FoMo: Forking Moment in Generative Trajectory as a Perceptual Distance

Jaihyun Lew[1], **Mingi Jung**[2], **Minjun Park**[1], **Wooseok Song**[2], **Sungroh Yoon**[1,2,3,*]
[1]Interdisciplinary Program in AI, Seoul National University
[2]Department of Electrical and Computer Engineering, Seoul National University
[3]AIIS, ASRI, INMC, and ISRC, Seoul National University
{fudojhl, 2019-16552, minjunpark, cody1129, sryoon}@snu.ac.kr

Abstract

Reference-based image quality assessment (IQA) metrics aim to reflect how humans perceive the perceptual distance between a pair of images. To learn how the human visual system (HVS) operates, recent reference-based IQA metrics heavily rely on human-annotated data. Mean opinion score (MOS)-based pointwise scoring, which assigns a scalar quality value per image, is preferable for annotation but is prohibitively expensive to collect at scale and is known to be noisy due to inconsistent human judgments. As an alternative, two-alternative forced choice (2AFC) pairwise labels have gained popularity due to their reliability and efficiency, but they capture only relative comparisons between pairs. In this paper, we propose a fully automated data generation pipeline that generates pointwise perceptual distance labels between image pairs without any human annotation. Our approach exploits the generative dynamics of diffusion models as a perceptual distance proxy, where the coarse structure of an image is generated in the early timesteps and the fine details are generated in the later timesteps. Images that fork early in the generation process share only coarse structure and are perceptually far apart; images that fork late differ only in fine detail. We demonstrate that the diffusion trajectory aligns well with the human visual system, and use this forking moment, FoMo, as a reference-grounded distance label to supervise the training of a reference-based IQA metric. The pointwise labels, which support universal comparison between arbitrary image pairs, enable an information-rich training objective. Extensive experiments across diverse backbone architectures confirm the effectiveness of our generation pipeline, outperforming human-annotated datasets in multiple benchmarks. Codes are publicly available at: https://github.com/JHLew/FoMo

1 Introduction

Reference-based image quality assessment (IQA) metrics [9, 13, 31, 43, 41] aim to quantify the perceptual difference between a reference image and its distorted counterpart. In particular, they play a central role in diverse image restoration tasks, such as super-resolution and denoising, where the goal is to compute the distance between a restored image and a reference target image. A reliable metric must therefore align closely with human perceptual judgments, not only distinguishing which of two images is closer to the reference, but also inducing a globally consistent ordering across diverse distortion types and severity levels.

To train such metrics, IQA datasets [43, 31, 23]have employed different forms of human supervision to approximate perceptual rankings. Mean opinion score (MOS) [30, 23] is the most direct and standard approach to produce such globally ordered labels. Multiple observers independently rate

*Corresponding Author

Preprint.

each distorted image on an absolute quality scale; the mean score is used as ground truth, and its global structure naturally supports rank-correlation training objectives [30]. However, reliable MOS collection is both expensive and fragile. Due to annotator bias and inter-session inconsistency, a large number of responses per image is required to suppress noise. KADID-10k [23], one of the largest MOS-annotated reference-based IQA datasets, required 30 crowdsourced ratings per image from over 2,200 subjects to produce 10,125 distorted images derived from only 81 reference images. Furthermore, labels are known to be inconsistent across datasets: the same distorted image can receive substantially different scores across datasets, because no common perceptual reference point exists across them. [31]

These difficulties have driven the field toward pairwise preference labeling. The two-alternative forced choice (2AFC) protocol, asking annotators which of two distorted images is more similar to a reference, is considerably more reliable than absolute rating, as relative judgments are less susceptible to individual-scale biases [31, 43]. BAPPS [43] and PieAPP [31] established 2AFC as the foundation for learning modern perceptual metrics, and both demonstrate that pairwise labels exhibit higher inter-annotator agreement than MOS under equivalent collection conditions. 2AFC has since become the dominant annotation paradigm for learning-based IQA. Yet, pairwise preference labels carry a structural limitation: a set of binary pairwise outcomes does not directly encode a global ordering, and global ranking of images are never taken into account in training. This tension between the practical tractability of pairwise labeling and the global ranking objective of IQA is a recognized open challenge [38, 5]. Ideally, one would have access to dense labels that directly support optimization toward global rank-correlation. In practice, however, this remains infeasible at scale under human annotation: collecting clean and consistent pointwise quality signals across thousands of images, while controlling for the annotator noise endemic to MOS, is prohibitively expensive.

In this paper, we propose to circumvent this bottleneck through an automated dataset generation pipeline grounded in the generative dynamics of diffusion models. [15, 36] Our key insight is that the generative dynamics of diffusion models provide a natural proxy for perceptual distance. During generation, coarse image structure is established in the early timesteps, while fine-grained details are resolved only in later timesteps. In this work, we define a fork as a controlled branching of the denoising process: two samples follow an identical trajectory up to a selected timestep and are then generated independently thereafter. Images that fork early in the generation process share only coarse structure and are perceptually far apart, whereas images that fork late differ primarily in fine detail. We use this forking moment, FoMo, as a reference-grounded distance label for training a reference-based IQA metric.

To validate this intuition, we conduct empirical analyses using a controlled set of image pairs generated via forking moments. We verify that their induced perceptual ordering aligns with human judgments, providing empirical grounding for using diffusion dynamics as a perceptual proxy. Building on this validation, we propose a fully automated and reference-grounded data generation pipeline that derives pointwise perceptual distance labels from diffusion forking moments. This enables global comparison across arbitrary image pairs without any human annotation, crowdsourcing infrastructure, or inter-annotator reconciliation.

Since FoMo is a pointwise score, it directly supports training consistent global ranking rather than binary pairwise preference. We extensively validate the effectiveness of our method across multiple benchmarks and diverse architectural backbones, from CNN-based models [19] to Transformer-based models. [40] In particular, our method outperforms human-annotated approaches, including KADID-10k's MOS labels, showing that automated diffusion-based labels can surpass large-scale human annotation as a strong and reliable training signal. These results demonstrate that the FoMo can serve as a scalable, annotation-free alternative to both MOS and pairwise human labeling paradigms. Our key contributions are summarized as below:

- We propose a fully automated and annotation-free data generation pipeline for reference-based IQA that derives pointwise perceptual distance labels from the forking moments of diffusion trajectories, eliminating the need for human annotation.

- We demonstrate that diffusion generative dynamics encode perceptual distance in a manner well aligned with human visual judgment, providing a scalable alternative to MOS and pairwise preference annotations.

- We show that FoMo supervision enables globally consistent ranking across distorted images, and training with a RankNet-style [2] global objective improves reference-based IQA performance across diverse benchmarks and model architectures.

2 Preliminary

Diffusion Models, Flow Matching, and Rectified Flows. Denoising diffusion probabilistic models (DDPM) [15] define a forward Markov process that gradually corrupts a clean image x_0 by adding Gaussian noise over T discrete timesteps, yielding a sequence of increasingly noisy images $x_1, x_2, \ldots, x_T$ where $x_T \sim \mathcal{N}(0, I)$. A neural network is trained to reverse this process, iteratively denoising x_T back to a clean sample x_0. Score-based generative models [36] generalize this to a continuous-time stochastic differential equation (SDE) framework, unifying many discrete diffusion variants under a single formalism. Flow matching [24] and rectified flows [25] instead parameterize the generative process as a probability flow ODE along linear interpolations between data and noise: $x_t = (1-t)x_0 + t\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $t \in [0, 1]$. The resulting trajectories are straighter and more sample-efficient, motivating their adoption in state-of-the-art models such as FLUX [20].

Despite differences in trajectory geometry and training formulation, all three families share the same fundamental structure: a forward process that progressively destroys image information from fine detail toward coarse structure, and a learned reverse process that recovers the image from a stochastically sampled intermediate state. FoMo is grounded in this shared structure. Throughout this paper, we describe our method in the continuous-time SDE framework, as it provides the most general formulation. Most experiments in this paper are conducted with FLUX, which uses a rectified flow formulation where forward corruption follows $x_t = (1-t)x_0 + t\epsilon$.

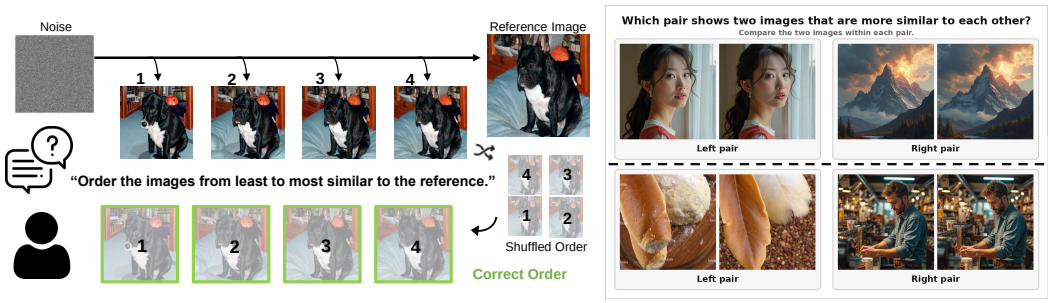
Perceptual Structure Along the Generative Trajectory A key premise of FoMo is that the generative trajectory encodes perceptual information in a structured, timestep-dependent manner. Choi et al. [6] provide a direct characterization: at low noise levels (high signal-to-noise ratio), the reverse process recovers imperceptible fine-grained details; at intermediate noise levels, it reconstructs perceptually rich and discriminative content such as object structure and texture; at high noise levels, it recovers only coarse global attributes such as color distribution. This stratification is not incidental, it is a structural consequence of the forward process, which destroys information roughly monotonically from high-frequency to low-frequency.

These observations directly motivate FoMo. If two images share a reverse trajectory up to t and then diverge via independent re-sampling, the perceptual content preserved up to t is shared between them, while content destroyed before t is independently regenerated. A late divergence (small t, little corruption) leaves most perceptual detail intact, yielding a perceptually close pair. An early divergence (large t, heavy corruption) destroys most discriminative content before re-sampling, producing a substantially different pair.

Trajectory Divergence as a Structural Prior The use of intermediate trajectory states to induce structured variation in generated outputs has appeared across several independent lines of work, lending support to the generality of the forking moment. [28, 33, 7] Most directly, Decatur et al. [7] demonstrate that when generating a collection of semantically related images, early denoising steps capture shared structure across similar prompts and need only be computed once; trajectories then branch independently from a later timestep onward. This explicitly instantiates a tree-structured forking process, and their findings confirm that the branching timestep controls the degree of visual similarity among the resulting outputs. FoMo builds on this foundation by converting the forking structure into an explicit, scalable source of perceptual distance labels, replacing human annotation with the generative process itself.

3 Empirical Grounding

Central to our approach is the hypothesis that the point of divergence in the diffusion sampling trajectory can serve as a meaningful perceptual distance label between image pairs. Before formalizing this as a metric, we first evaluate whether the divergence point serves as a reliable proxy for human perceptual similarity. That is, whether images forked earlier in the denoising process are consistently perceived as less similar to the reference than those forked later.



(a) (b)

Figure 1: (a) Single-reference study: participants rank synthesized variants to verify whether the divergence timestep in diffusion generative trajectory reflects the degree of perceptual similarity to the reference image. (b) Cross-reference study: participants are asked which of two pairs, built from different references, holds the two more similar images. Further examples in the Appendix.

3.1 Single-Reference Validation

Study Design To verify whether the divergence timestep provides perceptually meaningful guidance, we conducted a human study examining whether variants that diverge later in the denoising trajectory are consistently perceived as more similar to a reference image than those branching at earlier timesteps. For each reference image, we synthesized five variants by injecting Gaussian noise at five distinct timesteps and denoising from each of them, so that later injection timesteps correspond to smaller perturbations and higher expected perceptual similarity to the reference. Participants were shown a reference image alongside its five variants and were asked to rank the variants from most to least similar compared to the reference. The workflow of this study is illustrated in Fig. 1a.

Results According to our experimental analysis, human perceptual judgments demonstrate strong alignment with the divergence timestep ordering, yielding a Spearman rank correlation of 0.970 across 30 participants and 1,982 responses collected over 190 reference images sampled from the ImageNet [8] validation set. Given an inter-rater correlation of 0.960, this level of agreement suggests that observer judgments are both consistent and well-structured. Collectively, these results indicate that the diffusion forking timestep constitutes a reliable proxy for perceptual similarity, providing empirical grounding for its adoption as a distance label in the subsequent metric formulation.

3.2 Cross-Reference Validation

Study Design The study above establishes that the forking timestep orders variants of a single reference consistently with human perception. A perceptual distance, however, should also be globally consistent: if an image pair is labeled to be closer than another, they should look closer, regardless of which reference image anchors each pair. We therefore ran a second study in the strict two-alternative forced-choice (2AFC) format. Each item shows two reference: variant pairs built from two distinct references and asks which pair contains the images more similar to each other. Since each forking timestep is chosen before its variant is generated, the two timesteps alone determine which pair our label calls closer, and no human answer enters the label. How hard an item is depends on the gap between its two forking timesteps: a small gap means both pairs were forked at nearly the same point, so they are almost equally similar, whereas a large gap sets a barely altered pair against a heavily altered one. We constructed 250 items, stratified into five bins of 50 by this gap, used each reference in at most one item, randomized left/right placement per participant, and collected 5,713 responses from 28 participants. Example questions from this study are in Fig. 1b.

Results Human choices agree with the ordering induced by the forking timesteps in 90.2% of individual responses, with a Fleiss' κ [12] of 0.82 indicating almost perfect inter-rater reliability [21]. Because every item was judged by many participants, we can also ask what they concluded collectively rather than one response at a time: taking the majority answer for each item, the label agrees with the human consensus on 92.8% of items. Agreement rises monotonically with the gap: 64.3% for gaps of 1–10 timesteps, then 90.0% for gaps of 11–20, 97.0% for 21–30, and 99.5% and 99.8% for 31–40 and 41–50. Counting consensus rather than individual votes, the share of items whose

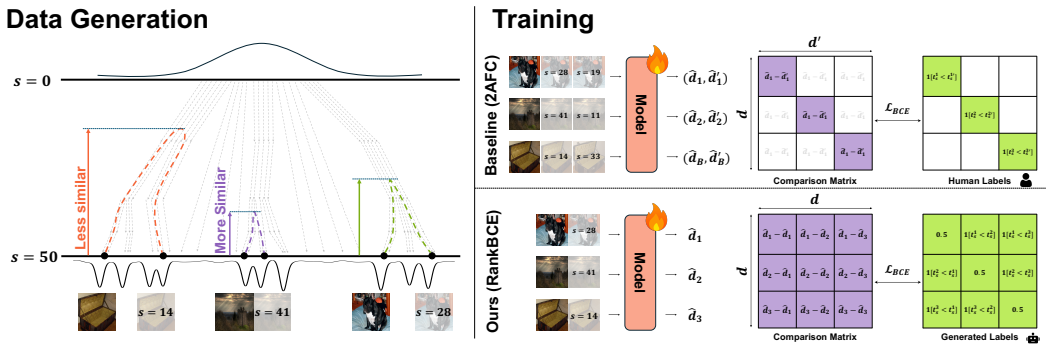


Figure 2: Visualization of our overall framework. (Left) Data generation trajectories, where the divergence point of the denoising trajectory (solid arrows) serves as a perceptual distance proxy. (Right) Unlike the traditional 2AFC framework, limited to within-anchor comparisons, our pointwise scoring system enables inter-anchor comparisons.

majority answer matches the label runs 69.4%, 94.0%, 100%, 100% and 100% across the same five bins: beyond a gap of 20 timesteps, every item is decided the way the label predicts. Where the label agrees with people least, people also agree least with one another: in the narrowest bin two randomly chosen participants give the same answer on only 75.4% of items and just 20% of items are decided unanimously ($\kappa = 0.50$), against 99.6%, 96% and $\kappa = 0.99$ in the widest. The forking timestep therefore induces a similarity ordering that holds across different reference images, breaking down only where the two labels are too close to call.

Validation at Scale Since it is extremely difficult to conduct these comparisons at scale by human annotation, we repeat the same test with established perceptual metrics standing in for the human observer. LPIPS-Alex [43], LPIPS-VGG [43], DISTS [9] and DreamSim [13] are all fitted to human judgments and widely used as proxies for them, which makes them a reasonable substitute here. We take 20,000 reference–variant pairs from our generated data and measure the distance each metric assigns to every pair. Pooling them into a single ranked list means that, as in the study above, almost every comparison is between pairs built on different references. Against that pooled ranking the forking label reaches a Spearman Rank Order Correlation Coefficient (SROCC) of 0.932 with LPIPS-Alex, 0.915 with LPIPS-VGG, 0.904 with DISTS and 0.904 with DreamSim. The ranking departs from the label only where the raters also hesitated, at near-ties. Once the two forking timesteps differ by more than 20 steps, the metrics agree with the label over 99.4% of the time. A pooled correlation of this magnitude is attainable only if the label is comparable across reference images rather than merely monotone within each one. The forking timestep behaves as a globally consistent distance label, not just a per-reference ranking.

4 Method

Based on the validation of Section 3, we present our dataset generation pipeline with fully automated labeling, and detail the training procedure for learning a perceptual distance metric from it.

4.1 Dataset Construction

For data generation, we use FLUX.1-dev [20] for its high-quality synthesis capability and broad coverage of visual content diversity. Given a reference image x_0, we aim to generate a perturbed variant paired with a distance label that is precise and requires no human annotation.

Assume a denoising process with S total steps. We uniformly sample a forking step $s \in [0, S-1]$, and obtain the corresponding interpolation factor from a predefined noise schedule $\mathcal{S}$, *i.e.*, $t_s = \mathcal{S}(s)$. We then construct a noisy latent as

$$x_{t_s} = (1 - t_s)x_0 + t_s\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

Starting from x_{t_s}, we perform the remaining $S - s$ denoising steps to obtain a perturbed image variant x_0^s. This yields a labeled pair (x_0, x_0^s, t_s), where t_s serves as the distance label. Intuitively, a larger t_s

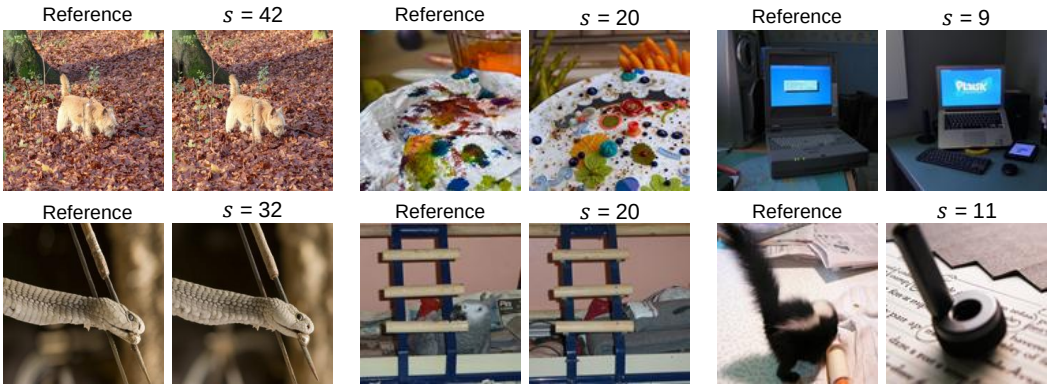


Figure 3: Data sample image pairs and their labels from our constructed dataset. s denote the forking step (smaller-the-further), or the distance label of the image with respect to the reference image.

corresponds to a higher noise level at the forking point and therefore to a greater perceptual deviation from the reference image. Although stochasticity is inherent to the diffusion process, the automated nature of label generation enables large-scale sampling, reducing label variance and leading to stable convergence (See Sec. F of Appendix.). Data samples from our constructed dataset provided in Fig. 3.

4.2 Objective Function

Conventional perceptual metrics such as LPIPS are trained on human-annotated 2AFC datasets, where each label encodes a relative preference between two distorted images given a shared reference. This relative structure constrains the loss to triplet-wise comparisons: given a triplet (x_0, x_0', x_0''), binary cross-entropy is applied to the predicted probability that one variant is closer to the reference than the other.

Our labels, by contrast, are pointwise: each pair (x_0, x_0^s) carries an independent distance value t_s, without requiring a shared anchor for comparison. This permits a more expressive training objective. Specifically, for a batch of B pairs with predicted distances $\{\hat{d}_i\}_{i=1}^B$ and labels $\{t_s^i\}_{i=1}^B$, we define a ground-truth comparison matrix $Y = [y_{ij}]$ as:

$$y_{ij} = \begin{cases} \mathbb{1}[t_s^i < t_s^j] & t_s^i \neq t_s^j, \quad i,j \in \{1,\dots,B\} \\ 0.5 & \text{otherwise} \end{cases} \tag{2}$$

where $y_{ij} = 1$ indicates that pair i has a smaller true distance than pair j. We then apply binary cross-entropy loss over all $B \times B$ comparisons:

$$\mathcal{L} = -\frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \Big[y_{ij} \log \sigma(\hat{d}_j - \hat{d}_i) + \texttt{sg}((1 - y_{ij}) \log \sigma(\hat{d}_i - \hat{d}_j)) \Big], \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\texttt{sg}$ stands for stop-gradient operation. This formulation, originally proposed for learning-to-rank in RankNet [2], is here adapted to perceptual distance learning, supervising the global ordering of distances across all $B \times B$ pairs rather than within isolated triplets. It makes full use of the pointwise label system from our data generation process.

5 Experiments

5.1 Experimental Settings

We use a maximum of $S = 50$ sampling steps with FLUX, and the forking moment is sampled from a uniform distribution: $s \sim U[0, 49]$. For training, we generate a total of 480k pairs and labels. Of these, 240k pairs use real images sampled from the ImageNet database [8] as references, and the other 240k pairs use synthetic images generated by FLUX as references. We mix these two domains of reference images in order to ensure a good coverage of real-world and synthetic domains.

We experiment both on models with CNN backbones and Transformer backbones. For CNN backbones, we experiment with the well-established LPIPS [43] and DISTS [9] backbones. For Transformer backbones, we base upon the strong foundational models, DINOv3 [35], CLIP [32] and MAE [14], along with a recent DreamSim [13] backbone. Following the experimental setting from Zhang et al. [43], we use the feature extractors fixed to the pre-trained state, and train the prediction head from scratch, except for DreamSim, which tunes the feature extractor with LoRA [17] in its specific setting. All evaluations are conducted under the native resolution of each image, except for DreamSim, which enforces 224 resolution images by resizing the input at all times. All experiments reported in this paper were trained with models with approximately 240k iterations of updates, using a single NVIDIA RTX A40 GPU. All quantitative results in the main manuscript is reported by Spearman Rank Order Correlation Coefficient (SROCC), unless otherwise mentioned. Further details on experimental setting are described in the Appendix.

5.2 Quantitative Evaluations

Table 1: Comparison of training datasets and their objectives on four reference-based IQA benchmarks, across CNN- and Transformer-based backbones. Mean SROCC over five random seeds; standard deviations are given in Table 6. † Evaluated under 224 resolution.

Train Data	Human Annot.	Objective	CNN-based			Transformer-based				Avg.
			LPIPS-Alex	LPIPS-VGG	DISTS	DINOv3	CLIP	MAE	DreamSim†	
Results on PIPAL [18]										
BAPPS [43]	✓	2AFC	0.622	0.634	0.489	0.287	0.151	0.338	0.760	0.469
PieAPP [31]	✓	2AFC	0.602	0.635	0.564	0.325	0.258	0.294	0.711	0.484
NIGHTS [13]	✓	2AFC	0.577	0.545	-0.294	0.234	0.277	0.234	0.662	0.319
KADID-10K [23]	✓	L1	0.577	0.590	0.518	0.298	0.312	0.253	0.617	0.452
FoMo (**Ours**)	✗	RankBCE	**0.733**	**0.683**	**0.615**	**0.699**	**0.644**	**0.632**	**0.776**	**0.683**
Results on TID2013 [30]										
BAPPS [43]	✓	2AFC	0.779	0.671	0.632	0.341	0.285	0.287	**0.813**	0.544
PieAPP [31]	✓	2AFC	0.761	0.722	0.680	0.315	0.302	0.467	0.767	0.573
NIGHTS [13]	✓	2AFC	0.779	0.653	-0.497	0.230	0.294	0.358	0.762	0.368
KADID-10K [23]	✓	L1	**0.793**	**0.757**	**0.834**	0.613	0.519	0.595	0.788	0.700
FoMo (**Ours**)	✗	RankBCE	0.785	0.663	0.691	**0.713**	**0.737**	**0.644**	0.801	**0.719**
Results on CSIQ [22]										
BAPPS [43]	✓	2AFC	0.943	0.887	0.855	0.411	0.427	0.455	**0.911**	0.699
PieAPP [31]	✓	2AFC	0.936	**0.896**	0.821	0.331	0.514	0.587	0.903	0.713
NIGHTS [13]	✓	2AFC	**0.945**	0.862	-0.560	0.335	0.364	0.427	0.867	0.463
KADID-10K [23]	✓	L1	0.935	0.895	**0.938**	0.616	0.604	0.698	0.874	0.794
FoMo (**Ours**)	✗	RankBCE	0.938	0.859	0.918	**0.811**	**0.901**	**0.795**	0.894	**0.874**
Results on LIVE [34]										
BAPPS [43]	✓	2AFC	**0.952**	**0.929**	0.843	0.688	0.425	0.298	0.927	0.723
PieAPP [31]	✓	2AFC	0.934	0.921	0.860	0.458	0.415	0.537	0.906	0.719
NIGHTS [13]	✓	2AFC	0.949	0.919	-0.774	0.439	0.560	0.376	0.867	0.477
KADID-10K [23]	✓	L1	0.942	0.912	0.950	0.765	0.827	0.797	0.896	0.870
FoMo (**Ours**)	✗	RankBCE	0.948	0.923	**0.954**	**0.896**	**0.925**	**0.907**	**0.931**	**0.926**

The main quantitative result of our approach, in comparison to existing dataset and objectives, is presented in Table 1. Across four benchmarks and seven backbone architectures, FoMo achieves the best overall performance. On PIPAL [18], the least saturated and the most important benchmark, FoMo outperforms the existing datasets and their objectives, in all seven backbones by a large margin, especially in Transformer-based backbones. On the other three benchmarks, TID2013 [30], CSIQ [22] and LIVE [34], FoMo performs comparable to baselines on CNN backbones, and outperforms most of them in Transformer backbones. These results reflect the effectiveness of our data generation pipeline and objective function, despite being the only approach that does not require human annotation. The entire numbers containing Kendall rank-order correlation coefficient (KROCC) and Pearson linear correlation coefficient (PLCC) on these benchmarks can be found in the Appendix.

5.3 Ablation study

In this section, we ablate and analyze the experimental choices in our pipeline. First, we discuss on the experimental analysis on the label and objective function used in training. Second, as our method is based on a RankNet-style binary cross-entropy loss which incorporates the entire ranking within a batch, computing a $B \times B$ comparison matrix, the batch size is expected to be an important

factor in our experiments, and we study on its effects. Third, we study to verify the generalizability of our approach, and check if the method could work with other diffusion models besides FLUX. Fourth, we study which timesteps are important in training, by dropping certain timestep ranges in training. Unless otherwise mentioned, we experiment on LPIPS-Alex and DINOv3 backbones, as representatives of each backbone styles, CNNs and Transformers. All the experiments in this section is evaluated with PIPAL benchmark, with SROCC scores.

Objective and label type We study on the choices on the objective function and the label types. In this experiment, we also experiment on KADID-10k dataset, for comprehensive analysis. First, our pipeline has two possible candidates of labels, using the interpolation factor t or the forking timestep s as the label, whereas KADID-10k labels are DMOS scores. The main difference with using t and step s as the label is resolution-variance. The main difference between using t and s as labels lies in their sensitivity to resolution changes, whether the label values are resolution-dependent. Refer to Sec. D of Appendix for further explanations. For each point-wise labels, we can apply three forms of objective function in training, ranked binary cross-entropy (RankBCE) loss, 2AFC-style pairwise binary cross-entropy loss, and a simple L1 regression loss on the label value. Our experimental results are presented in Table 2. According to our experiments, ranked binary cross-entropy loss as in RankNet [2] has clearly proved to be consistently superior than other objectives, supporting our claim on the importance of using rank-based objective. While direct DMOS regression is the default training convention [23, 9] for KADID-10k, our results show that a rank-based BCE objective is the stronger choice, and under this objective, FoMo still remains the better training source.

Table 2: Comparative experiment on the objective function and label type. Ranked binary cross-entropy loss (Rank) consistently shows stronger results, compared to triplet-based paired cross-entropy loss (2AFC) and L1 regression loss on the label (L1). Our label has shown to be more helpful in training, compared to the large-scale human annotated KADID-10k. Mean over five random seeds; full numbers with standard deviations can be found in Table 9.

	FoMo (Ours)						**KADID-10K [23]**		
Label	t			s			DMOS		
Objective	Rank	2AFC	L1	Rank	2AFC	L1	Rank	2AFC	L1
LPIPS-Alex [43]	**0.733**	<u>0.592</u>	0.440	**0.733**	<u>0.592</u>	0.390	**0.625**	<u>0.611</u>	0.577
DINOv3 [35]	**0.699**	0.622	<u>0.694</u>	**0.697**	0.614	<u>0.695</u>	**0.309**	0.145	<u>0.298</u>

Analysis on Batch Size We study on the effects of batch size in our ranked binary cross-entropy loss. Pairwise ranking losses that operate over all within-batch pairs are known to benefit from large batch sizes, since a larger batch means more comparisons and thus richer supervision per training step, a property well-documented in contrastive learning [4, 32]. As shown in Table 3, we observe a similar trend. Performance improves with batch size increases, especially in Transformer-based DINOv3 backbone, whereas in CNN-based backbone LPIPS-Alex, the performance peaks at batch size of 64. This reflects the potential that our approach could be scaled more with larger batch size and larger models.

Table 3: Effect of batch size on training performance. The Transformer-based backbone improves monotonically with batch size, while the CNN-based backbone saturates at 64 and degrades slightly beyond it. Mean $\pm$ standard deviation over five random seeds.

Backbone	**Batch size**				
	16	32	64	128	256
LPIPS-Alex [43]	0.719 ± 0.006	0.729 ± 0.006	$\mathbf{0.733 \pm 0.006}$	0.728 ± 0.006	0.712 ± 0.006
DINOv3 [35]	0.479 ± 0.080	0.650 ± 0.018	0.696 ± 0.007	0.705 ± 0.003	$\mathbf{0.707 \pm 0.005}$

Cross Model Validation To study the generalizability of our method, we adopt diverse diffusion models into our pipeline in replacement to FLUX, which we used in our main experiments. For this experiment, we use three pre-trained diffusion models publicly available, *i.e.*, Stable Diffusion 1.5 (SD 1.5) [33], SD-XL [29], and SD3 [11]. Using these pre-trained models, we generate image pair

sets in the same manner. For this experiment we build a pool of 50k pairs per generator and draw five disjoint 10k subsets from it, training on each subset with the training seed held fixed. Every generator is thus matched at 10k training pairs, roughly 2% of our main training set. The experimental results can be found in Table 4. Across generators the ordering of the two backbones is preserved, and every generator yields a working metric. Even at that reduced budget, training on any of the four surpasses the best human-annotated dataset for the CNN backbone (0.622, Table 1), and with the Transformer backbone every generator except SD-XL at least matches its best human-annotated baseline (0.508). The approach therefore does not depend on FLUX specifically, but the choice of generator is not immaterial either. FLUX.1 is the strongest option for both backbones, by margins well outside the data-sampling error bars, and the Transformer-based models benefit the most from the strong generator.

Table 4: Experimental results using diverse diffusion models for data generation. Our pipeline is not specific to FLUX, and every generator yields a working metric, although FLUX.1 is the strongest choice. Cells report mean $\pm$ standard deviation over five train runs, each using disjoint 10k subsets from the generator.

Backbone	**Data Generator** (10k pairs)			
	SD-1.5 [33]	SD-XL [29]	SD-3 [11]	FLUX.1 [20]
LPIPS-Alex [43]	0.687 ± 0.008	0.662 ± 0.008	0.695 ± 0.007	0.741 ± 0.002
DINOv3 [35]	0.574 ± 0.078	0.418 ± 0.044	0.517 ± 0.017	0.672 ± 0.016

Timestep Range Selection We study on how the performance of models change depending on the sampling range of forking timesteps. In this experiment, we use the three CNN backbones, LPIPS-Alex, VGG and DISTS models, along with a Transformer backbone, DINOv3. The results are displayed in Table 5. Within a 50-step process, the initial 35 steps, which are closer to the noise state than the clean image state, has shown to be the most important. The final 15 steps of the generation process are dedicated to refining fine-grained details that are barely perceptible to the human eye. Since such subtle refinements carry little perceptual significance, incorporating these steps into training may introduce noise into the learning signal, and the two LPIPS backbones indeed peak without them. DISTS and DINOv3, however, perform best over the full range. Considering the overall robustness, in our main experiments, we used the full 50 steps in training.

Table 5: Experiment on different sampling ranges of forking timesteps. At a matched window width the earlier window is the most useful one, and the full schedule overall provides the most stable result. Mean $\pm$ standard deviation over five random seeds.

Backbone	**All**	**Sampling Range of Forking Timesteps**				
	$[0, 50]$	$[0, 25]$	$[0, 35]$	$[13, 37]$	$[15, 50]$	$[25, 50]$
LPIPS-Alex [43]	0.733 ± 0.006	$\underline{0.748 \pm 0.008}$	$\mathbf{0.757 \pm 0.007}$	0.729 ± 0.008	0.731 ± 0.007	0.720 ± 0.006
LPIPS-VGG [43]	0.683 ± 0.008	$\underline{0.688 \pm 0.005}$	$\mathbf{0.697 \pm 0.006}$	0.685 ± 0.004	0.666 ± 0.006	0.626 ± 0.005
DISTS [9]	$\mathbf{0.615 \pm 0.001}$	0.464 ± 0.002	$\underline{0.583 \pm 0.001}$	0.569 ± 0.002	0.556 ± 0.001	0.485 ± 0.003
DINOv3 [35]	$\mathbf{0.703 \pm 0.006}$	0.620 ± 0.015	0.696 ± 0.015	0.700 ± 0.009	$\underline{0.702 \pm 0.015}$	0.680 ± 0.014

6 Related Work

6.1 Reference-based Image Quality Assessment

A reliable perceptual metric must produce a globally consistent quality ordering, one that mirrors how humans rank distortions across the full image set, not just on isolated pairs. This requirement has driven the evolution of IQA annotation methodology. Early benchmark datasets such as LIVE [34], CSIQ [22], TID2013 [30] and KADID-10k [23] collected Mean Opinion Scores (MOS) as ground truth, implicitly assuming that averaged absolute ratings constitute a reliable global quality scale. PieAPP [31] challenged this directly, arguing that MOS-based annotations are unreliable because absolute quality ratings are inconsistent across raters and sessions: observers apply different scale calibrations and anchor their ratings differently, making cross-image comparisons ambiguous. This motivated a shift toward 2-alternative-forced-choice (2AFC) annotation, where raters compare two

distortions relative to a reference rather than score in isolation. Concurrent to this finding, LPIPS [43] also adopts the 2AFC paradigm for labeling and constructs a large-scale dataset gathered from human annotations. PIPAL [18] extends it to GAN-based restorations via Elo-ranked pairwise judgments to expand the coverage of IQA evaluations. DreamSim [13] also collects a new dataset based on the 2AFC paradigm, but its main focus is on mid-level similarity rather than low-level similarity.

Despite this recent popularity of 2AFC-style labeling and training, the 2AFC setting carries its own structural weakness. Talebi et al. [38] point out that mini-batch pairwise optimization never explicitly sees the global ranking of images, and each gradient step accounts for only a small fraction of all possible comparisons; they demonstrate that regularizing with rank-centrality aggregation consistently improves human preference prediction. dipIQ [27] reinforces this directly by comparing pairwise RankNet and listwise ListNet training objectives on the same automatically generated quality-discriminable pairs, finding that the listwise variant consistently outperforms its pairwise counterpart, a clear empirical signal that optimizing global ordinal structure is superior to aggregating independent pairwise decisions. Yet all of these methods remain bottlenecked by human annotation cost and coverage.

6.2 Diffusion Models as Perceptual Signal

There have been several efforts in exploiting diffusion models for perceptual similarity. DIFT [39] and Diffusion Hyperfeatures [26] employ diffusion features for structural similarity, but their methods are targeted for geometric correspondence task, rather than low-level perceptual distance. DiffSim [37] uses attention layer features in Stable Diffusion [33] to measure visual similarity, but targets style and instance-level consistency in generative customization settings rather than low-level perceptual fidelity. AnoDDPM [42] partially diffuses an image to an intermediate timestep and measures the pixel-level reconstruction divergence after denoising as an anomaly score, but restricted to detecting distributional outliers in medical images. DifFIQA [1] applies perturbation-robustness under DDPM noising as a face image quality signal, scoring faces by the shift in identity embedding between input and reconstructed image, an NR-IQA approach specific to facial content.

Our approach differs from all of the above on two axes. First, unlike feature-based approaches, our approach does not use diffusion model representations at inference time; the diffusion model is used only for generating the data samples and their labels. Second, unlike scoring approaches like AnoDDPM and DifFIQA, our approach uses the denosing generative process as a label generation mechanism, instead of using the generation result in distance computation.

7 Conclusion

In this paper, we introduced a data generation pipeline that reframes perceptual distance as a forking moment in a diffusion denoising trajectory. By forward-diffusing a reference image to a sampled timestep and denoising it back, we automatically synthesize training pairs with calibrated perceptual distances, no human annotation required. Our human study verified that using forking moments in diffusion trajectory aligns well with human perception, supporting our argument. We further show that RankNet-style supervision over our generated data substantially outperforms 2AFC-style binary classification, yielding richer gradient signal and implicit transitivity enforcement. Extensive experiments on diverse backbones and evaluation benchmarks demonstrate consistent improvements over strong baselines, validating both the dataset generation pipeline and the ranking-based objective.

Acknowledgments and Disclosure of Funding

This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2026; the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (Nos. 2022R1A3B1077720 and 2022R1A5A7083908); and the Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program, Seoul National University). The authors also thank Yongsung Kim for valuable discussions and support.

References

[1] Žiga Babnik, Peter Peer, and Vitomir Štruc. Diffiqa: Face image quality assessment using denoising diffusion probabilistic models. In *2023 IEEE international joint conference on biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[3] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[5] Zewen Chen, Juan Wang, Bing Li, Chunfeng Yuan, Weiming Hu, Junxian Liu, Peng Li, Yan Wang, Youqun Zhang, and Congxuan Zhang. Gmc-iqa: Exploiting global-correlation and mean-opinion consistency for no-reference image quality assessment. *arXiv preprint arXiv:2401.10511*, 2024.

[6] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11472–11481, 2022.

[7] Dale Decatur, Thibault Groueix, Wang Yifan, Rana Hanocka, Vladimir Kim, and Matheus Gadelha. Reusing computation in text-to-image diffusion for efficient generation of image sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16482–16491, 2025.

[8] 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*, pages 248–255. Ieee, 2009.

[9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

[10] 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. In *International Conference on Learning Representations*.

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

[12] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378, 1971.

[13] 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. *Advances in Neural Information Processing Systems*, 36:50742–50768, 2023.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

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

[16] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[18] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European conference on computer vision*, pages 633–651. Springer, 2020.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[20] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[21] GG Landis JRKoch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159174, 1977.

[22] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010.

[23] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

[25] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*.

[26] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

[27] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on image processing*, 26(8): 3951–3964, 2017.

[28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.

[29] 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 *The Twelfth International Conference on Learning Representations*.

[30] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

[31] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018.

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

[33] 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*, pages 10684–10695, 2022.

[34] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

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

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

[37] Yiren Song, Xiaokang Liu, and Mike Zheng Shou. Diffsim: Taming diffusion models for evaluating visual similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16904–16915, 2025.

[38] Hossein Talebi, Ehsan Amid, Peyman Milanfar, and Manfred K Warmuth. Rank-smoothed pairwise learning in perceptual quality assessment. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3413–3417. IEEE, 2020.

[39] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36:1363–1389, 2023.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[42] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 650–656, 2022.

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

A Experimental Details

Model Training The prediction heads of LPIPS-Alex/VGG are linear layers from each level of the feature pyramid, and DISTS learn the α and β in integration of the extracted intermediate features. For three Transformer backbones, DINOv3, CLIP and MAE, the feature extractor is also fixed, and we append three ViT layers [10] for distance logit prediction. The prediction head layers take the feature tokens extracted from the backbone as the input, along with a `[CLS]` token which make the prediction of the distance logit. For all experiments besides DreamSim, we use a batch size of 64, and a fixed learning rate of 2e-4 and 1e-4 respectively for CNN backbones and the three Transformer backbones. For DreamSim, we follow its original configuration without any change, from LoRA [17] configuration (r=16, α=1, dropout 0.3 on the qkv projections), input protocol and optimizer, and only the training set is varied; the sample budget matches every other column.

Ensuring symmetry in Transformer backbone models We write $\hat{d}(u, v)$ for the distance the model assigns to an image pair (u, v); this is the quantity written $\hat{d}_i$ in Sec. 4.2. For the Transformer backbones, DINOv3, CLIP and MAE, it is produced by the prediction head, whose output needs an adjustment for symmetry. The head reads the two images as a single concatenated sequence and is thus asymmetric, $\hat{d}(u, v)$ and $\hat{d}(v, u)$ returning non-identical values. Therefore we report the results acquired from $\frac{1}{2}\big[\hat{d}(u, v) + \hat{d}(v, u)\big]$, which removes the asymmetry. During training, each training pair is presented in a random order to make the model work in both orders, and be naturally symmetric. This way, we observe the distance logits to be similar in both input orders, although inherently they cannot be perfectly identical. Neither adjustment applies to LPIPS, DISTS or DreamSim, which compute a symmetric distance directly and are reported as they are.

B Full Results

Table 6: Comparison of training datasets and their objectives on four reference-based IQA benchmarks, across CNN- and Transformer-based backbones. Mean ± standard deviation of SROCC over five random seeds; full-statistics version of Table 1. †Evaluated under 224 resolution.

Train Data	CNN-based			Transformer-based				Avg.
	LPIPS-Alex	LPIPS-VGG	DISTS	DINOv3	CLIP	MAE	DreamSim†	
Results on PIPAL [18]								
BAPPS [43]	0.622 ± 0.005	0.634 ± 0.018	0.489 ± 0.003	0.287 ± 0.020	0.151 ± 0.040	0.338 ± 0.013	0.760 ± 0.007	0.469 ± 0.008
PieAPP [31]	0.602 ± 0.008	0.635 ± 0.016	0.564 ± 0.003	0.325 ± 0.017	0.258 ± 0.037	0.294 ± 0.023	0.711 ± 0.007	0.484 ± 0.007
NIGHTS [13]	0.577 ± 0.006	0.545 ± 0.011	-0.294 ± 0.003	0.234 ± 0.024	0.277 ± 0.025	0.234 ± 0.029	0.662 ± 0.019	0.319 ± 0.005
KADID-10K [23]	0.577 ± 0.025	0.590 ± 0.034	0.518 ± 0.002	0.298 ± 0.010	0.312 ± 0.018	0.253 ± 0.041	0.617 ± 0.003	0.452 ± 0.006
FoMo (**Ours**)	**0.733 ± 0.006**	**0.683 ± 0.008**	**0.615 ± 0.001**	**0.699 ± 0.006**	**0.644 ± 0.024**	**0.632 ± 0.020**	**0.776 ± 0.010**	**0.683 ± 0.004**
Results on TID2013 [30]								
BAPPS [43]	0.779 ± 0.003	0.671 ± 0.002	0.632 ± 0.001	0.341 ± 0.022	0.285 ± 0.030	0.287 ± 0.022	**0.813 ± 0.004**	0.544 ± 0.006
PieAPP [31]	0.761 ± 0.006	0.722 ± 0.005	0.680 ± 0.004	0.315 ± 0.042	0.302 ± 0.025	0.467 ± 0.027	0.767 ± 0.004	0.573 ± 0.008
NIGHTS [13]	0.779 ± 0.003	0.653 ± 0.001	-0.497 ± 0.005	0.230 ± 0.022	0.294 ± 0.035	0.358 ± 0.033	0.762 ± 0.007	0.368 ± 0.007
KADID-10K [23]	**0.793 ± 0.004**	**0.757 ± 0.012**	**0.834 ± 0.001**	0.613 ± 0.024	0.519 ± 0.016	0.595 ± 0.013	0.788 ± 0.004	0.700 ± 0.006
FoMo (**Ours**)	0.785 ± 0.003	0.663 ± 0.003	0.691 ± 0.001	**0.713 ± 0.003**	**0.737 ± 0.016**	**0.644 ± 0.052**	0.801 ± 0.003	**0.719 ± 0.007**
Results on CSIQ [22]								
BAPPS [43]	0.943 ± 0.001	0.887 ± 0.001	0.855 ± 0.001	0.411 ± 0.027	0.427 ± 0.015	0.455 ± 0.051	**0.911 ± 0.000**	0.699 ± 0.007
PieAPP [31]	0.936 ± 0.002	**0.896 ± 0.007**	0.821 ± 0.003	0.331 ± 0.042	0.514 ± 0.016	0.587 ± 0.012	0.903 ± 0.003	0.713 ± 0.005
NIGHTS [13]	**0.945 ± 0.001**	0.862 ± 0.002	-0.560 ± 0.006	0.335 ± 0.031	0.364 ± 0.022	0.427 ± 0.033	0.867 ± 0.007	0.463 ± 0.006
KADID-10K [23]	0.935 ± 0.002	0.895 ± 0.016	**0.938 ± 0.000**	0.616 ± 0.009	0.604 ± 0.026	0.698 ± 0.035	0.874 ± 0.009	0.794 ± 0.010
FoMo (**Ours**)	0.938 ± 0.001	0.859 ± 0.006	0.918 ± 0.001	**0.811 ± 0.002**	**0.901 ± 0.014**	**0.795 ± 0.038**	0.894 ± 0.002	**0.874 ± 0.005**
Results on LIVE [34]								
BAPPS [43]	**0.952 ± 0.001**	**0.929 ± 0.002**	0.843 ± 0.001	0.688 ± 0.032	0.425 ± 0.059	0.298 ± 0.062	0.927 ± 0.001	0.723 ± 0.013
PieAPP [31]	0.934 ± 0.013	0.921 ± 0.013	0.860 ± 0.002	0.458 ± 0.022	0.415 ± 0.030	0.537 ± 0.016	0.906 ± 0.004	0.719 ± 0.006
NIGHTS [13]	0.949 ± 0.002	0.919 ± 0.002	-0.774 ± 0.004	0.439 ± 0.021	0.560 ± 0.057	0.376 ± 0.058	0.867 ± 0.007	0.477 ± 0.015
KADID-10K [23]	0.942 ± 0.013	0.912 ± 0.032	0.950 ± 0.000	0.765 ± 0.023	0.827 ± 0.009	0.797 ± 0.015	0.896 ± 0.012	0.870 ± 0.012
FoMo (**Ours**)	0.948 ± 0.003	0.923 ± 0.004	**0.954 ± 0.000**	**0.896 ± 0.002**	**0.925 ± 0.011**	**0.907 ± 0.026**	**0.931 ± 0.002**	**0.926 ± 0.003**

C Human Alignment Experiment

We present the web user-interface used for single-reference human study from the experiment of Sec. 3.1 are in Fig. 4 and 5, and additional examples of experiment from Sec. 3.2 are in Fig. 6 and 7.

Table 7: Comparison of training datasets and their objectives on four reference-based IQA benchmarks, across CNN- and Transformer-based backbones. Mean ± standard deviation of KROCC over five random seeds. †Evaluated under 224 resolution.

Train Data	CNN-based			Transformer-based				Avg.
	LPIPS-Alex	LPIPS-VGG	DISTS	DINOv3	CLIP	MAE	DreamSim†	
Results on PIPAL [18]								
BAPPS [43]	0.440 ± 0.004	0.455 ± 0.014	0.337 ± 0.002	0.194 ± 0.014	0.101 ± 0.027	0.229 ± 0.009	0.564 ± 0.007	0.332 ± 0.006
PieAPP [31]	0.420 ± 0.007	0.450 ± 0.013	0.389 ± 0.002	0.222 ± 0.013	0.175 ± 0.026	0.197 ± 0.016	0.511 ± 0.006	0.338 ± 0.005
NIGHTS [13]	0.405 ± 0.005	0.385 ± 0.008	-0.204 ± 0.002	0.159 ± 0.017	0.186 ± 0.017	0.157 ± 0.020	0.471 ± 0.018	0.223 ± 0.004
KADID-10K [23]	0.403 ± 0.020	0.419 ± 0.027	0.360 ± 0.001	0.204 ± 0.007	0.212 ± 0.012	0.170 ± 0.028	0.440 ± 0.003	0.315 ± 0.005
FoMo (**Ours**)	**0.536 ± 0.006**	**0.497 ± 0.007**	**0.435 ± 0.001**	**0.500 ± 0.005**	**0.459 ± 0.019**	**0.444 ± 0.015**	**0.571 ± 0.010**	**0.492 ± 0.003**
Results on TID2013 [30]								
BAPPS [43]	0.584 ± 0.003	0.497 ± 0.002	0.450 ± 0.001	0.235 ± 0.017	0.194 ± 0.021	0.192 ± 0.015	**0.619 ± 0.003**	0.396 ± 0.004
PieAPP [31]	0.567 ± 0.006	0.540 ± 0.004	0.486 ± 0.004	0.217 ± 0.030	0.207 ± 0.017	0.321 ± 0.019	0.574 ± 0.003	0.416 ± 0.006
NIGHTS [13]	0.582 ± 0.003	0.480 ± 0.001	-0.345 ± 0.004	0.157 ± 0.016	0.199 ± 0.024	0.243 ± 0.023	0.566 ± 0.006	0.269 ± 0.005
KADID-10K [23]	**0.596 ± 0.005**	**0.567 ± 0.011**	**0.640 ± 0.001**	0.448 ± 0.021	0.368 ± 0.012	0.426 ± 0.009	0.598 ± 0.005	0.520 ± 0.005
FoMo (**Ours**)	0.586 ± 0.003	0.489 ± 0.003	0.512 ± 0.001	**0.526 ± 0.002**	**0.547 ± 0.016**	**0.473 ± 0.043**	0.605 ± 0.004	**0.534 ± 0.006**
Results on CSIQ [22]								
BAPPS [43]	0.786 ± 0.002	0.701 ± 0.003	0.654 ± 0.001	0.287 ± 0.020	0.298 ± 0.011	0.315 ± 0.036	**0.741 ± 0.001**	0.540 ± 0.005
PieAPP [31]	0.773 ± 0.004	**0.713 ± 0.009**	0.611 ± 0.004	0.229 ± 0.029	0.361 ± 0.012	0.405 ± 0.009	0.724 ± 0.005	0.545 ± 0.003
NIGHTS [13]	**0.790 ± 0.002**	0.664 ± 0.001	-0.384 ± 0.005	0.230 ± 0.021	0.244 ± 0.016	0.291 ± 0.024	0.678 ± 0.009	0.359 ± 0.004
KADID-10K [23]	0.768 ± 0.005	0.705 ± 0.020	**0.776 ± 0.001**	0.447 ± 0.007	0.430 ± 0.021	0.510 ± 0.029	0.682 ± 0.012	0.617 ± 0.010
FoMo (**Ours**)	0.781 ± 0.002	0.672 ± 0.006	0.750 ± 0.001	**0.626 ± 0.002**	**0.723 ± 0.019**	**0.609 ± 0.030**	0.709 ± 0.003	**0.696 ± 0.004**
Results on LIVE [34]								
BAPPS [43]	**0.800 ± 0.001**	**0.757 ± 0.005**	0.644 ± 0.001	0.508 ± 0.029	0.298 ± 0.043	0.203 ± 0.041	0.766 ± 0.002	0.568 ± 0.008
PieAPP [31]	0.769 ± 0.018	0.746 ± 0.022	0.649 ± 0.003	0.324 ± 0.018	0.288 ± 0.023	0.370 ± 0.015	0.728 ± 0.005	0.554 ± 0.006
NIGHTS [13]	0.797 ± 0.003	0.744 ± 0.004	-0.574 ± 0.003	0.307 ± 0.016	0.392 ± 0.045	0.256 ± 0.041	0.679 ± 0.009	0.372 ± 0.012
KADID-10K [23]	0.789 ± 0.017	0.734 ± 0.044	0.801 ± 0.001	0.580 ± 0.021	0.627 ± 0.011	0.595 ± 0.015	0.713 ± 0.016	0.691 ± 0.014
FoMo (**Ours**)	0.791 ± 0.007	0.748 ± 0.007	**0.804 ± 0.001**	**0.720 ± 0.003**	**0.754 ± 0.017**	**0.734 ± 0.038**	**0.769 ± 0.004**	**0.760 ± 0.005**

Table 8: Comparison of training datasets and their objectives on four reference-based IQA benchmarks, across CNN- and Transformer-based backbones. Mean ± standard deviation of PLCC over five random seeds. †Evaluated under 224 resolution.

Train Data	CNN-based			Transformer-based				Avg.
	LPIPS-Alex	LPIPS-VGG	DISTS	DINOv3	CLIP	MAE	DreamSim†	
Results on PIPAL [18]								
BAPPS [43]	0.659 ± 0.005	0.676 ± 0.014	0.495 ± 0.002	0.314 ± 0.015	0.186 ± 0.043	0.348 ± 0.018	**0.779 ± 0.007**	0.494 ± 0.010
PieAPP [31]	0.607 ± 0.010	0.637 ± 0.014	0.558 ± 0.003	0.364 ± 0.014	0.284 ± 0.029	0.338 ± 0.025	0.707 ± 0.005	0.499 ± 0.003
NIGHTS [13]	0.634 ± 0.006	0.605 ± 0.007	0.361 ± 0.002	0.296 ± 0.028	0.324 ± 0.030	0.259 ± 0.037	0.669 ± 0.022	0.450 ± 0.008
KADID-10K [23]	0.599 ± 0.022	0.623 ± 0.034	0.562 ± 0.001	0.351 ± 0.021	0.346 ± 0.019	0.274 ± 0.034	0.619 ± 0.007	0.482 ± 0.008
FoMo (**Ours**)	**0.768 ± 0.005**	**0.732 ± 0.008**	**0.656 ± 0.001**	**0.706 ± 0.006**	**0.652 ± 0.023**	**0.625 ± 0.016**	0.764 ± 0.010	**0.700 ± 0.003**
Results on TID2013 [30]								
BAPPS [43]	0.814 ± 0.003	0.755 ± 0.002	0.664 ± 0.001	0.488 ± 0.028	0.343 ± 0.037	0.330 ± 0.030	**0.850 ± 0.002**	0.606 ± 0.012
PieAPP [31]	0.798 ± 0.006	0.777 ± 0.005	0.743 ± 0.003	0.405 ± 0.024	0.344 ± 0.020	0.513 ± 0.030	0.808 ± 0.008	0.627 ± 0.006
NIGHTS [13]	0.805 ± 0.004	0.733 ± 0.003	0.686 ± 0.003	0.391 ± 0.036	0.386 ± 0.056	0.434 ± 0.045	0.791 ± 0.005	0.604 ± 0.011
KADID-10K [23]	0.815 ± 0.007	**0.785 ± 0.014**	**0.845 ± 0.001**	0.694 ± 0.019	0.608 ± 0.014	0.643 ± 0.027	0.825 ± 0.006	0.745 ± 0.006
FoMo (**Ours**)	**0.817 ± 0.003**	0.753 ± 0.005	0.770 ± 0.001	**0.768 ± 0.002**	**0.788 ± 0.018**	**0.657 ± 0.046**	0.831 ± 0.003	**0.769 ± 0.006**
Results on CSIQ [22]								
BAPPS [43]	0.945 ± 0.001	**0.908 ± 0.003**	0.850 ± 0.001	0.561 ± 0.036	0.483 ± 0.024	0.468 ± 0.051	**0.932 ± 0.001**	0.735 ± 0.007
PieAPP [31]	0.934 ± 0.004	0.907 ± 0.008	0.862 ± 0.002	0.437 ± 0.044	0.594 ± 0.015	0.594 ± 0.011	0.920 ± 0.004	0.750 ± 0.006
NIGHTS [13]	**0.945 ± 0.002**	0.877 ± 0.002	0.769 ± 0.003	0.450 ± 0.060	0.399 ± 0.033	0.455 ± 0.042	0.878 ± 0.008	0.682 ± 0.012
KADID-10K [23]	0.932 ± 0.004	0.889 ± 0.027	**0.934 ± 0.000**	0.730 ± 0.018	0.727 ± 0.029	0.735 ± 0.034	0.891 ± 0.007	0.834 ± 0.011
FoMo (**Ours**)	0.945 ± 0.001	0.888 ± 0.006	0.931 ± 0.001	**0.874 ± 0.002**	**0.923 ± 0.011**	**0.816 ± 0.030**	0.905 ± 0.003	**0.897 ± 0.004**
Results on LIVE [34]								
BAPPS [43]	**0.946 ± 0.001**	**0.929 ± 0.003**	0.839 ± 0.000	0.760 ± 0.031	0.462 ± 0.054	0.366 ± 0.035	0.935 ± 0.001	0.748 ± 0.011
PieAPP [31]	0.923 ± 0.016	0.916 ± 0.016	0.870 ± 0.002	0.493 ± 0.023	0.469 ± 0.047	0.550 ± 0.021	0.912 ± 0.002	0.733 ± 0.011
NIGHTS [13]	0.946 ± 0.002	0.922 ± 0.003	0.753 ± 0.004	0.564 ± 0.020	0.586 ± 0.062	0.402 ± 0.055	0.883 ± 0.006	0.722 ± 0.015
KADID-10K [23]	0.935 ± 0.018	0.903 ± 0.036	0.947 ± 0.000	0.816 ± 0.016	0.853 ± 0.008	0.809 ± 0.013	0.906 ± 0.010	0.881 ± 0.011
FoMo (**Ours**)	0.941 ± 0.004	0.923 ± 0.004	**0.950 ± 0.001**	**0.913 ± 0.001**	**0.928 ± 0.011**	**0.913 ± 0.027**	**0.935 ± 0.002**	**0.929 ± 0.004**

Table 9: Comparative experiment on the objective function and label type. Ranked binary cross-entropy loss (Rank) consistently shows stronger results, compared to triplet-based paired cross-entropy loss (2AFC) and L1 regression loss on the label. Our label has shown to be more helpful in training, compared to the large-scale human annotated KADID-10k [23]. PIPAL SROCC, mean ± standard deviation over five random seeds.

	FoMo (Ours)						KADID-10K [23]		
Label	t			s			DMOS		
Objective	Rank	2AFC	L1	Rank	2AFC	L1	Rank	2AFC	L1
LPIPS-Alex [43]	0.733 ± 0.006	0.592 ± 0.002	0.440 ± 0.059	0.733 ± 0.006	0.592 ± 0.002	0.390 ± 0.049	0.625 ± 0.004	0.611 ± 0.003	0.577 ± 0.025
DINOv3 [35]	0.699 ± 0.006	0.622 ± 0.020	0.694 ± 0.008	0.697 ± 0.006	0.614 ± 0.009	0.695 ± 0.013	0.309 ± 0.030	0.145 ± 0.010	0.298 ± 0.010



Figure 4: Screenshot of the human annotation website for experiments in Sec. 3

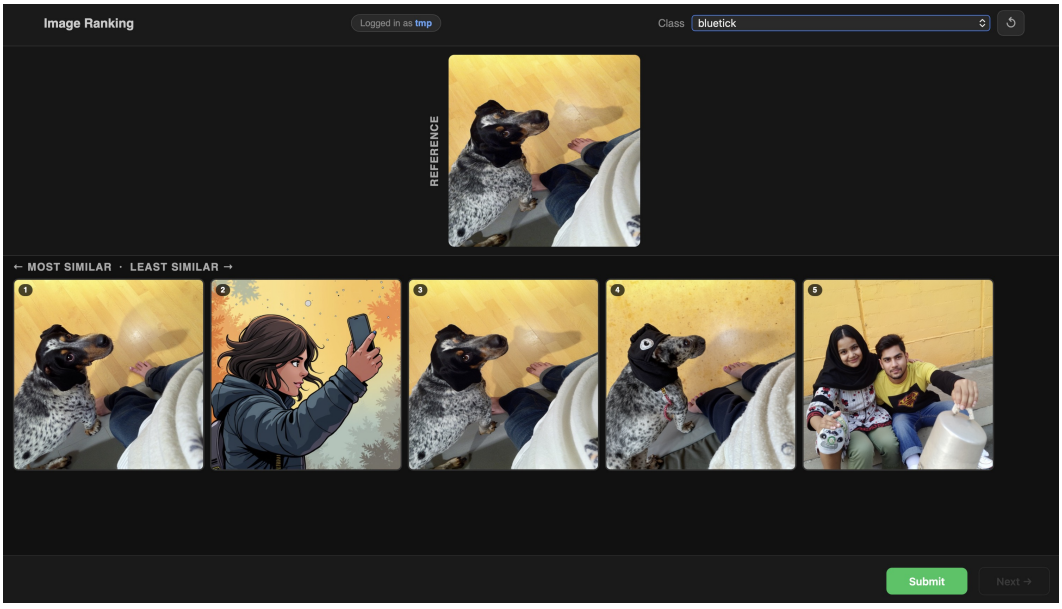


Figure 5: Screenshot of the human annotation website for experiments in Sec. 3

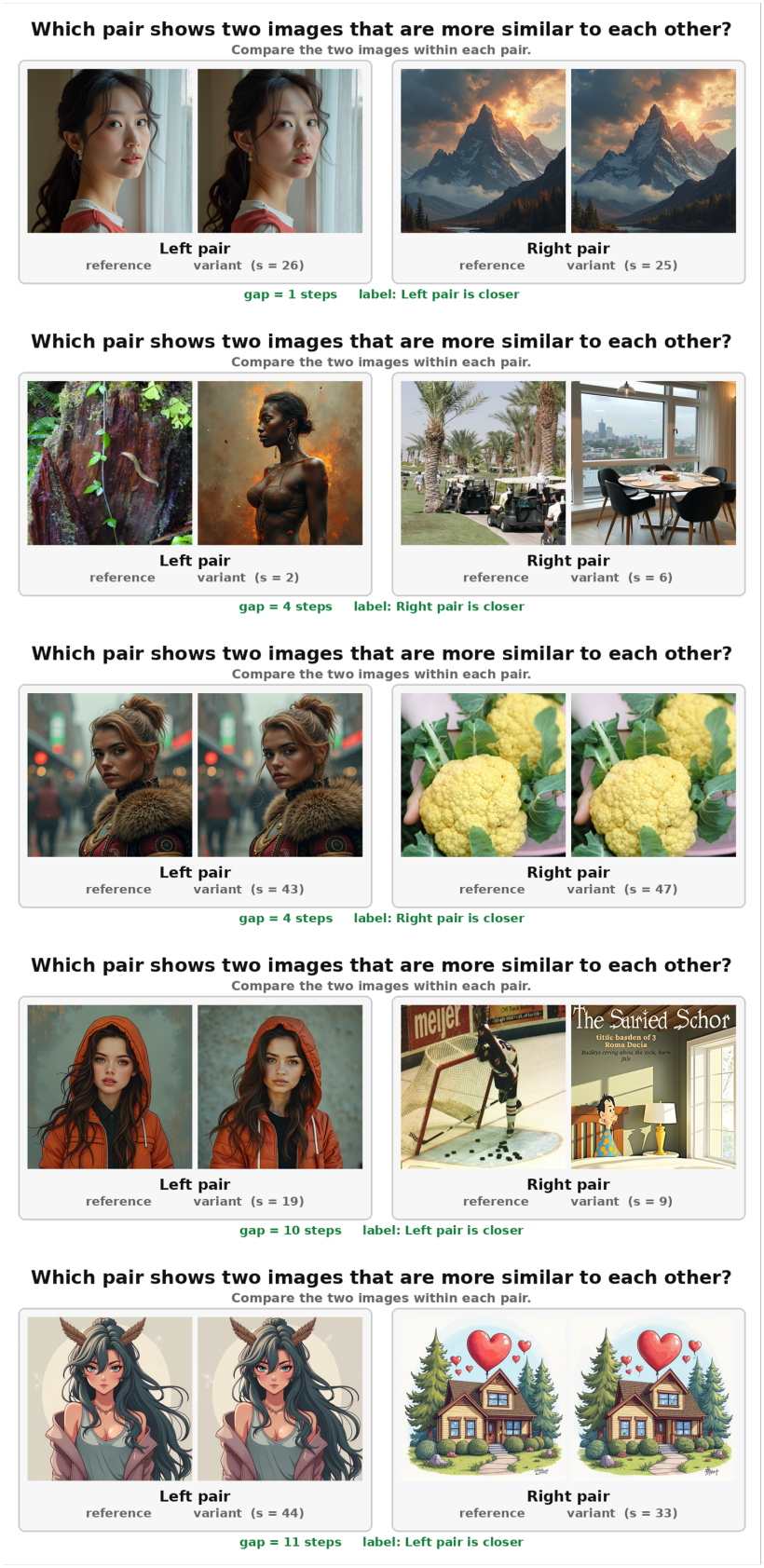


Figure 6: Examples from cross-reference human study, with each samples' forked moments, the gap of forking moments and labels from FoMo provided.

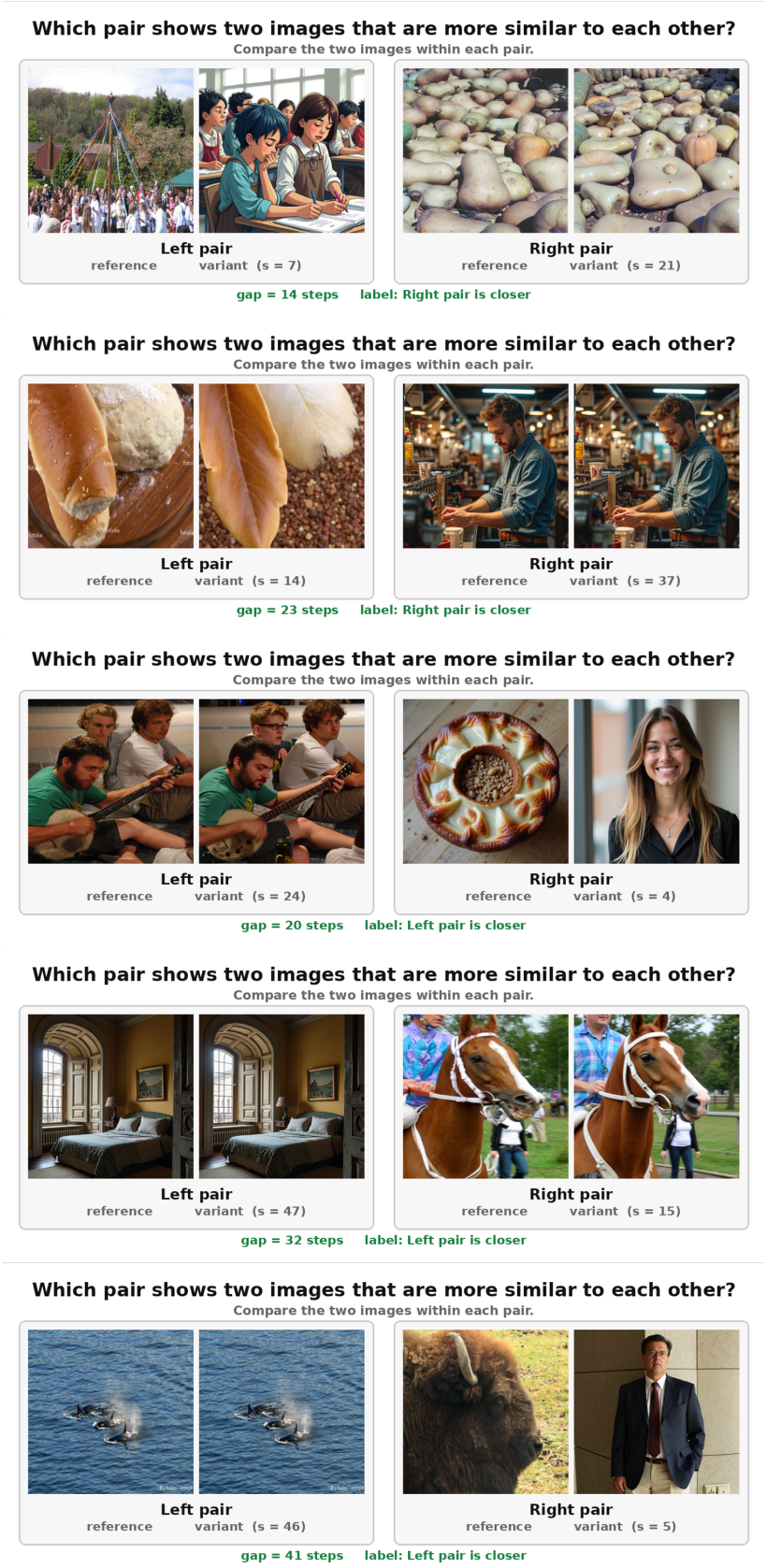


Figure 7: Examples from cross-reference human study, with each samples' forked moments, the gap of forking moments and labels from FoMo provided.

D Data Augmentation

One of a key advantage of our approach is its compatibility with a substantially broader range of data augmentation strategies than human-annotated alternatives permit. Pairwise preference datasets such as BAPPS [43] and NIGHTS [13] impose strict constraints on applicable augmentations: while label-preserving transformations such as random horizontal flips and in-plane rotations can be safely applied without altering perceptual judgments, aggressive spatial augmentations, most notably random resize cropping, are not applicable. Annotations in these datasets reflect global perceptual preferences elicited at a fixed resolution over entire image triplets; a crop that exposes only a local region may induce a different perceptual ordering and thereby contradict the original annotation.

Our approach enjoys better flexibility because the forking moment label is derived from a predefined noise schedule and can therefore be recomputed analytically under any spatial transformation. We describe how two canonical augmentations are accommodated within our framework. Both derivations rely on the established result that the diffusion noise schedule must be rescaled with image resolution [16, 3]. Concretely, the log signal-to-noise ratio (log-SNR) of the variance-preserving forward process shifts with resolution as

$$\log \text{SNR}(t; r) = \log \text{SNR}(t; r_0) + 2 \log(\frac{r_0}{r}), \tag{4}$$

where r denotes the image resolution (*e.g.*, the shorter spatial dimension in pixels), r_0 is a reference resolution at which the baseline schedule is defined, and $t \in [0, 1]$ is the continuous interpolation factor. Given the variance-preserving constraint $a_t^2 + b_t^2 = 1$, the forward-process noise coefficients at resolution r are recovered as

$$a_t^2(r) = \sigma(\log \text{SNR}(t; r)), b_t^2(r) = 1 - a_t^2(r), \tag{5}$$

where $\sigma(\cdot)$ denote the sigmoid operator.. We refer to the log-SNR value at the forking moment as $\lambda_s = \log \text{SNR}(t_0;\ r_0)$, which serves as a resolution-invariant proxy for perceptual divergence in both cases below.

Resizing Rescaling an image from r_0 to a new resolution r_{resize} preserves the global scene content and structure. The relative noise level at which content is destroyed therefore remains unchanged, and so does the perceptual divergence between the reference and distorted images. Consequently, the forking timestep s is retained as the label after resizing. However, because the noise schedule is resolution-dependent (Eq. 4), the continuous interpolation factor t corresponding to s shifts implicitly. If s maps to interpolation factor t_0 under the original schedule, then at resolution r_{resize} the factor t_0 must be updated to t_{resize} such that

$$\log \text{SNR}(t_{\text{resize}}; r_{\text{resize}}) = \log \text{SNR}(t_0, r_0), \tag{6}$$

which amounts to a shift of t_0 by $2 \log(r_0/r_{\text{resize}})$ in log-SNR space.

Random Resize Cropping Cropping simultaneously alters the spatial resolution and the visible image content. Since the cropped region depicts only a portion of the original scene, the forking timestep s can no longer be assumed invariant. However, the continuous interpolation factor t, which encodes the relative signal-to-noise level at which the two images perceptually diverge, is a resolution-agnostic quantity and is preserved across the crop. This follows directly from the pixel-wise nature of the diffusion forward process: a noised image at step s is formed as $x_s = a_{t_0} x_0 + b_{t_0} \epsilon$, where the mixing ratio t_0 is applied uniformly across all spatial locations. Any crop of x_s is therefore a crop of the same mixture at factor t_0, irrespective of the global image extent or resolution.

The label for the cropped region at resolution r_{crop} is recomputed as follows. The interpolation factor t_0 is retained from the original annotation, and λ_s is computed as above. The noise schedule is then rescaled to resolution r_{crop} via Eq. 4, yielding a new log-SNR curve. Finally, the forking timestep s_{crop} is recovered by identifying the discrete step $s \in \{0, \ldots, S-1\}$ whose normalized time s/S maps to log-SNR closest to λ_s under the rescaled schedule. In the continuous-time limit this reduces to an exact inversion of the log-SNR function at r_{crop}.

In summary, resizing preserves the forking timestep s while the interpolation factor t shifts, whereas random resize cropping preserves t while s must be recomputed. In both cases the true invariant is λ_s: the log-SNR at the moment of perceptual divergence.

E Comparison against Off-the-Shelf Metrics

Directly comparing off-the-shelf metrics to ours could make it hard to strictly ablate the benefits of our proposed approach. Thus, the results in the main paper, such as Table 1 is intended to be ablative. Most of the experimental settings are shared, from frozen backbone, prediction head initialization, optimizer, schedule and budget, etc., leaving the training data and the objective as the only variables. Yet, how our approach performs in comparison to off-the-shelf metrics is nonetheless worth establishing, and Table 10 presents the results, placing the public LPIPS-Alex, LPIPS-VGG, DISTS and DreamSim checkpoints beside the same architectures trained with FoMo supervision and evaluated under the same protocol. (In DreamSim, all images were resized to 224 resolution, following the protocol of DreamSim.)

Across the four architectures, FoMo supervision is comparable to the released checkpoints and slightly ahead overall, leading in half of the cells (25 of 48) and in all twelve for LPIPS-Alex. It is worth noting that this result is obtained under a single configuration applied to every backbone, shared unchanged across the three CNN backbones, with nothing tuned per architecture or per benchmark and without any human-annotated supervision. The released checkpoints, by contrast, each reflect considerable per-metric care: their own design choices, hyper-parameters and training sets. This result again proves the strength of FoMo, and further implies the potential that the metrics could still improve more with careful tuning of hyper-parameters for each model.

Table 10: Released off-the-shelf metrics compared with the same architecture trained with FoMo supervision. Better of each pair in **bold**. Released checkpoints are single deterministic models whereas FoMo columns are 5-seed means.

		LPIPS-Alex [43]		**LPIPS-VGG** [43]		**DISTS** [9]		**DreamSim** [13]	
		Released	FoMo	Released	FoMo	Released	FoMo	Released	FoMo
PIPAL [18]	SROCC	0.620	**0.733**	0.612	**0.683**	**0.672**	0.615	0.759	**0.776**
	KROCC	0.435	**0.536**	0.438	**0.497**	**0.482**	0.435	0.563	**0.571**
	PLCC	0.623	**0.767**	0.668	**0.712**	**0.685**	0.654	**0.780**	0.762
TID2013 [30]	SROCC	0.745	**0.785**	**0.670**	0.663	**0.708**	0.691	**0.812**	0.801
	KROCC	0.548	**0.586**	**0.497**	0.489	**0.521**	0.512	**0.614**	0.605
	PLCC	0.753	**0.816**	**0.749**	0.746	0.755	**0.766**	0.746	**0.831**
CSIQ [22]	SROCC	0.923	**0.938**	**0.883**	0.859	**0.930**	0.918	**0.911**	0.894
	KROCC	0.750	**0.781**	**0.697**	0.672	**0.764**	0.750	**0.738**	0.709
	PLCC	0.920	**0.944**	**0.906**	0.888	**0.938**	0.931	**0.928**	0.900
LIVE [34]	SROCC	0.924	**0.948**	**0.932**	0.923	0.948	**0.954**	0.910	**0.931**
	KROCC	0.751	**0.791**	**0.765**	0.748	0.793	**0.804**	0.745	**0.769**
	PLCC	0.916	**0.940**	**0.934**	0.923	0.945	**0.949**	0.918	**0.934**

F Per-Sample Label Variance

The forking construction is stochastic: two variants generated from the same reference at the same forking step are not identical, because the noise re-injected at the fork differs, yet both carry the same label. To quantify the resulting spread we take 120 ImageNet references, generate $K = 8$ variants at each of five forking steps, 4,800 images in total, and measure every variant's distance to its reference. For one reference at one forking step this gives eight distances, of which we take the mean and the standard deviation. Table 11 reports both averaged over the 120 references, together with their ratio, the coefficient of variation (CoV). The spread is small in every regime: the standard deviation stays below 0.05 in absolute terms and at most 10.1% of the distance it accompanies, while the mean distance itself changes six- to eightfold across the schedule. Re-running the generator therefore perturbs a sample by far less than the label separates it from its neighbors, and the perturbation reorders two variants of the same reference in at most 2.8% (LPIPS-Alex) and 7.0% (DISTS) of comparisons. Since this noise is independent across the 480k training pairs and averages out over them, we regard it as negligible for training.

Table 11: Spread of the measured distance across generation seeds. 120 references $\times$ 5 forking steps $\times$ $K = 8$ seeds. For each reference we take the eight distances obtained at one forking step and compute their mean and standard deviation; the table reports these averaged over the 120 references, with CoV their ratio. A larger s denotes a later fork, and hence a variant closer to the reference.

Forking step s	**LPIPS-Alex**			**DISTS**		
	mean d	std	CoV	mean d	std	CoV
5	0.692	0.043	6.4%	0.391	0.034	8.7%
15	0.518	0.046	9.1%	0.306	0.031	10.1%
25	0.371	0.030	7.9%	0.225	0.021	9.2%
35	0.233	0.011	4.7%	0.150	0.011	6.9%
45	0.090	0.002	2.7%	0.067	0.004	5.6%

G Per-Distortion-Type Analysis

Table 1 reports one SROCC per benchmark. This appendix asks which distortion families that number is built from. We recompute SROCC within each distortion type of all four benchmarks and compare FoMo against the strongest human-annotated recipe for the same backbone on the same benchmark. PIPAL is evaluated on its training split, the only one whose distortion types are identifiable, so its full-set values are not comparable to the validation numbers of Table 1.

With the Transformer backbone, FoMo supervision leads on almost every distortion family of every benchmark (Table 12). The gain is largest exactly where the backbone on its own is weakest, the super-resolution families of PIPAL, and the noise and contrast families of CSIQ. So the supervision, not the architecture, is what supplies the perceptual ordering. With the CNN backbone the picture is narrower, as one would expect of a network whose ImageNet features already encode a perceptual prior. FoMo leads throughout PIPAL, but on the three legacy synthetic-distortion benchmarks it trails on a majority of families, by margins of hundredths of a point.

The failures are consistent across both backbones and concentrate in three families (Table 13 lists every type): additive pixel noise, corruption confined to a small region, and global photometric shifts such as contrast change and mean shift. None of these occurs in our training data. A diffusion model re-synthesizes an image as a whole, so it never adds pixel-independent noise, never corrupts an isolated rectangle, and never applies a purely photometric change; supervision cannot teach what it never shows.

Table 12: Per-distortion-type summary on all four benchmarks. "Best baseline" is the strongest of the four human-annotated recipes for that backbone on that benchmark, selected independently per benchmark. "Full" is the score over the whole benchmark and "led" the number of distortion types on which FoMo is ahead. PIPAL uses its training split, the only one that exposes distortion types. Symmetrized native-resolution protocol, seed-0 checkpoints.

Benchmark	**types**	**LPIPS-Alex**			**DINOv3**		
		best baseline	FoMo	led	best baseline	FoMo	led
PIPAL	7	0.611	**0.681**	7/7	0.259	**0.525**	7/7
TID2013	24	**0.794**	0.783	10/24	0.592	**0.710**	22/24
CSIQ	6	**0.945**	0.936	1/6	0.613	**0.808**	6/6
LIVE	5	**0.952**	0.942	2/5	0.748	**0.896**	5/5

H Limitations

Despite the excellent performance in existing benchmarks. Our approach has a few limitations. One major limitation is that, since our approach generated distorted images with diffusion models, the metric models trained with our data and objective may fail in unseen image domains that are

Table 13: Every distortion type of all four benchmarks, against the strongest human-annotated recipe for the same backbone on the same benchmark. Bold marks the better of each pair. FoMo supervision leads on 20 of the 42 types with the CNN backbone and on 40 of the 42 with the Transformer backbone.

Distortion type	**LPIPS-Alex**		**DINOv3**	
	best baseline	FoMo (**Ours**)	best baseline	FoMo (**Ours**)
PIPAL [18] (training split)				
SR (traditional)	0.577	**0.669**	0.098	**0.618**
SR (PSNR-oriented)	0.707	**0.785**	0.315	**0.721**
SR (kernel mismatch)	0.560	**0.650**	0.349	**0.537**
SR (GAN-based)	0.501	**0.565**	0.172	**0.471**
Denoising	0.688	**0.757**	0.437	**0.693**
Mixture	0.570	**0.665**	0.370	**0.599**
Traditional	0.586	**0.628**	0.125	**0.346**
TID2013 [30]				
Additive Gaussian noise	**0.809**	0.766	0.432	**0.808**
Additive noise, colour comp.	**0.742**	0.690	0.352	**0.735**
Spatially correlated noise	0.717	**0.744**	0.659	**0.794**
Masked noise	**0.785**	0.770	0.133	**0.608**
High-frequency noise	**0.847**	0.806	0.506	**0.848**
Impulse noise	**0.552**	0.527	0.589	**0.633**
Quantisation noise	**0.786**	0.764	0.657	**0.828**
Gaussian blur	0.929	**0.931**	0.488	**0.785**
Image denoising	0.857	**0.871**	0.740	**0.868**
JPEG	**0.897**	0.887	0.710	**0.889**
JPEG2000	0.914	**0.934**	0.756	**0.881**
JPEG transmission errors	0.882	**0.898**	0.645	**0.803**
JPEG2000 transmission errors	**0.799**	0.791	0.568	**0.696**
Non-eccentricity pattern noise	0.782	**0.822**	0.279	**0.800**
Local block-wise distortion	0.335	**0.349**	**0.401**	0.277
Mean shift	**0.778**	0.737	0.126	**0.576**
Contrast change	**0.434**	0.410	**-0.029**	-0.159
Colour saturation change	**0.791**	0.782	0.271	**0.757**
Multiplicative Gaussian noise	**0.742**	0.691	0.478	**0.748**
Comfort noise	0.874	**0.877**	0.628	**0.890**
Lossy compression of noisy img.	**0.914**	0.901	0.734	**0.885**
Colour quantisation with dither	**0.812**	0.786	0.592	**0.837**
Chromatic aberrations	0.880	**0.890**	0.605	**0.783**
Sparse sampling and reconstr.	0.925	**0.943**	0.828	**0.915**
CSIQ [22]				
AWGN	**0.940**	0.913	0.459	**0.891**
Gaussian blur	**0.960**	0.956	0.625	**0.941**
Contrast change	**0.949**	0.929	0.076	**0.863**
Pink noise	**0.945**	0.917	0.593	**0.891**
JPEG	**0.958**	0.949	0.800	**0.955**
JPEG2000	0.942	**0.945**	0.781	**0.947**
LIVE [34]				
Fast fading	0.962	**0.965**	0.704	**0.960**
Gaussian blur	0.962	**0.964**	0.489	**0.903**
JPEG2000	**0.950**	0.941	0.711	**0.933**
JPEG	**0.965**	0.959	0.841	**0.962**
White noise	**0.961**	0.876	0.908	**0.966**

out-of-distribution, such as artificial distortions uncommon in nature. For scenarios like those, one could train a diffusion model on the new domain, and use it the trained diffusion model for generating new samples to train in that domain. This way, our approach can overcome its own limitation.

I Ethics Statement

The human annotation study involved only perceptual preference judgments on image pairs, with no deception, no collection of personally identifiable information, and no risk beyond normal screen use. Participants were informed of the task nature prior to annotation. The study was determined exempt from formal IRB review under the minimal-risk behavioral research exemption.

J Broader Impacts

Positive Societal Impact Automating perceptual label generation reduces reliance on costly human annotation, lowering the barrier to building human-aligned IQA metrics. Better metrics improve evaluation pipelines across image restoration and synthesis, benefiting applications in medical imaging, compression, and accessibility.

Negative Societal Impact More accurate perceptual metrics could be exploited to optimize generative models toward visually convincing outputs that conceal manipulations, potentially aiding synthetic media misuse. The metric may also inherit perceptual biases from the diffusion model used to generate training labels.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The method itself is experimented, and the results verify our contributions and scope, in the Experiments section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our method is described in detail in both Experiments section and the Appendix. Our code (`https://github.com/JHLew/FoMo`) and data (`https://huggingface.co/datasets/JHLew/fomo_480k`) are publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The main dataset and model used are ImageNet-1K [8] and FLUX [20], which are all publicly available, and could be fully reproduced using these. The code (`https://github.com/JHLew/FoMo`) and data (`https://huggingface.co/datasets/JHLew/fomo_480k`) are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The details are described in the Experiments section and the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide error bars or confidence bars in Tables 3,4,5,6,7,8,9,11.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, it is mentioned in the Experimental settings section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the authors have reviewed and the paper conforms the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: both are discussed in the Appendix of the paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: we have appropriately cited and credited the datasets and models used in the research of this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the synthetic dataset (`https://huggingface.co/datasets/JHLew/fomo_480k`) and the trained models (`https://github.com/JHLew/FoMo`) with well documentations.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide the screenshot and user interface of our annotation site in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We provide an ethics statement section in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used for writing and coding assistance, and does not impact the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.