls the balance between energy-driven refinement and stochastic exploration. It provides control over sample quality, without retraining or additional unconditional network evaluations.

Increasing the step size η amplifies the deterministic energy-gradient update relative to the fixed noise. Similarly, extending sampling beyond the training horizon allows further refinement of intermediate samples. Empirically, both adjustments can sharpen individual samples and produce more class-typical images at the cost of mode coverage.

4 Experiments

4.1 Main Results

Unless otherwise noted, all TEM experiments use batch size 256, learning rate 2×10^{-4}, and 50K iterations, training with 40 steps sampling and inference with 60 steps sampling. Main benchmark numbers for TEM are obtained with inference-time scaling but without classifier-free guidance. All conditions within each ablation are controlled. Full training details are provided in Appendix B.

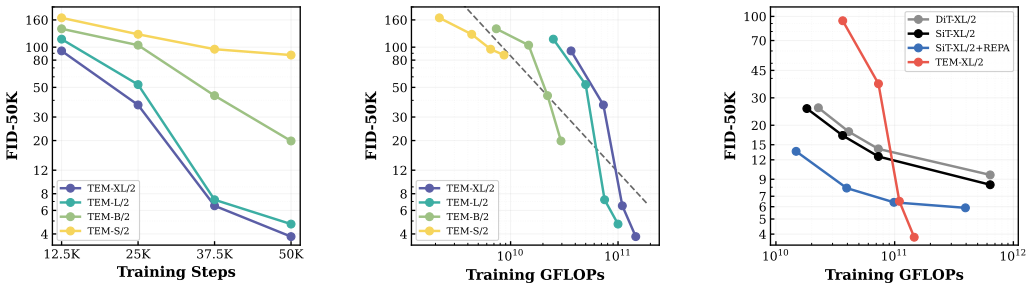
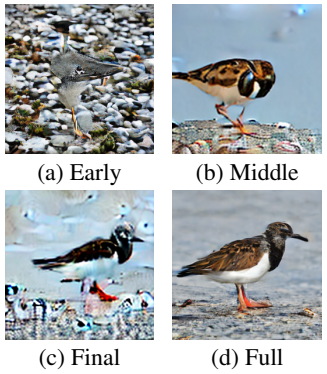
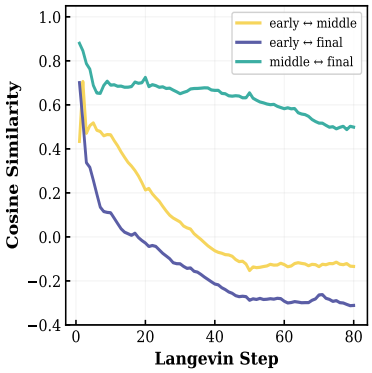


Figure 3: **Scaling behavior of TEM on class-conditional ImageNet 256×256.** *Left*: FID-50K as a function of training steps across model sizes. Larger models consistently achieve lower FID at every checkpoint; 40 sampling steps are used for sampling during training. *Middle*: FID-50K as a function of cumulative training GFLOPs. All four model sizes follow a consistent compute–quality trend (dashed line). *Right*: FID-50K versus training GFLOPs for TEM-XL compared against DiT-XL, SiT-XL and SiT-XL+REPA showing that TEM tracks, and eventually achieves lower FID than unguided diffusion baselines with less training compute.





Variant	FID ↓
Loss terms (3 heads)	
$\mathcal{L}_{\text{contrast}}$ only	121.11
$+\,\mathcal{L}_{\text{opt}}$	46.30
$+\,\mathcal{L}_{\text{ent}}$	42.76
$+\,\mathcal{L}_{\text{PRA}}$	**19.86**
One head	20.95
Two heads	20.50

Figure 4: **Ablations on Multi-energy Head and Loss Terms.** *Left*: generated images from different energy heads, where (a), (b), (c) decompose the energy into early, middle, and final heads respectively, and (d) uses the full energy head. *Middle*: cosine similarity between energy head gradients across Langevin steps. *Right*: FID-50K at 50K steps for TEM-B. The upper block adds each loss term cumulatively with three heads; the lower block varies the number of energy heads under the full loss, against the three-head configuration above.

Our TEM-XL model achieves a guidance-free FID-50K of 3.82 on ImageNet 256×256 and 8.24 on 512×512 (Table 1). To our knowledge, these are the best reported large-scale results for a contrastively trained EBM using its own MCMC negatives. Evaluation uses persistent sampling initialized from Gaussian noise and outperforms unguided diffusion transformers of comparable size. This brings contrastive EBMs into the performance range of modern large-scale generative models. Although EBMs are often considered expensive, TEM's training and inference costs are comparable to, or even lower than, those of diffusion transformers of similar size (Appendix C). As shown in Figure 3, TEM scales with model size in both training steps and training compute (left, middle), and scales more steeply than other baselines, overtaking them at the highest budget (right).

4.2 Ablations

Analysis of Multi-Head Energy Composition. We validate the use of multiple energy functions through three analyses (Figure 4). The three-head model reaches a lower FID than a single- or double-head baseline at 50K steps (right). Fitting three energy functions whose sum must jointly explain the data yields a better energy landscape.

Sampling from individual heads (left) suggests complementary roles. The early head recovers local texture without global structure (a), the middle head captures object layout without fine detail (b), and the final head preserves coarse semantics alone (c), while their sum (d) satisfies all three simultaneously. Gradient cosine similarity between heads (middle) is high early in sampling and decreases in later steps, suggesting a shared initial direction followed by more specialized contributions.

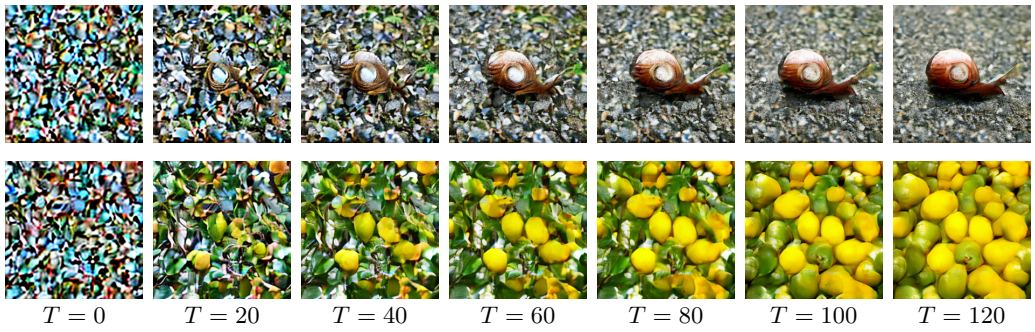


Figure 5: **Progressive sample refinement via extended sampling steps.** Two generation trajectories (top: snail, bottom: lemon) visualized from $T = 0$ (pure noise) to $T = 120$. At the training horizon ($T = 40$), coarse structure is established but residual noise remains. Extending beyond the training horizon produces increasingly sharp, class-typical images, demonstrating inference-time scaling without retraining.

Algorithm 1 Sampling Algorithm

if $|\mathcal{B}| = N$ **then** ▷ $\mathcal{B} \leftarrow \emptyset$ at beginning
  Sample $x_0 \sim \mathcal{B}$
else
  Sample $x_0 \sim \mathcal{N}(0, I)$
end if
for $t = 1$ to T **do**
  $x_t \leftarrow x_{t-1} - \eta \nabla_x E_\theta(x_{t-1}) + \sigma\epsilon$
  if $t = 40$ **then**
    Add x_{40} to replay buffer $\mathcal{B}$
  end if
end for
return x_T

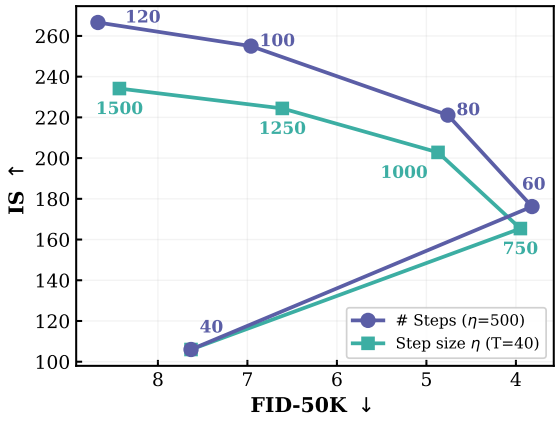


Figure 6: **Inference-time scaling pseudocode and quality–diversity trade-off.** *Left:* MCMC sampling procedure of TEM. Chains can run beyond 40 steps, but only the step-40 state is inserted into the replay buffer, so the buffer stays matched to the training dynamics. *Right:* FID-50K vs. Inception Score for TEM-XL, increasing the number of sampling steps (blue, η=500) or the step size (green, T=40) from the training configuration (T=40, η=500). Moderately increasing the number of steps improves both metrics (FID 3.82 at 60 steps). Beyond that, IS keeps rising while FID degrades as samples concentrate on high-density modes. This mirrors the fidelity–diversity trade-off controlled by classifier-free guidance, but here it comes from a single network with no retraining.

Ablation on Representation Supervision. As shown in Figure 7, PRA lowers FID-50K of TEM-B from 42.8 to 19.9 at 50K steps. Additionally aligning negative samples (Pos+Neg RA) still improves over no alignment but reaches only 36.6, erasing most of this gain. As discussed in Section 3.2, negatives provide a wrong alignment target, and aligning them can even be harmful.

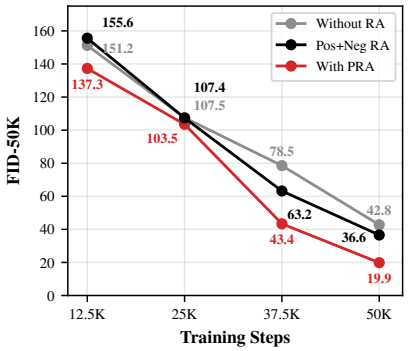


Figure 7: **Effect of PRA on training.** FID-50K across training steps for TEM-B with no representation alignment, Pos+Neg RA, and PRA. PRA consistently outperforms both, reaching 19.9 FID at 50K steps versus 42.8 without alignment and 36.6 with Pos+Neg RA. This supports aligning only positive samples in our method.

4.3 Test-Time Scaling Analysis

We investigate the effect of scaling sampling dynamics at inference time by varying both the number of sampling steps and the step size. Figure 5 visualizes the progressive refinement: at $T = 40$ (the training horizon), samples contain residual noise, but extended sampling produces increasingly sharp, class-typical images.

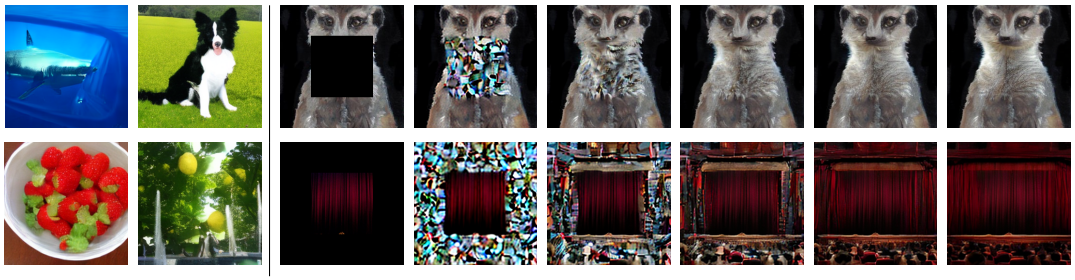


Figure 8: **Application of TEM for Image Composition and Inpainting.** *Left*: composed ImageNet images, car mirror + white shark and border collie + rapeseed (top), strawberry + petri dish and lemon + fountain (bottom). *Right*: inpainting of a meerkat (top) and a movie theater (bottom) across sampling steps. Both applications require no retraining or architectural modification.

Figure 6 quantifies this trade-off. Increasing the number of steps from 40 to 120 and scaling the step size from 500 to 1500 yields higher Inception Scores (indicating more recognizable samples) at the cost of slightly elevated FID (indicating reduced diversity). Notably, this FID-IS trade-off curve closely resembles that of classifier-free guidance, despite requiring no architectural modifications or unconditional network evaluations, but rather just inference-time scaling.

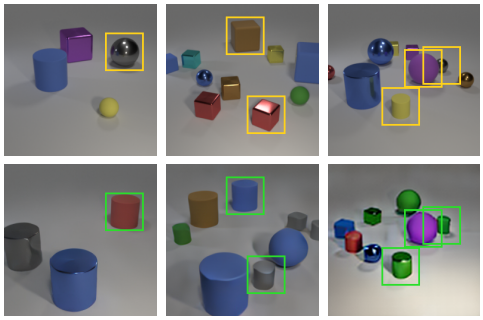
This controllability lets users tune sampling to the application, with longer, larger-step trajectories for fidelity and shorter, smaller-step ones for diversity.

4.4 Applications of TEM

Compositional Generation. Compositional generation arises naturally in the energy-based formulation. Let $p_1(x) \propto \exp(-E_1(x))$ and $p_2(x) \propto \exp(-E_2(x))$ denote two energy-defined distributions. Their composition is defined as the product distribution $p_{\text{comp}}(x) \propto p_1(x)p_2(x)$, which can be equivalently written as $p_{\text{comp}}(x) \propto \exp(-(E_1(x) + E_2(x)))$. Consequently, compositional sampling is obtained by replacing the original energy with the summed energy $E_1(x) + E_2(x)$, enabling joint generation that simultaneously satisfies both energy functions.

We train TEM-B conditioned on object location and generate compositions of multiple objects at multiple specified positions with CLEVR dataset. Figure 9 shows qualitative samples alongside a comparison against Composable Diffusion (Liu et al., 2022), where we report object detection accuracy for 1-, 2-, and 3-object compositions. Following their evaluation protocol, we train the classifier ourselves using their training recipe. TEM outperforms Composable Diffusion at every composition count. Experimental details are provided in Appendix E.

Out-of-Distribution Detection. EBMs naturally provide a density-based score for OOD detection. $E_\theta(x)$ assigns lower energy to in-distribution data and higher energy to OOD inputs (Du & Mordatch, 2019). Thus, a single forward pass of $E_\theta(x)$ directly provides an OOD score without any task-specific training. In contrast, diffusion models do not directly provide such a scalar energy and require an additional scoring procedure based on their score function. For the diffusion baselines, we follow Wyatt



Models	1 obj.	2 obj.	3 obj.
Comp. Diff.	0.863	0.598	0.262
TEM (Ours)	**0.956**	**0.766**	**0.274**

Figure 9: **CLEVR 256×256 compositional generation.** *Top:* conditioning inputs, yellow boxes indicate specified object positions; second row shows TEM-B samples, green boxes indicate detected objects. *Bottom:* compositional generation accuracy (↑), object detection success rate by a trained classifier over 1000 scenes for 1-, 2-, and 3-object compositions.

Table 2: **Out-of-distribution detection on ImageNet benchmarks (AUROC↑).** Comparison of DiT-XL/2, SiT-XL/2, SiT-XL/2+REPA and our TEM-XL.

OOD Dataset	DiT	SiT	SiT-REPA	Ours
iNaturalist	0.777	0.733	0.780	**0.904**
Textures	0.439	0.426	0.438	**0.687**
Places365	**0.669**	0.621	0.481	0.636
SUN	0.505	0.513	0.500	**0.640**
Average	0.598	0.573	0.549	**0.717**

et al. (2022). Using the ImageNet-1k validation set as the in-distribution reference, we report AUROC (↑) on four standard benchmarks (Huang & Li, 2021). Experiment details are in Appendix F.

Inpainting. Inpainting is particularly simple in energy-based models. Our EBM defines a single energy over image latents, allowing refinement from any intermediate completion. We mask the input image before VAE encoding, preventing access to ground-truth content in the missing region. We then apply sampling updates to the missing latent region, restoring the observed-region latents to their initial encoded values after each step. Figure 8 shows image composition and inpainting results.

5 Related Work

Foundations of Energy-Based Models. Energy-based models assign an unnormalized scalar energy to each input and define a Gibbs distribution over the data space, an idea with deep roots in statistical physics and undirected graphical models (Hinton et al., 2006; LeCun et al., 2006). Classical instantiations such as Restricted Boltzmann Machines relied on tractable architectures to make learning feasible, which in turn limited their representational capacity. Du & Mordatch (2019) showed that short-run Langevin MCMC combined with a replay buffer yields a practical training loop, producing coherent samples on CIFAR-10. Follow-up work refined the training recipe through improved MCMC initialization, sampler matching terms, and regularization strategies (Du et al., 2020; Nijkamp et al., 2019; 2020), yet these methods operated at resolutions of 32×32 or less diverse dataset, leaving the scalability of the paradigm unresolved.

Latent-Space Energy-Based Models. Pang et al. (2020) proposed training an EBM as a prior over the latent codes of a variational autoencoder, while Hill et al. (2022) explored energy-based priors for hierarchical latent variable models. Nie et al. (2021) further demonstrated controllable generation by manipulating energies in a learned latent manifold. VAEBM (Xiao et al., 2020) combines a VAE with an EBM refinement in data space, reparameterizing the MCMC chain through the VAE's latent space to make sampling practical at higher resolutions. In contrast, we train a single scalar output Transformer energy function in the frozen latent space of a pretrained autoencoder (Rombach et al., 2022), keeping the EBM as the sole generative component.

Scaling Energy-Based Models. Several recent efforts have pushed EBMs toward higher resolution and stronger performance, though each introduces auxiliary mechanisms that depart from the pure energy-based formulation. EGC (Guo et al., 2023) hybridizes energy-based learning with diffusion modeling and can produce 256×256 ImageNet samples. However, its sampling procedure relies on the full diffusion denoising trajectory. On the adversarial side, Grathwohl et al. (2019) reinterpret classifier logits as energies, and Yin et al. (2026) scale this idea to ImageNet 256×256. However, their sampler is initialized from an auxiliary image dataset at both training and inference, and is never seeded from noise at this scale.

Trained contrastively, TEM generates high-fidelity 256×256 images, enabling inference-time scaling and all EBM native applications. Appendix G provides an extended discussion of related work.

6 Limitations and Conclusion

Limitations. TEM's primary limitation is the computational cost of iterative sampling during both training and inference. One promising direction is to train a feed-forward proposal model to generate candidates in low-energy regions, reducing the sampling needed to refine them. Extended sampling also introduces a fidelity–diversity trade-off: longer trajectories sharpen images but can reduce diversity and worsen FID. Adaptive noise schedules and proposals that enable movement between distinct modes could help preserve diversity while improving sample quality.

Conclusion. In this work, we introduce the Transformer Energy-Based Model (TEM), an approach that addresses long-standing scalability issues in contrastive training of energy-based models. With a modern transformer backbone, multi-layer energy extraction, and representation-aware training, we demonstrate that a contrastively trained EBM can synthesize high-fidelity ImageNet images at 256×256 and 512×512, competitive with other generative models. We also show that the properties that motivate energy-based modeling in the first place, compositional generation, inpainting, and out-of-distribution detection, remain available at this resolution. We hope this work serves as a starting point for further scaling of contrastively trained EBMs.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, including suggestions for additional ablations and evaluation protocols, and to check the derivation in Appendix A for correctness and clarity. We have not used generative AI tools to generate synthetic data sets, develop theoretical models, formulate mathematical claims, interpret results, propose or refine hypotheses, implement methods, or provide critical ingredients for proofs, and translation, dataset cleaning, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to review the manuscript, and edit parts of the paper to improve readability. AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. *Advances in Neural Information Processing Systems*, 38:8583–8609, 2025. 19

Jan Beirlant, Edward J Dudewicz, László Györfi, Edward C Van der Meulen, et al. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997. 4, 14

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm. 19

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022. 19

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014. 18

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 2

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 19

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019. 1, 2, 3, 4, 8, 9, 15

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*, 2020. 1, 3, 4, 9, 14, 15

Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023. 1, 19

Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, and Tariq Iqbal. Energy-based transformers are scalable learners and thinkers. *arXiv preprint arXiv:2507.02092*, 2025. 15, 19

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019. 9

Qiushan Guo, Chuofan Ma, Yi Jiang, Zehuan Yuan, Yizhou Yu, and Ping Luo. Egc: Image generation and classification via a diffusion energy-based model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22952–22962, 2023. 9

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. 3

Mitch Hill, Erik Nijkamp, Jonathan Mitchell, Bo Pang, and Song-Chun Zhu. Learning probabilistic models from generator latent spaces with hat ebm. *Advances in Neural Information Processing Systems*, 35:928–940, 2022. 9

Geoffrey Hinton, Simon Osindero, Max Welling, and Yee-Whye Teh. Unsupervised discovery of nonlinear structure using contrastive backpropagation. *Cognitive science*, 30(4):725–731, 2006. 2, 9

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8710–8719, 2021. 9, 18

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017. 17

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10124–10134, 2023. 19

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 16

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 1, 9

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022. 8, 17, 18

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024. 18, 19

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 18

Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021. 9

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 9

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5272–5280, 2020. 9

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 15

Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems*, 33:21994–22008, 2020. 9

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023. 3, 4, 15, 16, 18, 19

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 3

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 4, 9, 15, 19

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 3

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? *arXiv preprint arXiv:2512.10794*, 2025. 15

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015. 1

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 19

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018. 18

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025. 19

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025. 4

Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 649–655. IEEE, 2022. 8, 18

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010. 18

Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. Vaebm: A symbiosis between variational autoencoders and energy-based models. *arXiv preprint arXiv:2010.00654*, 2020. 9

Xuwang Yin, Claire Zhang, Julie Steele, Nir Shavit, and Tony Wang. Scalable energy-based models via adversarial training: Unifying discrimination and generation. In *International Conference on Learning Representations*, volume 2026, pp. 68385–68424, 2026. 9

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. *arXiv preprint arXiv:2410.06940*, 2024. 4, 18, 19

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 18

A THE SAMPLER-MATCHING LOSS

Maximum likelihood training of an EBM minimizes $\mathcal{L}_{\mathrm{NLL}} = \mathbb{E}_{p_{\mathcal{D}}}[E_\theta(x)] + \log Z(\theta)$, whose gradient is given in Eq. 1: a positive phase over data and a negative phase over samples from p_θ. The negative phase is the obstacle, since exact samples from p_θ require MCMC run to convergence.

In practice we replace p_θ by the distribution q_θ of the final state of a short Langevin chain initialized at $\mathcal{N}(0, I)$ or at a past sample. The resulting gradient,

$$\mathcal{L}_{\mathrm{contrast-only}} = \mathbb{E}_{p_{\mathcal{D}}(x)}[E_\theta(x)] - \mathbb{E}_{\mathrm{sg}(q_\theta(x'))}[E_\theta(x')], \tag{9}$$

is therefore biased by the gap between q_θ and p_θ, and the bias is largest exactly when the sampler fails to reach the model's low-energy regions.

Following Du et al. (2020), we introduce a loss motivated by this gap that encourages low-energy, diverse sampler outputs:

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(q_\theta \,\|\, p_{\mathrm{sg}(\theta)}) + \mathrm{const} = \mathbb{E}_{q_\theta(x)}[E_{\mathrm{sg}(\theta)}(x)] + \mathbb{E}_{q_\theta(x)}[\log q_\theta(x)], \tag{10}$$

where the stop-gradient $\mathrm{sg}(\cdot)$ freezes the energy parameters so that gradients flow only through the sampler, and the constant $-\log Z(\mathrm{sg}(\theta))$ does not depend on the differentiated parameters. Minimizing $\mathcal{L}_{\mathrm{KL}}$ makes the sampler's outputs both low-energy and diverse, which is precisely what an accurate negative phase requires. It also counteracts the failure mode in which the energy function shapes a landscape that its own sampler cannot descend.

A.1 ESTIMATING THE SAMPLER-ENERGY TERM

Eq. 10 splits into two pieces, $\mathcal{L}_{\mathrm{KL}} = \mathcal{L}_{\mathrm{opt}} + \mathcal{L}_{\mathrm{ent}}$, with $\mathcal{L}_{\mathrm{opt}} = \mathbb{E}_{q_\theta(x)}[E_{\mathrm{sg}(\theta)}(x)]$ and $\mathcal{L}_{\mathrm{ent}} = \mathbb{E}_{q_\theta(x)}[\log q_\theta(x)] = -\mathbb{H}(q_\theta)$.

The first is directly tractable: each sampling step is a differentiable function of θ, so

$$\frac{\partial \mathcal{L}_{\mathrm{opt}}}{\partial \theta} = \mathbb{E}_{q_\theta(x^0,\ldots,x^T)}\left[\frac{\partial\, E_{\mathrm{sg}(\theta)}\left(x^{T-1} - \eta \nabla_x E_\theta\left(x^{T-1}\right) + \omega^T\right)}{\partial \theta}\right], \tag{11}$$

which is a second-order gradient term. Backpropagating through all T steps is memory-prohibitive at our sequence lengths, so we differentiate only through the final sampling step, following (Du et al., 2020).

A.2 ESTIMATING THE ENTROPY TERM

The second piece is the obstacle. Evaluating $\log q_\theta(x)$ exactly is intractable: q_θ is the pushforward of the buffer distribution through T sampling steps, and its density at a point requires marginalizing over all trajectories that could have produced it.

We instead use the non-parametric nearest-neighbor entropy estimator (Beirlant et al., 1997), which estimates $\mathbb{H}(p)$ from n samples by their average log-distance to the nearest other sample. Up to factors independent of θ, minimizing $\mathcal{L}_{\mathrm{ent}}$ therefore amounts to

$$\mathcal{L}_{\mathrm{ent}} \;=\; \mathbb{E}_{q_\theta(x)}[-\log \mathrm{NN}(x, \mathcal{B})]\,, \qquad \mathrm{NN}(x, \mathcal{B}) = \min_{x' \in \mathcal{B}} \|x - x'\|_2\,, \tag{12}$$

where $\mathcal{B}$ is a set of past MCMC samples treated as constants (stop-gradient) and distances are L_2 in latent space. Repelling samples from one another increases their diversity. This class of estimators scale poorly with dimension, and the buffer samples are neither independent nor drawn from the current q_θ. As in Du et al. (2020), we use it only as a diversity regularizer rather than as an accurate estimate of $\mathbb{H}(q_\theta)$.

B TRAINING & INFERENCE DETAILS

B.1 TRAINING DETAILS.

The energy head consists of a two-layer MLP with SiLU activations. Details of the transformer architecture are provided in Table 3. All experiments operate in the latent space of a pre-trained Stable

Diffusion VAE (Rombach et al., 2022), following Peebles & Xie (2023) and are trained with bfloat16 mixed precision. The latent space is clipped by norm 100 to stabilize sampling for 256×256 and by norm 200 for 512×512, which is chosen as the threshold containing over 99.99% of ImageNet training latents, which stabilizes sampling. A smaller threshold clips real data points, while a larger one expands the region the sampler has to search, which slightly degrades performance. For sampling, we use a step size of 500 and noise scale of 0.05. With a small step size, the network would have to produce very large input gradients to move samples meaningfully within a limited number of steps, and we found that this destabilizes training. With an excessively large step size, however, each update is coarse, and the model struggles to learn a fine-grained energy landscape. Following prior work (Du & Mordatch, 2019; Du et al., 2020; Gladstone et al., 2025), who likewise decouple the step size from the noise scale and use a step size much larger than the standard Langevin prescription, we select the step size empirically to balance these two effects. During training, 40-step sampling is employed.

The replay buffer's replay probability is set to 0.99 and capacity is 16000. Class labels are drawn independently for every chain, including chains initialized from the buffer. Buffer states therefore serve as class independent initializations near the data manifold, which the conditional energy refines toward the requested class. Once full, each newly added sample replaces a uniformly randomly selected buffer entry. For the sampler matching term, the distance with the nearest neighbor term is calculated based on the distance from 100 randomly sampled seeds to the replay buffer following (Du et al., 2020). We use DINOv2-large (Oquab et al., 2024) as the VFM and implement the projection network P_ϕ as a lightweight CNN with spatial normalization following Singh et al. (2025) attached after the block nearest to one-third of the total network depth. We use $\lambda = 1$ for energy L2 regularization, $\lambda_1 = 1$ and $\lambda_2 = 0.6$ for overall objective's coefficient at Equation 7 following Du et al. (2020) and Singh et al. (2025). We optimize the model using Adam with a learning rate of 2×10^{-4}, momentum parameters $(\beta_1, \beta_2) = (0, 0.999)$, $\epsilon = 10^{-8}$, and no weight decay. We keep an exponential moving average (EMA) of the weights with decay 0.9999, updated after every optimizer step. All reported samples and metrics use the EMA weights. During training, we replace the class label with a null token with probability 0.15, enabling unconditional energy evaluation for OOD detection.

Table 3: **TEM architecture configurations.** Model variants from TEM-S to TEM-XL. All variants use a patch size of 2×2 and operate in the latent space of a pretrained Stable Diffusion VAE. Scaling from S to XL increases both depth and hidden dimension, yielding monotonic FID improvements (see Figure 3).

Model	**TEM-S**	**TEM-B**	**TEM-L**	**TEM-XL**
Params	33M	122M	415M	609M
Depth (layers)	12	12	24	28
Hidden size	384	768	1024	1152
Attention heads	6	12	16	16

B.2 Inference details.

Evaluation uses persistent chains: the first 5,000 chains are initialized from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and subsequent chains continue from buffered 40-step states (Algorithm 1), so no external data enters sampling. Once full, each newly added 40 step state replaces a uniformly randomly selected buffer entry. Class labels are sampled independently for each chain and are randomly mixed throughout evaluation. We apply the same latent-norm clipping threshold used during training. Each output requires T additional sampling steps, including for noise-initialized chains, giving an amortized cost of T steps per sample under this persistent evaluation protocol.

To discuss more about details about fidelity-diversity trade-off at Section 4.3, Table 4 directly quantifies this trade-off: increasing sampling steps from 40 to 80 raises precision from 0.65 to 0.81 while reducing recall from 0.65 to 0.51. This trend is similar to the fidelity-diversity trade-off observed with increasing classifier-free guidance in diffusion models.

Table 4: Fidelity-diversity trade-off. Top: TEM-XL with varying numbers of sampling steps, precision/recall with $k = 3$. Bottom: DiT-XL/2 with varying classifier-free guidance scale g results from Peebles & Xie (2023)

Model	# Steps	FID ↓	IS ↑	Precision ↑	Recall ↑
TEM-XL	40	7.63	106.0	0.65	0.65
TEM-XL	60	3.82	176.2	0.76	0.57
TEM-XL	80	4.72	221.1	0.81	0.51
DiT-XL (g=1.0)	250	9.62	121.5	0.67	0.67
DiT-XL (g=1.25)	250	3.22	201.8	0.76	0.62
DiT-XL (g=1.5)	250	2.27	278.2	0.83	0.57

C Computational Cost

Training the TEM-XL model for 50K steps on ImageNet 256×256 dataset takes approximately 6 days using 4 H200 GPUs. For sampling, it takes about an hour to sample 50K images with 1 H200 GPU for 60 sampling steps. Higher resolution training on 512×512 takes roughly 3–4 times longer than 256×256 on both training and inference.

C.1 Training Cost

Training cost. We follow the FLOP accounting convention of Kaplan et al. (2020), letting F denote the cost of a single forward pass. A standard training step for a diffusion transformer costs $3F$ per image: one forward pass and a backward pass of roughly $2F$, since gradients are taken with respect to both the activations and the parameters. A TEM step has four components:

- **Negative sampling.** Each of the T_{train}=40 sampling steps requires $\nabla_{\mathbf{z}} E_\theta(\mathbf{z})$, a gradient with respect to the input only, which costs approximately F in addition to the forward pass. The chain therefore costs about $2FT_{\text{train}} = 80F$.
- **Sampler matching loss.** This term calculates the second-order gradient for last sampling step which takes about $8F$ based on measurement.
- **Contrastive loss and parameter update.** The energies of the positive and negative batches are each evaluated and backpropagated through, about $3F$ each.
- **Representation alignment.** One forward pass of the frozen VFM, which is not backpropagated through. For DINOv2-L at our token count this is approximately $0.9F$.

Despite this per-step overhead, TEM reaches a lower guidance-free FID with less total training compute, because it converges in far fewer steps. Table 5 compares the total training cost of TEM with that of representative models.

Table 5: **Training cost on ImageNet 256×256.** All values use the same accounting convention. VFM forwards are included for TEM and SiT+REPA.

Model	Training steps	Training FLOPs	FID ↓
DiT-XL/2	7M	638 E	9.62
SiT-XL/2	7M	638 E	8.30
SiT-XL/2+REPA	4M	393 E	5.90
TEM-XL	50K	146 E	3.82

C.2 Inference Cost

Because each sampling requires the input gradient $\nabla_{\mathbf{z}} E_\theta(\mathbf{z})$ at every step, the number of steps is not directly comparable to that of diffusion models, we report FLOPs instead. Following the accounting convention of Kaplan et al. (2020), let F denote the cost of one forward pass. Sampling requires only the gradient with respect to the input, which costs approximately F, as opposed to roughly $2F$ when parameter gradients are also required. Each sampling step therefore costs about $2F$, and

Algorithm 2 Transformer Energy-Based Model (TEM) Training

Input: Data distribution $p_D(\mathbf{x})$, encoder $\mathcal{E}$, frozen VFM h_{VFM}, sampling step size η, noise scale σ, number of steps T, loss coefficients $\lambda, \lambda_1, \lambda_2$, replay probability p_{replay}
Initialize: Replay buffer $\mathcal{B} \leftarrow \emptyset$
while not converged **do**
  # 1. Sample positive data and initialize negative chains
  Sample real images $\mathbf{x}^+ \sim p_D(\mathbf{x})$; $\mathbf{z}^+ \leftarrow \mathcal{E}(\mathbf{x}^+)$
  Sample $\tilde{\mathbf{z}}^0 \sim \mathcal{B}$ with probability p_{replay}, else $\tilde{\mathbf{z}}^0 \sim \mathcal{N}(0, I)$
  # 2. Sampling
  for $t = 1$ **to** T **do**
    $\tilde{\mathbf{z}}^t \leftarrow \tilde{\mathbf{z}}^{t-1} - \eta \nabla_{\mathbf{z}} E_\theta(\tilde{\mathbf{z}}^{t-1}) + \omega, \quad \omega \sim \mathcal{N}(0, \sigma^2)$
  end for
  $\hat{\mathbf{z}}^- \leftarrow \tilde{\mathbf{z}}^T$; $\mathbf{z}^- \leftarrow \text{sg}(\tilde{\mathbf{z}}^T)$ ▷ with and without gradient
  # 3. Compute Training Objectives
  $\mathcal{L} \leftarrow \frac{1}{N} \sum \Big(E_\theta(\mathbf{z}^+) - E_\theta(\mathbf{z}^-) + \lambda \big(E_\theta(\mathbf{z}^+)^2 + E_\theta(\mathbf{z}^-)^2 \big)$
    $+ \lambda_1 \big(E_{\text{sg}(\theta)}(\hat{\mathbf{z}}^-) - \log \text{NN}(\hat{\mathbf{z}}^-, \mathcal{B}) \big) + \lambda_2 \mathcal{L}_{\text{PRA}}(\mathbf{x}^+, \mathbf{z}^+) \Big)$
  # 4. Backpropagation and Parameter Update
  $\Delta\theta \leftarrow \nabla_\theta \mathcal{L}$
  # 5. Update persistent buffer
  $\mathcal{B} \leftarrow \mathcal{B} \cup \{\mathbf{z}^-\}$
end while

sampling with K steps costs $2FK$. Table 6 compares the sampling cost of TEM with that of other models.

Table 6: **Sampling cost on ImageNet 256×256.** All results are without classifier-free guidance.

Model	Steps / Sample	FLOPs / Sample	FID ↓
DiT-XL/2	250	29.66 T	9.62
SiT-XL/2	250	29.66 T	8.30
TEM-L	60	9.87 T	4.74
TEM-XL	60	14.48 T	3.82

Although each sampling step costs roughly twice the cost of a single forward pass, TEM uses far fewer steps while reaching a lower guidance-free FID. Classifier-free guidance would improve the baselines' quality but doubles their sampling cost. Although accelerated solvers can further reduce the cost of diffusion sampling, the overhead of computing input gradients is modest in practice, and TEM attains a lower guidance-free FID at a comparable sampling cost.

D Training Algorithm

The full training procedure for TEM is given in Algorithm 2.

E Compositional Generation in 256x256 CLEVR Dataset

To quantitatively evaluate TEM's compositional performance, we train TEM on a 256×256 location-conditioned CLEVR (Johnson et al., 2017) dataset consisting of 70,000 images. Prior work on compositional generation evaluates at 128×128 with a classifier trained at that resolution. We extend this setting to 256×256 to test whether the compositional advantage of energy-based models persists at higher resolution, which is a harder generative setting. The following describes the training setup and the evaluation protocol used to compare against Composable Diffusion (Liu et al., 2022).

Training a location-conditioned TEM on CLEVR 256×256. We train our TEM-B model and evaluate it by providing object locations from real images and composing images according to the number of objects. We condition the energy network on a target object location $\mathbf{p} = (x, y) \in [0, 1]^2$ (normalized image coordinates) through two complementary pathways: a global coordinate code

injected by adaptive layer normalization, and a spatial bias added directly to the patch-token grid. To condition TEM on location, we encode each coordinate with a NeRF-style Fourier feature encoding,

$$\gamma(c) = \left[c,\ \sin(2^0\pi c),\ \cos(2^0\pi c),\ \ldots,\ \sin(2^{L-1}\pi c),\ \cos(2^{L-1}\pi c)\right], \tag{13}$$

which gives a richer positional signal than raw coordinates (Mildenhall et al., 2021). With L=8 frequency bands, (x, y) maps to a $2(1+2L)$=34-dimensional code, which is linearly projected, combined with the (optional) attribute embeddings, and injected into every transformer block via AdaLN. We train for 150k steps with batch size 64.

Evaluation protocol. Composable Diffusion (Liu et al., 2022) operates at 128×128, and its evaluation classifier is trained at this resolution. We use their released 128×128 CLEVR checkpoint directly, and train the classifier ourselves following their training recipe with 128×128 CLEVR dataset. TEM generates at 256×256, and we downsample its samples to 128×128 only for scoring, so both methods are judged by the same classifier at the same resolution. The classifier reaches over 99.9% detection accuracy on real 128×128 CLEVR images, so the reported accuracies reflect generation quality rather than detector error. We report detection success rate over 1,000 scenes for 1-, 2-, and 3-object compositions.

F Out-of-Distribution Detection

TEM as an OOD Detector. TEM provides a scalar energy that we evaluate as an OOD score, using lower energy to indicate more ID-like inputs. We evaluate this property as a secondary capability of TEM. Specifically, we use a TEM-XL model trained with class-label dropout: during training, the class condition is randomly replaced with a null token, allowing the model to learn an unconditional energy function $E_\theta(\mathbf{x})$ alongside the conditional one. At test time, we query the unconditional branch and use the scalar energy $S(\mathbf{x}) = -E_\theta(\mathbf{x})$ as the OOD score in a single forward pass, with no class labels, post-hoc calibration, or OOD exposure.

Diffusion Baselines. As baselines we use pretrained DiT-XL/2 (Peebles & Xie, 2023), SiT-XL/2 (Ma et al., 2024), and SiT-XL/2+REPA (Yu et al., 2024) as ImageNet-trained generative baselines. We use the official released checkpoints: 7M training iterations for DiT-XL/2 and SiT-XL/2, and 4M for SiT-XL/2+REPA. Each test image x is encoded to a latent $z_0 = \mathcal{E}(x)$ by the Stable Diffusion VAE. For DiT, we construct $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, and the network is queried under the null class token $y = \varnothing$ (the learned embedding at index 1000, used for classifier-free guidance dropout during training). The OOD score is the denoising residual

$$s(x) = \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(z_t, t, \varnothing)\right\|^2, \tag{14}$$

where lower $s(x)$ indicates a more in-distribution input, following Wyatt et al. (2022). We score DiT using its noise-prediction residual and SiT-based models using their velocity-prediction residuals, with each model's native interpolation and time convention. All baselines use a single network evaluation, matching the single-forward-pass cost of TEM. For DiT, we use timestep t=350. For SiT+REPA, we select the timestep that matches the same noise level.

We include SiT-XL/2+REPA to separate two possible explanations for TEM's performance. Since TEM is trained with PRA, its energy function inherits semantic structure from VFM, and one might attribute the gain to that supervision rather than to the energy score itself. SiT-XL/2+REPA receives the same DINOv2 alignment signal but is scored by its denoising residual, and its AUROC is close to that of SiT-XL/2 (Table 2). Alignment to a vision foundation model therefore does not by itself account for the gap. What distinguishes TEM is that the energy provides a single scalar that directly ranks inputs by fit to the learned distribution.

Evaluation Protocol. Following Huang & Li (2021), we use the ImageNet-1k validation set (50,000 images) as the in-distribution reference and evaluate on four OOD benchmarks spanning a range of semantic distances from ImageNet: *iNaturalist* (Van Horn et al., 2018), *SUN* (Xiao et al., 2010), *Places365* (Zhou et al., 2017), and *Textures* (Cimpoi et al., 2014). We report AUROC (↑).

G Additional related work

G.1 Baselines in Table 1

All baseline numbers are taken from the original papers; a dash indicates that the corresponding setting was not reported. Among GANs we compare against BigGAN-deep (Brock et al., 2019) and GigaGAN (Kang et al., 2023). For autoregressive and masked models we report MaskGIT (Chang et al., 2022) and VAR-d30 (Tian et al., 2024). The diffusion and flow baselines are ADM (Dhariwal & Nichol, 2021), LDM-4 (Rombach et al., 2022), DiT-XL/2 (Peebles & Xie, 2023), SiT-XL/2 (Ma et al., 2024), and SiT-XL/2 with REPA (Yu et al., 2024). Parameter counts refer to the 256×256 configuration unless a resolution-specific model was released.

G.2 Scaling Energy-based models

A parallel line of work attains scale by learning energy landscapes from predictive targets rather than from negative samples drawn from the model itself. Du et al. (2023) assign a scalar energy to each diffusion timestep via denoising score matching, recovering MCMC-based composition but forgoing a single global energy. Equilibrium Matching (Wang & Du, 2025) fits a time-invariant equilibrium gradient through a regression objective, defining the energy only implicitly through its gradient field. Its explicit scalar-energy variants (EqM-E) substantially degrade generation quality (FID 73.40 and 75.53 in the B/4 setting). Energy-Based Transformers (Gladstone et al., 2025) learn an energy landscape by backpropagating a reconstruction loss through unrolled inner-loop optimization. Energy Matching (Balcerak et al., 2025) explicitly parameterizes a scalar potential, but achieves its high-fidelity generation results by combining an optimal-transport objective with contrastive energy learning. Its image generation experiments are limited to 32×32 resolution (CIFAR-10 and unconditional ImageNet 32×32), so it is not included in Table 1.

In contrast, TEM directly parameterizes a single scalar energy and trains it contrastively using only negatives produced by its own MCMC sampler, without a transport or flow-matching objective. Thus, while prior work demonstrates that energy parameterizations can scale when coupled with predictive, transport-based, or reconstruction-style objectives, our results show that an explicit, contrastively trained EBM can itself scale to ImageNet 256×256.