NearID: Identity Representation Learning via Near-identity Distractors

Aleksandar Cvejic[1], Rameen Abdal[2⋆], Abdelrahman Eldesokey[1], Bernard Ghanem[1], and Peter Wonka[1]

[1] King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[2] Snap Research, Palo Alto, CA, USA

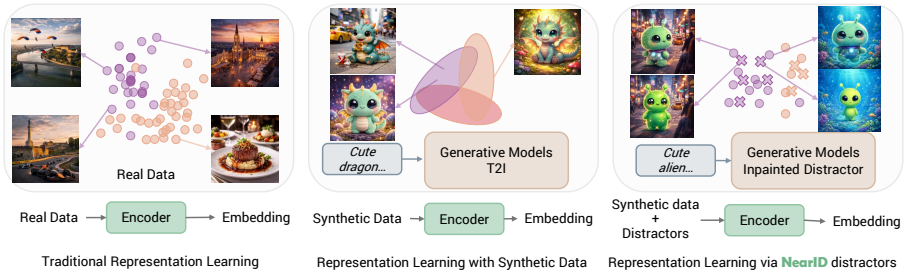


Fig. 1: NearID: From Context to Identity. (Left) Traditional representations entangle object identity with background context. **(Middle)** Synthetic Data lacks explicit control over visually similar distractors. **(Right)** *NearID* introduces matched-context distractors to remove contextual shortcuts and isolate intrinsic identity signals.

Project page: `https://gorluxor.github.io/NearID/`

Abstract. When evaluating identity-focused tasks such as personalized generation and image editing, existing vision encoders entangle object identity with background context, leading to unreliable representations and metrics. We introduce a principled framework for matched-context identity confusion using Near-identity (NearID) distractors, where semantically similar but distinct instances are placed on the exact same background as a reference image, eliminating contextual shortcuts and isolating identity as the sole discriminative signal. Based on this principle, we present the NearID dataset (19K identities, 316K matched-context distractors) together with a strict margin-based evaluation protocol. Under this setting, pre-trained encoders perform poorly, achieving Sample Success Rates (SSR), a strict margin-based identity discrimination metric, as low as 30.7% and often ranking distractors above true cross-view matches. We address this by learning identity-aware representations on a frozen backbone using a two-tier contrastive objective enforcing the hierarchy: same identity > NearID distractor > random negative. This improves SSR to 99.2%, enhances part-level discrimination by 28.0%, and yields stronger alignment with human judgments on DreamBench++, a human-aligned benchmark for personalization.

Keywords: Representation Learning · Identity Representation

⋆ Served in an advisory role.

1 Introduction

How reliably can a vision encoder tell whether two images depict the *same* object? When the backgrounds differ, modern foundation models, including CLIP [51], DINOv2 [45], SigLIP2 [66], and even large vision-language models such as Qwen3-VL [3], often succeed. Yet these same models are systematically fooled by a simple adversarial test: replace the object with a NearID (*different but visually similar*) instance while keeping the background *identical*. Under this condition, the shared context dominates the embedding, causing the impostor to score *higher* than the true identity viewed against a different backdrop. Such background-entangled metrics actively inflate the CLIP-I and DINO scores that the majority of personalization and editing literature relies on for automated identity preservation evaluation [22, 48, 53].

We provide a principled framework for matched-context identity confusion through Near-identity (NearID) (See Figure 1): semantically similar but distinct instances inpainted into the exact same background as the reference, removing all contextual shortcuts and isolating identity as the sole discriminative signal. Under the resulting NearID evaluation protocol, the frozen SigLIP2 backbone achieves only 30.74% Sample Success Rate (SSR), threshold-based similarity based on Near-identity distractors and positives defined in Sec. 3.4, and on part-level manipulations from the Mind the Glitch (MTG) dataset [14], *every* standard encoder (CLIP, DINOv2, SigLIP2) scores exactly 0.0% SSR.

To close this gap we introduce **NearID**, comprising three tightly coupled contributions: (i) a large-scale dataset of 19k object identities with multi-view positives and over 316k NearID distractors synthesized from four generative models, providing the training signal and evaluation benchmark; (ii) a two-tier contrastive objective that enforces an explicit similarity hierarchy (same identity $>$ NearID distractor $>$ random batch negative), teaching the model *what* to be invariant to and *what* to discriminate; and (iii) a lightweight training recipe that keeps the foundation encoder frozen and learns only a Multi-head Attention Pooling (MAP) [25, 82] projection head ($\sim$3.6% of total parameters), preserving the general-purpose priors while reshaping the similarity geometry for identity.

Empirically, NearID resolves the identified failure at the object level on NearID, where SSR rises from 30.74% to 99.17%, and at the part level on Mind-the-Glitch (MTG), where SSR improves from 0.0% to 35.0% [14], a setting in which even the dedicated Visual Semantic Matching (VSM) metric reaches only 7.0% under the same training data and pairwise evaluation protocol [14]. Beyond SSR, we demonstrate substantially improved alignment with benchmark oracles and human judgments despite training exclusively on near-identity distractors: on MTG, Pearson correlation to the metric oracle increases from 0.180 to 0.465 under the full evaluation with view and background changes, and from 0.366 to 0.486 under the paired background context protocol (Sec. 3.4); on DreamBench++ [48], our identity-specific tuning improves overall alignment with human concept-preservation judgments even though our training data excludes style prompts and live subjects, and it generalizes to animals and humans with correlation gains of 0.105 and 0.065, respectively. These results indicate that the

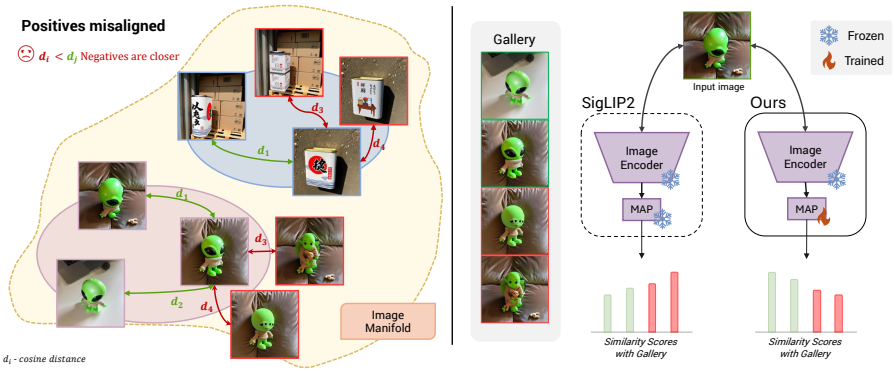


Fig. 2: NearID overview. Left: In the pretrained image-embedding manifold (e.g., SigLIP2 [66]), identity-consistent *positives* can be misaligned: edited or re-rendered views of the same instance may lie farther from the anchor than visually confusable *negatives* (NearID distractors illustrated by $d_i < d_j$), which degrades retrieval and scoring reliability. **Right:** We keep the SigLIP2 image encoder frozen and train only a lightweight MAP [25, 82] projection head to reshape the similarity geometry for the target task. Compared to the frozen baseline, the trained head increases similarity for true positives (green) while pushing NearID distractors and unrelated gallery items lower (red), improving the desired ordering of similarity scores with respect to a gallery.

learned representations capture genuine identity cues rather than overfitting to synthetic artifacts.

In summary, our contributions are:

1. We formalize **Near-identity distractors**: confounders that share the exact background as the reference while changing the identity signal. We introduce **NearID-bench**, a matched-context evaluation protocol (SSR and PA) that quantifies identity context entanglement in modern vision encoders, and we release a large-scale **NearID** dataset (19k identities, 316k+ distractors, 4 generative models) to support research in this setting.
2. Building on the standard **MAP projection head** used with SigLIP2, we train a lightweight identity adapter on a frozen SigLIP2 backbone using a **two-tier contrastive objective** that enforces a strict similarity ordering: same identity > NearID distractor > random batch negative.
3. We show that this training yields embeddings that substantially improve matched-context NearID distractor rejection (NearID SSR and PA), increase sensitivity to localized identity edits in part-level evaluation (MTG), and increase metric-to-human agreement on DreamBench++, indicating that reducing identity context coupling leads to more reliable personalization evaluation.

Table 1: Near-identity discrimination and alignment. We evaluate matched-context Near-identity discrimination at the object level (NearID) and part level (MTG [14]), and measure correlation with MTG oracle scores and human concept-preservation judgments on DreamBench++ (DB++) [48], verifying alignment with both oracle-defined edit severity and human perception. **Top:** Existing embedding and VLM methods under the Near-identity protocol. **Bottom:** Training in our setting, showing improvements on NearID and gains on MTG and DB++. MTG reports metric-to-oracle alignment (M–O/M–O_{PAIR}); DB++ reports metric-to-human alignment (M–H). Deltas are relative to the frozen SigLIP2 baseline. * denotes a method trained directly on MTG. SSR and PA are averaged across seven inpainting settings (three excluded from training; see Sec. 3.4).

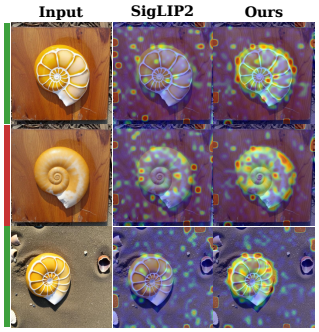
Scoring Model	NearID		MTG				DB++
	SSR ↑	PA ↑	M–O ↑	M–O_{PAIR} ↑	SSR ↑	PA ↑	M–H ↑
Qwen3VL30B [3]	49.73	69.20	0.219	0.329	17.0	26.0	–
CLIP [51]	10.31	20.92	0.239	0.484	0.0	0.0	0.493
DINOv2 [45]	20.43	34.55	0.324	**0.519**	0.0	0.0	0.492
VSM* [14]	32.13	46.70	0.394	0.445	7.0	24.5	0.190
SigLIP2 [66]	30.74	48.81	0.180	0.366	0.0	0.0	0.516
NearID (Ours)	**99.17** (+68.43)	**99.71** (+50.90)	**0.465** (+0.285)	0.486 (+0.120)	**35.0** (+35.0)	**46.5** (+46.5)	**0.545** (+0.029)

2 Related Work

Visual foundation models and identity representation: Representation learning has advanced rapidly through self-supervised [6, 9, 15, 19, 21], multimodal [2,26,38,51,79], and generative pretraining [52,53,68,69], driving progress across discriminative [32,39] and generative tasks. Modern vision-language models (VLMs) [1, 3, 11, 37, 40, 70] are now widely adopted as general-purpose encoders [56, 66, 73, 83]. Contrastive language-image models [51, 66] excel at broad category-level alignment, and the self-supervised DINO family [6,45] learns powerful dense features and patch-level structural representations. However, because these models are trained for broad semantic grouping, they often entangle instance identity with context [57], yielding high similarity for images with similar backgrounds despite different instances. Recent bias analyses confirm that VLM behavior can be strongly influenced by background patterns, and that removing the background can substantially reduce such biases [67]. Specialized identity encoders [13, 31, 42, 46, 55] achieve strong identity separation but remain confined to narrow domains such as human faces. Region-focused approaches like Alpha-CLIP [61] augment CLIP with an auxiliary alpha channel to extract features from user-specified regions (e.g., masks or bounding boxes) while retaining contextual awareness, but still lack explicit identity-level supervision. In contrast, our work seeks to extract open-domain, instance-level identity representations. Rather than risking catastrophic forgetting by fine-tuning a foundation model, we leverage the robust priors of a frozen backbone and utilize a lightweight Multi-head Attention Pooling (MAP) head [25,82], which we train to explicitly attend to identity-salient features while suppressing contextual background cues.

Deep metric learning and hierarchical contrastive objectives: Standard deep metric learning relies on pulling positive pairs together while pushing negatives apart. Traditional formulations utilize batch-based contrastive losses [9, 43] or supervised extensions that group multiple positives SupContrast [29]. While effective for instance discrimination, these binary objectives treat all negatives equally, forcing the model to push semantically related intra-class (NearID) samples and completely unrelated random samples to the same near-zero similarity. To inject structure into the negative space, prior works have explored triplet mining [55], multi-negative objectives [58], and dynamic pair weighting [60, 71]. Hierarchical metric learning further attempts to impose discrete geometric boundaries using ranked lists or tree-based proxies [72, 77]. However, these methods typically rely on online algorithmic selection of informative negatives, which is unstable during training, or on predefined categorical taxonomies. Our approach differs fundamentally by utilizing *explicitly curated* near-identity distractors generated via controlled inpainting. We enforce a strict three-tier structural hierarchy (positives > Near-identity negatives > batch negatives) through our dual-tier objective, providing stable, constant gradient pressure without the instability of algorithmic mining. Closest to our use of contrastive signals for identity, CustomContrast [8] employs a multi-level contrastive objective for subject-driven generation, but co-trains an encoder *within* a text-to-image generator; in contrast, NearID learns a standalone, generator-agnostic identity *metric* purely from near-identity distractor margins. **Continuous calibration and learning-to-rank:** Beyond strict categorical ranking, identity preservation often exhibits varying degrees of severity, requiring continuous calibration of similarity scores. Soft contrastive learning and contrastive regression frameworks [28,64,81] replace binary targets with continuous variables, adjusting the contrastive push based on label distance. Similarly, Learning-to-Rank (LTR) objectives optimize for monotonic sorting across retrieved items rather than absolute distances [4,5,76]. While post-hoc calibration strategies like temperature scaling or isotonic regression [20, 80] can adjust score distributions after training, they cannot fix underlying manifold collisions. In contrast to methods requiring explicit continuous regression targets, our approach achieves data-driven structural calibration. By subjecting a diverse curriculum of part-level edits (via the MTG dataset [14]) to our ranking regularizer ($\mathcal{L}_{\text{rank}}$), the embedding magnitudes naturally correlate with physical edit severity. This smoothly interpolates between identical instances and unrelated batch negatives without relying on manual numeric margin during optimization.

Evaluation of personalized image generation: The rapid advancement of subject-driven generation and image editing [10, 18, 41, 50, 53, 69, 75, 78] has exposed a critical vulnerability in current evaluation protocols. The vast majority of personalization literature relies on out-of-the-box CLIP [51] cosine similarity (CLIP-I) or DINO-scores to quantify identity preservation [22]. However, as our experiments demonstrate, these generic semantic metrics are easily fooled by near-identity confounders, leading to inflated scores when a model generates a different instance on the correct background. Recent work by Kilrain et al. [30] further shows that pairwise CLIP/DINO similarities can remain near-perfect



	DINOv2	SigLIP2	VSM	NearID
NearID object-level	✗	✗	✗	✓
NearID part-level	✗	✗	✓	✓
Mask-free	✓	✓	✗	✓
Human alignment	✓	✓	✗	✓
Oracle alignment	✗	✗	✓	✓

Fig. 3: Left: Attention map comparison vs baseline on positives and negatives. **Right:** Summary of properties evaluated in this paper. NearID improves matched-context Near-ID rejection on NearID-bench and transfers to part-level identity evaluation on MTG, while remaining mask-free at inference.

when layout and background match while identity-defining details drift, and proposes a gallery-based retrieval protocol to expose this failure mode. In personalization pipelines, this identity and context coupling manifests in both learning and inference: optimizing textual identifiers with standard diffusion losses can entangle foreground and background information [7], and diffusion-transformer architectures with global attention can conflate semantic identity with spatial layout [24]. Current evaluation protocols using VLMs as judges further compound these issues, since they provide the full image as input rather than isolating the subject [33, 85].

Recent efforts to bridge this gap have introduced perceptual metrics calibrated to human judgments, ranging from low-level distortion penalties [84] to mid-level holistic similarity metrics like DreamSim [17]. While DreamSim successfully aligns with human preferences regarding layout, pose, and overall composition, it evaluates *holistic* scene similarity rather than isolating *strict object identity*. Consequently, it remains susceptible to background conflation when a structurally similar distractor shares the reference context. Furthermore, recent analysis by PercepAlign [62] demonstrates that fully fine-tuning vision encoders to align with generic human perceptual judgments actively degrades their underlying high-level semantic representations.

Our proposed framework addresses these dual challenges. To prevent representation degradation, we freeze the foundation backbone and tune only a lightweight MAP head to project features into a specialized identity subspace. Concurrently, our NearID protocol provides the first rigorous, automated evaluation standard explicitly designed to penalize background conflation and reward true object-level identity preservation rather than holistic visual similarity.

3 NearID

3.1 Overview and Problem Formulation

We formulate identity preservation as a structured metric learning problem designed to explicitly disentangle object identity from background context (see Figure 2). **NearID** is a unified framework comprising three tightly coupled components: (i) a hierarchical contrastive objective that enforces an explicit similarity ordering between positives, near-identity distractors, and random negatives (Section 3.2), (ii) a large-scale matched-context dataset with curated identities and synthesized near-identity distractors that instantiate this training setting (Section 3.3), and (iii) a rigorous evaluation protocol based on discriminability margins and alignment with human and oracle judgments (Section 3.4).

Together, these elements isolate identity as the sole discriminative signal and enable principled training and reliable evaluation under matched-context confounders.

Training Tuple Construction. Each training sample is constructed as a tuple $\mathcal{T}_i = \{a_i, \{g_{i,p}\}_{p=1}^{P}, \{r_{i,k}\}_{k=1}^{K}\}$ where a_i denotes an anchor image of object identity i, $g_{i,p}$ are up to P positive views of the *same* identity against different backgrounds, $r_{i,k}$ are K *near-identity distractors*. A NearID distractor consists of a semantically similar but distinct object instance that is inpainted into the *exact same background* as the anchor a_i. This matched-context construction removes all contextual shortcuts and forces the model to discriminate based solely on intrinsic identity cues.

Learning Objective. The goal is to learn an embedding function $f_\theta : \mathcal{I} \to \mathbb{R}^d$ such that, for a given anchor a_i, in terms of cosine similarity, $\cos(a_i, p_i) > \cos(a_i, n_i^{\text{near}}) > \cos(a_i, n_i^{\text{rand}})$, i.e., same identity > NearID distractor > random batch negative. This explicitly enforces a structured similarity hierarchy rather than treating all negatives equally.

Parameter-Efficient Adaptation. To avoid catastrophic forgetting and preserve the robust semantic priors of large-scale vision encoders, we freeze a pre-trained backbone (SigLIP2 [66]) and optimize only a lightweight Multi-head Attention Pooling (MAP) projection head [25, 82].

Given spatial patch embeddings extracted from the frozen backbone, the MAP head selectively aggregates identity-salient features and projects them into a refined metric space. This design updates only $\sim$3.6% of total parameters while reshaping the embedding geometry to satisfy the identity hierarchy defined above (See Figure 3 left pane). In the following subsection, we formalize the extended hierarchical contrastive objective used to enforce this structure.

3.2 NearID Loss

Standard contrastive objectives treat all negatives uniformly, failing to penalize background-induced correlations that allow models to shortcut on contextual

cues rather than intrinsic identity features (illustrated in Figure 1 and Figure 2). We address this with the **NearID loss** ($\mathcal{L}_{\text{NearID}}$), a two-component objective that enforces a structured three-level similarity hierarchy: true identity matches above NearID distractors, and NearID distractors above generic batch negatives.

Notation. Let $\phi(x)$ denote frozen backbone features and let f denote the trainable MAP projection head applied on top of ϕ. For an image x, we define the ℓ_2-normalized embedding $\mathbf{z}_x = \frac{f(\phi(x))}{\|f(\phi(x))\|_2} \in \mathbb{R}^d$. All pairwise comparisons use temperature-scaled similarity logits $\ell_{u,v} = \frac{\mathbf{u}^\top \mathbf{v}}{\tau}$, where $\tau > 0$ is a fixed temperature hyperparameter ($\tau = 0.07$). Since embeddings are ℓ_2-normalized, dot products equal cosine similarities.

For each anchor image a_i, $\{g_{i,p}\}_{p=1}^{P}$ positive views, and $\{r_{i,k}\}_{k=1}^{K}$ NearID distractors (See Section 3.1), we define their embeddings as $\mathbf{a}_i = \mathbf{z}_{a_i}$, $\mathbf{g}_{i,p} = \mathbf{z}_{g_{i,p}}$, and $\mathbf{r}_{i,k} = \mathbf{z}_{r_{i,k}}$. Let $\mathcal{G} = \{\mathbf{g}_{j,q}\}$ denote the global set of all positive embeddings gathered across devices (DDP), with $|\mathcal{G}| = M$. Define the anchor's own positive set $\mathcal{G}_i = \{\mathbf{g}_{i,q}\}_{q=1}^{P} \subset \mathcal{G}$ and the batch-negative pool $\mathcal{B}_{\text{neg}}^{(i)} = \mathcal{G} \setminus \mathcal{G}_i$. Note that the NearID distractor set for anchor i, $\mathcal{R}_i = \{\mathbf{r}_{i,k}\}_{k=1}^{K}$, is defined per-anchor and is not included in $\mathcal{G}$. In practice, we average only over valid positives and distractors using masks (omitted for clarity).

Discrimination term ($\mathcal{L}_{disc}$). For each positive index p, the discrimination term applies a softmax cross-entropy over the global positive pool augmented with the NearID distractors as $\mathcal{L}_{\text{disc},p}^{(i)} = -\log \frac{\exp(\ell_{\mathbf{a}_i,\mathbf{g}_{i,p}})}{\sum_{g\in\mathcal{G}} \exp(\ell_{\mathbf{a}_i,g}) + \sum_{k=1}^{K} \exp(\ell_{\mathbf{a}_i,\mathbf{r}_{i,k}})}$.

This formulation is structurally related to InfoNCE [44] and multi-class N-pair objectives [58], but with two key differences. First, the NearID distractors $\mathbf{r}_{i,k}$ are explicitly included in the denominator alongside global batch candidates, providing targeted identity-confusing competition that standard batch negatives cannot supply. Second, unlike Supervised Contrastive Learning [29], which places all same-class positives in the numerator, we retain the other $P-1$ positive views in the denominator. This per-index formulation mitigates multi-positive collapse and ensures that each view is individually discriminable under background variation. The discrimination loss averages over valid positive indices:

$$\mathcal{L}_{\text{disc}} = \frac{1}{P} \sum_{p=1}^{P} \mathbb{E}_i \left[\mathcal{L}_{\text{disc},p}^{(i)} \right]. \tag{1}$$

Ranking regularizer ($\mathcal{L}_{rank}$). The discrimination term alone treats NearID distractors simply as additional negatives to push away. Under strong gradient pressure from highly confusable distractors, this can collapse the local semantic neighborhood, destroying the graded similarity structure that distinguishes "nearby but different" from "unrelated", a known concern in metric learning with highly confusable negatives [72]. To preserve this structure, we add a ranking regularizer that encourages each NearID distractor to be ranked *above* the generic batch-negative pool. Define the batch-negative log-sum-exp as $\text{LSE}_i =$

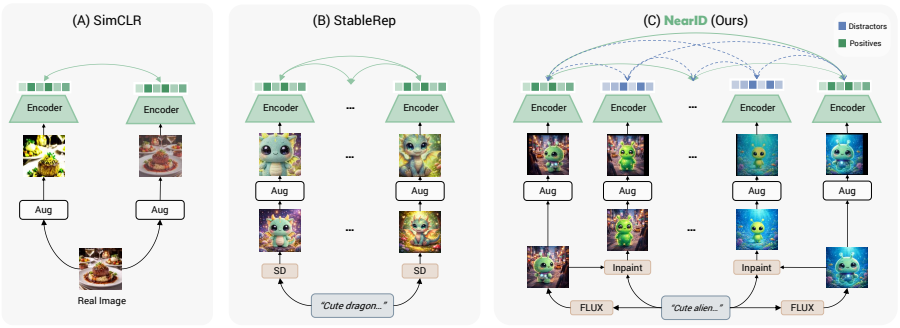


Fig. 4: Contrastive training paradigms. Unlike **(A)** SimCLR [9] (augmented views of a single image) and **(B)** StableRep [65] (same-caption generations), **(C)** NearID uses multi-view positives of the *same* instance and adds *near-identity distractors*, visually similar but distinct instances inpainted into the anchor's background, a structured negative signal absent from both. See the supplement for a full discussion.

$\log \sum_{g \in \mathcal{B}_{\text{neg}}^{(i)}} \exp(\ell_{\mathbf{a}_i, g})$. which acts as a smooth maximum over the batch-negative pool.

For each NearID distractor $r_{i,k}$, the ranking regularizer takes the form, $\mathcal{L}_{\text{rank}}^{(i,k)} = \log\Big(1 + \exp\big(\text{LSE}_i - \ell_{\mathbf{a}_i, \mathbf{r}_{i,k}}\big)\Big)$.

which is a softplus penalty that decreases smoothly as the NearID distractor logit exceeds the aggregate batch-negative signal, and increases when batch negatives dominate. This is equivalent to the cross-entropy form, $\mathcal{L}_{\text{rank}}^{(i,k)} = -\log \frac{\exp(\ell_{\mathbf{a}_i, \mathbf{r}_{i,k}})}{\exp(\ell_{\mathbf{a}_i, \mathbf{r}_{i,k}}) + \sum_{g \in \mathcal{B}_{\text{neg}}^{(i)}} \exp(\ell_{\mathbf{a}_i, g})}$.

However, the softplus view makes the listwise ranking nature explicit: LSE_i acts as a smooth maximum over the batch-negative pool, and the penalty enforces that each NearID distractor logit exceeds this threshold. Aggregating:

$$\mathcal{L}_{\text{rank}} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_i \Big[\mathcal{L}_{\text{rank}}^{(i,k)}\Big]. \tag{2}$$

NearID loss. The complete objective combines discrimination and ranking:

$$\mathcal{L}_{\text{NearID}} = \mathcal{L}_{\text{disc}} + \alpha \mathcal{L}_{\text{rank}}, \tag{3}$$

where $\alpha = 0.5$ controls the ranking regularizer strength. Together, the two terms enforce the intended similarity ordering $\ell_{\mathbf{a}_i, \mathbf{g}_{i,p}} > \ell_{\mathbf{a}_i, \mathbf{r}_{i,k}} \gtrsim \text{LSE}_i$.

ensuring that true positives are selected over all candidates ($\mathcal{L}_{\text{disc}}$), while NearID distractors remain closer than generic batch negatives ($\mathcal{L}_{\text{rank}}$), preserving the graded semantic structure of the embedding space.

3.3 NearID Dataset Construction

Source datasets. SynCD [34] renders up to three views of the *same* instance across backgrounds from Objaverse [12] 3D assets via conditioned diffusion; all such views, including front and back, are positives, as identity rather than viewpoint defines a positive pair. We extend SynCD by adding *near-identity distractors* as a structured negative signal absent from prior contrastive paradigms (Figure 4). **MTG** [14] pairs the *same* object under a localized part edit, providing a fine-grained calibration signal.

Dataset Foundation and Curation. Constructing a robust benchmark for object-level identity requires data with high viewpoint diversity and strict identity separation. While datasets such as Mind-The-Glitch (MTG) [14] and Subjects200k [63] offer valuable baselines, they are limited by scale (e.g., 5K samples in MTG), a lack of multi-view diversity, and a focus on subject-level matching against a single plain-background reference, rather than instance-level object identity under viewpoint and background changes. Furthermore, MTG explicitly cites significant dataset quality issues within Subjects200k, which further motivates our decision to improve upon it.

To address these limitations, we build our synthesis pipeline on the SynCD dataset [34] (90K initial samples). To guarantee absolute identity fidelity and prevent semantic overlap during training and evaluation, we apply a rigorous filtering protocol. We first discard the "deformable" partition (approximately half of the dataset) because the publicly released version lacks the exact foreground text descriptions required for precise generative conditioning and low diversity with 16 base classes. Within the remaining "rigid" object subset, we enforce strict uniqueness constraints based on paired textual descriptions and `ObjaverseID` [12]. This curation process yields our proposed dataset, **NearID**, comprising 19,386 unique object identities and 45,215 distinct images. Notably, 12,943 identities (67%) possess two distinct views, and 6,433 identities (33%) provide three distinct views. This scale and viewpoint coverage significantly surpasses the entire MTG dataset (limited to 5K identities with only 2 views). We create mutually exclusive train/val/test splits with 18,786/100/500 identities, respectively.

Multi-Model Synthesis Pipeline. Using this curated core subset, we generate near-identity distractors employing a diverse ensemble of SoTA diffusion models. This pipeline encompasses Stable Diffusion XL (SDXL) [49], FLUX.1 [35], Qwen-Image [74], and PowerPaint [86] (utilizing the BrushNet v2.1 pipeline [27]), which ensures a wide variance in generative priors, structural adherence, and artifact distributions. Overall, this yields 7 inpainting configurations and 316,505 near-identity distractor images, disentangling inpainting artifacts, model-specific biases, and source-dependent generative fingerprints.

To supplement our object-level distractors with fine-grained variations, we concatenate the MTG dataset, which inherently consists of part-level inpaintings from a single source (SDXL). To ensure equal sampling probability during training, the MTG subset is repeated four times to match the scale of our syn-

thesized data. Generation is conducted using mixed precision, where supported by the model architecture, to optimize throughput during large-scale dataset construction. We use default parameters from publicly available code for each method, with slight modifications when utilizing adapters. Exact inference hyperparameters are detailed in the supplementary material.

Data Processing Pipeline. Generation is performed at two distinct resolutions: 512^2 and the respective model-native resolutions (e.g., spatial scaling to align the shortest side with 640 for SD 1.5-based variants like BrushNet, or strictly 1024^2 for SDXL), except for BrushNet, which operates exclusively at its native resolution. Following the synthesis of the near-identity distractors, spatial standardization is enforced by downsampling any high-resolution outputs to 512^2 utilizing Lanczos resampling. This ensures consistent dimensions for downstream evaluation metrics and contrastive training while mitigating aliasing artifacts.

3.4 Evaluation Protocol

Given the tuple formulation introduced in Section 3.1, we evaluate identity discrimination under a matched-context setting designed to eliminate background shortcuts. The protocol analyzes similarity margins within strict identity–background triplets. For a given identity, let $\mathbf{p}_i$ and $\mathbf{p}_j$ denote two positive views of the same object captured against distinct backgrounds i and j. Let $\mathbf{n}_i$ denote a NearID distractor: a semantically similar but distinct instance inpainted into the exact background of $\mathbf{p}_i$.

Bidirectional Discriminability Margin. We measure whether identity similarity across backgrounds exceeds similarity to a matched-context distractor. For a pair (i, j), we define the directed margins:

$$\delta_{i\to j}^{(ij)} = s(\mathbf{p}_i, \mathbf{p}_j) - s(\mathbf{p}_i, \mathbf{n}_i), \qquad \delta_{j\to i}^{(ij)} = s(\mathbf{p}_i, \mathbf{p}_j) - s(\mathbf{p}_j, \mathbf{n}_j). \tag{4}$$

A margin trial is considered successful if $\delta > 0$, indicating that cross-background identity similarity overrides matched-context confounders.

Micro-pooled Metrics. We aggregate directed margin trials using two measures:

Sample Success Rate (SSR). A sample is counted as successful only if *all* valid directed margins are positive (up to six per identity):
Pairwise Accuracy (PA). The proportion of successful directed margin trials across the dataset, treating each margin independently:

$$\mathrm{SSR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big[\forall\,(j,d):\ \delta_d^{(ij)} > 0\Big], \quad \mathrm{PA} = \frac{\big|\{m \in \mathcal{M} : \delta_m > 0\}\big|}{|\mathcal{M}|}, \tag{5}$$

where N is the number of identities and $\mathcal{M}$ is the set of all directed margin trials.

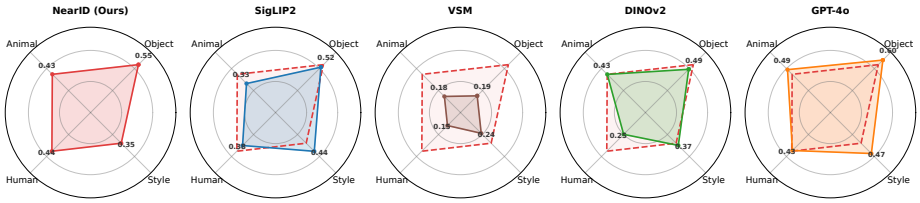


Fig. 5: Per-category human alignment ($M-H$) on DreamBench++ [48]. Each radar shows the Pearson correlation between a metric and human concept-preservation judgments across four DB++ categories; NearID (Ours) is repeated as a dashed reference in each subplot. Despite training exclusively on rigid objects, NearID improves over the frozen SigLIP2 baseline on *Animal* (+0.105) and *Human* (+0.065), indicating that disentangling identity from context transfers across semantic domains. The expected decrease on *Style* (−0.092), a category entirely absent from training, confirms that the gains reflect genuine identity learning rather than general score inflation.

Table 2: DreamBench++ concept-preservation human alignment. Pearson correlations between human (H) and metric scores: G = GPT, D = DINO, C = CLIP, S = SigLIP2, NearID (Ours). Bolded numbers indicate improvement over SigLIP2.

Method	T2I Model	Concept Preservation (Metric–Human Alignment)					
		H–H	G–H	D–H	C–H	S–H	NearID–H
Textual Inversion [18]	SD v1.5	0.684	0.625 ± 0.032	0.611 ± 0.016	0.610 ± 0.001	0.660 ± 0.003	**0.675** ± 0.007
DreamBooth [53]	SD v1.5	0.731	0.722 ± 0.011	0.632 ± 0.010	0.623 ± 0.015	0.652 ± 0.012	**0.655** ± 0.023
DreamBooth LoRA [23, 53]	SDXL v1.0	0.651	0.637 ± 0.004	0.446 ± 0.013	0.481 ± 0.007	0.480 ± 0.006	**0.546** ± 0.006
BLIP-Diffusion [36]	SD v1.5	0.630	0.466 ± 0.019	0.311 ± 0.024	0.290 ± 0.004	0.363 ± 0.016	**0.382** ± 0.002
Emu2 [59]	SDXL v1.0	0.763	0.777 ± 0.004	0.774 ± 0.018	0.744 ± 0.021	0.766 ± 0.021	**0.768** ± 0.028
IP-Adapter-Plus ViT-H [78]	SDXL v1.0	0.530	0.424 ± 0.045	0.212 ± 0.017	0.262 ± 0.002	0.253 ± 0.002	**0.306** ± 0.018
IP-Adapter ViT-G [78]	SDXL v1.0	0.485	0.401 ± 0.021	0.261 ± 0.004	0.283 ± 0.031	0.248 ± 0.019	**0.316** ± 0.041
Fisher-z ($\bar{r}$ - Sec. 4.2)		0.648 ± 0.038	0.597 ± 0.056	0.492 ± 0.081	0.493 ± 0.074	0.516 ± 0.079	**0.545** ± 0.071

Alignment with Human and Oracle Judgments. To ensure our learned embeddings align with fine-grained identity preservation, we evaluate against the MTG dataset [14] and DreamBench++ (DB++) [48]. We utilize the MTG part-level oracle score, defined as $1-(|\Omega_p|/|\Omega_o|)$, where Ω_p and Ω_o represent the part and object masks, respectively. Hence, we also compute two Pearson [47] correlations:

M − O. Correlations between the cosine similarities and the oracle scores.
M − H. Correlations between the cosine similarities and human scores on DB++.

All correlations are reported as the mean Pearson correlation, $\bar{r}$, computed via Fisher's z-transformation [16] over per-sample coefficients (see supplementary material for the aggregation formula).

4 Experiments

4.1 Implementation Details

Architecture and Optimization. All embedding models in our proposed pipeline share a frozen `SigLIP2-so400m-patch14-384` backbone, unless otherwise spec-

Table 3: Training objective ablation. All variants share a frozen SigLIP2 backbone with a trainable MAP head, trained on NearID + MTG data for 11 epochs with identical hyperparameters. More details in the supplementary material about the specifics of each loss and the extension. †Representation collapse: high discrimination but degraded general-purpose alignment (see text). We observe that our configurations pose the best balance between the different adaptations to the NearID setting.

Training Loss	NearID		MTG				DB++
	SSR ↑	PA ↑	M–O ↑	M–O_{PAIR} ↑	SSR ↑	PA ↑	M–H ↑
None (frozen)	30.74	48.81	0.180	0.366	0.0	0.0	0.516
InfoNCE (sym., 1-pos)	60.97	75.26	0.267	0.418	8.0	12.0	0.555
+ $\mathcal{R}$ neg	99.57	99.79	0.236	0.267	64.0	74.0	0.251
+ Oracle Ranking†	86.34	92.25	0.299†	0.444†	7.0†	11.0†	0.167†
+ $\mathcal{R}$ neg + Oracle Ranking	99.60	99.89	0.247	0.277	65.0	74.5	0.227
SigLIP (BCE, ext., + hard>batch rank)	61.40	77.81	0.213	0.385	6.0	9.5	0.530
Circle (ext.,+hard>batch margin rank)†	*99.97*	*99.99*	0.264†	0.303†	*67.0*†	*76.5*†	0.141†
$\mathcal{L}_{\text{NearID}}$ **(Ours)**	**99.17**	**99.71**	**0.465**	**0.486**	35.0	46.5	**0.545**
+ Pos. Cohesion	99.31	99.78	0.459	0.485	36.0	47.0	0.541

ified. We train only the Multihead Attention Pooling (MAP) projection head, which outputs a 1152 dimensional ℓ_2-normalized embedding. This head-only tuning regime updates approximately 15M parameters out of the roughly 428M total model parameters. By training merely ∼3.6% of the overall network capacity, we ensure minimal computational overhead while preserving the robust zero-shot priors of the foundation model. Models are optimized using AdamW ($\eta = 10^{-4}$, weight decay=10^{-4}) in mixed-precision (`fp16`). We employ a cosine annealing schedule with a 100-step linear warmup, utilizing a global batch size of 128 across 11 epochs (approximately $3,350$ gradient steps).

Role-Aware Data Augmentation. We apply role-aware stochastic foreground masking: backgrounds are blacked out for anchors (p=0.5), positives (p=0.2), and distractors (p=0.6), the latter to prevent trivial rejection via background border artifacts. Positives receive the lightest masking because their cross-background variation is itself the identity signal; a mask-free ablation (see supplementary material) confirms this masking is essential for background-invariant, part-level transfer. Standard augmentations (color jitter, flipping, translation/scale) are applied per slot.

Joint Training. The NearID dataset is interleaved with the MTG training split (upsampled 4×). MTG's part-level edits are subjected to $\mathcal{L}_{\text{rank}}$ without explicit regression targets, enabling continuous calibration from the data distribution.

4.2 Results

We evaluate against frozen contrastive models (CLIP [51], SigLIP2 [66]), self-supervised transformers (DINOv2 [45]), a VLM (Qwen3-VL 30B [3]), and VSM [14],

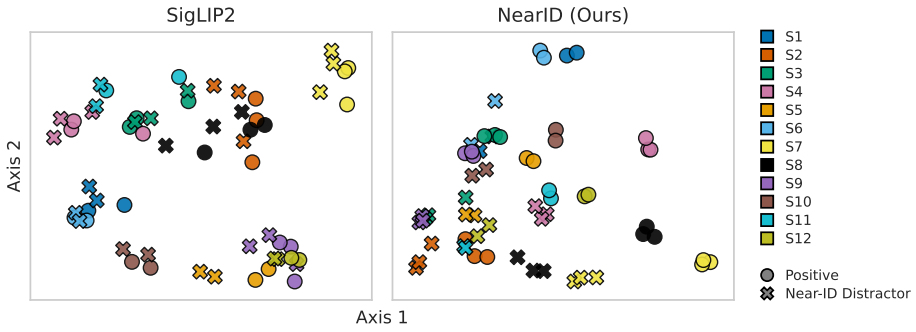


Fig. 6: KernelPCA visualization of Near-Identity separation. We project embeddings for n=12 identities (S1–S12) into 2D using identical KernelPCA [54] settings in both panels. Circles denote positives (same identity) and crosses denote matched-context NearID distractors (negatives). Compared to the frozen SigLIP2 baseline, NearID increases separation by pushing the distractors away from the corresponding positive clusters, providing visual evidence of improved near-identity discrimination.

trained on MTG. SSR and PA are pooled across distractor sources via support-weighted averaging.

As shown in Table 1, frozen encoders fail under matched-context evaluation: SigLIP2 achieves only 30.74% SSR. NearID resolves this, reaching 99.17% SSR and 99.71% PA. On MTG, all standard encoders score 0.0% SSR; even VSM reaches only 7.0%. NearID achieves 35.0% SSR with substantially improved oracle alignment (M–O: 0.465 vs. 0.180). On DB++ (Table 2), NearID improves metric-to-human correlation to 0.545 vs. 0.516 for SigLIP2. As shown in Figure 5, these gains generalize across DB++ categories, including animals and humans absent from training, confirming that identity-specific training transfers to real-world evaluation. Figure 3 right pane summarizes the resulting property profile across methods.

4.3 Visualization of Embeddings

Figure 6 shows KernelPCA [54] projections of NearID validation embeddings. The frozen baseline (left) entangles distractors with positives, while NearID (right) separates them into distinct clusters, with positives grouped but individually discriminable.

4.4 Training Objective Ablation

We train the same MAP head on identical data with different loss functions (Table 3); only the loss differs.

Standard contrastive training is insufficient. InfoNCE improves SSR from 30.74% to 60.97% and achieves the highest M–H (0.555), but 61% SSR means near-identity distractors still frequently outscore true positives.

Hierarchy is necessary but collapse is a risk. Circle + Ranking achieves 99.97% NearID SSR but *collapses* to M–H $= 0.141$, well below the 0.516 frozen baseline. Oracle Ranking similarly collapses (M–H $= 0.167$). These aggressive formulations over-specialize the embedding, motivating the softer softplus formulation in $\mathcal{L}_{\text{rank}}$.

NearID loss balances discrimination and alignment. $\mathcal{L}_{\text{NearID}}$ attains 99.17% SSR while maintaining M–H $= 0.545$, trading only 0.010 M–H relative to InfoNCE (which fails at 61% SSR).

Adding positive cohesion yields marginal gains (99.31% SSR) with comparable MTG SSR (36.0% vs 35.0%), confirming that the base $\mathcal{L}_{\text{NearID}}$ suffices for the primary setting.

5 Conclusion

We presented **NearID**, a unified framework for learning identity-discriminative representations under matched-context confounders. We showed that widely used vision encoders systematically conflate object identity with background context: under our evaluation protocol, even the best frozen encoder achieves only 30.74% SSR when near-identity distractors share the reference background.

To address this, we introduced three tightly coupled components: (i) a large-scale dataset with 19K identities and 316K+ matched-context distractors from four generative models, (ii) a two-component contrastive objective that enforces a strict similarity hierarchy without hard margins or oracle supervision, and (iii) a discriminability-based evaluation protocol grounded in similarity margins. Training only a lightweight MAP head on a frozen backbone, NearID achieves 99.17% SSR, improves part-level discrimination on MTG, and increases alignment with human judgments on DreamBench++.

The NearID dataset and evaluation protocol can serve as a diagnostic tool for any vision encoder, providing a rigorous test of whether learned representations capture genuine identity or merely exploit contextual shortcuts.

Limitations. Our scope targets common rigid objects rather than specialized domains such as faces or industrial inspection; distractor realism is bounded by the fidelity of the inpainting models used to synthesize them; and NearID is never trained on stylized data, leaving style-conditioned identity to future work.

We hope NearID establishes a stronger standard for evaluating and learning identity-aware representations in personalized generation and image editing.

Acknowledgements

This work was supported by the King Abdullah University of Science and Technology (KAUST) Center of Excellence for Generative AI (award 5940) and Snap Research. We thank Rapidata.ai for API access.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning. p. 89–96. ICML '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1102351.1102363, https://doi.org/10.1145/1102351.1102363
5. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning. p. 129–136. ICML '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1273496.1273513, https://doi.org/10.1145/1273496.1273513
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chen, H., Zuo, Z., Zhao, L., Li, J., Yang, J.: Conceptcraft: One-shot personalized text-to-image generation via object-background disentanglement. IEEE Transactions on Circuits and Systems for Video Technology **36**, 133–146 (2026), https://api.semanticscholar.org/CorpusID:280589664
8. Chen, N., Huang, M., Chen, Z., Zheng, Y., Zhang, L., Mao, Z.: Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2123–2131 (2025)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
10. Cvejic, A., Eldesokey, A., Wonka, P.: Partedit: Fine-grained image editing using pre-trained diffusion models. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
11. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023), https://arxiv.org/abs/2305.06500
12. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects (2022), https://arxiv.org/abs/2212.08051
13. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)

14. Eldesokey, A., Cvejić, A., Ghanem, B., Wonka, P.: Mind-the-glitch: Visual correspondence for detecting inconsistencies in subject-driven generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
15. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19358–19369 (2023)
16. Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika **10**(4), 507–521 (1915)
17. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
18. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
19. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
20. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
22. Hochberg, D.C., Anschel, O., Shoshan, A., Kviatkovsky, I., Aggarwal, M., Medioni, G.: Towards quantitative evaluation metrics for image editing approaches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7892–7900 (2024)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
24. Hu, J., Han, T., Ma, K., Gao, J., Yang, S., He, X., Luo, J., Wei, X., Zhang, W.: Positionic: Unified position and identity consistency for image customization. arXiv preprint arXiv:2507.13861 (2025)
25. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)
26. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
27. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: European Conference on Computer Vision. pp. 150–168. Springer (2024)
28. Keramati, M., Meng, L., Evans, R.D.: Conr: Contrastive regularizer for deep imbalanced regression. arXiv preprint arXiv:2309.06651 (2023)
29. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)

30. Kilrain, C., Carlyn, D., Chae, J., Beery, S., Chao, W.L., Gu, J.: Finer-personalization rank: Fine-grained retrieval examines identity preservation for personalized generation. arXiv preprint arXiv:2512.19026 (2025)
31. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18729–18738 (2022), https://api.semanticscholar.org/CorpusID:247940176
32. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643
33. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12268–12290 (2024)
34. Kumari, N., Yin, X., Zhu, J.Y., Misra, I., Azadi, S.: Generating multi-image synthetic data for text-to-image customization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16524–16534 (2025)
35. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
36. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36**, 30146–30166 (2023)
37. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/li23q.html
38. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
39. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training (2022), https://arxiv.org/abs/2112.03857
40. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), https://arxiv.org/abs/2304.08485
41. Mao, Z., Huang, M., Ding, F., Liu, M., He, Q., Zhang, Y.: Realcustom++: Representing images as real textual word for real-time customization. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
42. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment (2021), https://arxiv.org/abs/2103.06627
43. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
44. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
45. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
46. Papantoniou, F.P., Lattas, A., Moschoglou, S., Deng, J., Kainz, B., Zafeiriou, S.: Arc2face: A foundation model for id-consistent human faces. In: European Conference on Computer Vision. pp. 241–261. Springer (2024)

47. Pearson, K.: Vii. note on regression and inheritance in the case of two parents. proceedings of the royal society of London **58**(347-352), 240–242 (1895)
48. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. arXiv preprint arXiv:2406.16855 (2024)
49. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
50. Qian, G., Wang, K.C., Patashnik, O., Heravi, N., Ostashev, D., Tulyakov, S., Cohen-Or, D., Aberman, K.: Omni-id: Holistic identity representation designed for generative tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8786–8795 (2025)
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
52. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
53. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
54. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural computation **10**(5), 1299–1319 (1998)
55. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
56. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3 (2025), https://arxiv.org/abs/2508.10104
57. Singhania, A., Malani, K., Dhawan, R., Jain, A., Tandon, G., Sharma, N., Chakraborty, S., Batra, V., Phogat, A.: Beyond the pixels: Vlm-based evaluation of identity preservation in reference-guided synthesis (2025), https://arxiv.org/abs/2511.08087
58. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 1857–1865. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
59. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14398–14409 (2024)
60. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6398–6407 (2020)

61. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want (2023), https://arxiv.org/abs/2312.03818
62. Sundaram, S., Fu, S., Muttenthaler, L., Tamir, N., Chai, L., Kornblith, S., Darrell, T., Isola, P.: When does perceptual alignment benefit vision representations? Advances in Neural Information Processing Systems **37**, 55314–55341 (2024)
63. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
64. Thoma, J., Paudel, D.P., Gool, L.V.: Soft contrastive learning for visual localization. Advances in neural information processing systems **33**, 11119–11130 (2020)
65. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. Advances in Neural Information Processing Systems **36**, 48382–48402 (2023)
66. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
67. Vo, A., Nguyen, K.N., Taesiri, M.R., Dang, V.T., Nguyen, A.T., Kim, D.: Vision language models are biased. arXiv preprint arXiv:2505.23941 (2025)
68. Wang, Q., Cvejic, A., Eldesokey, A., Wonka, P.: Editclip: Representation learning for image editing (2025), https://arxiv.org/abs/2503.20318
69. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
70. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models (2024), https://arxiv.org/abs/2311.03079
71. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5022–5030 (2019)
72. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5022–5030 (2019)
73. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders (2023), https://arxiv.org/abs/2301.00808
74. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
75. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18682–18692 (2025)
76. Xia, F., Liu, T.Y., Wang, J., Zhang, W., Li, H.: Listwise approach to learning to rank: theory and algorithm. In: Proceedings of the 25th international conference on Machine learning. pp. 1192–1199 (2008)
77. Yang, Z., Bastan, M., Zhu, X., Gray, D., Samaras, D.: Hierarchical proxy-based loss for deep metric learning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1859–1868 (2022)

78. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
79. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Trans. Mach. Learn. Res. **2022** (2022), https://api.semanticscholar.org/CorpusID:248512473
80. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 694–699 (2002)
81. Zha, K., Cao, P., Son, J., Yang, Y., Katabi, D.: Rank-n-contrast: Learning continuous representations for regression. Advances in Neural Information Processing Systems **36**, 17882–17903 (2023)
82. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12104–12113 (2022)
83. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), https://arxiv.org/abs/2303.15343
84. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
85. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v(ision) as a generalist evaluator for vision-language tasks (2023), https://arxiv.org/abs/2311.01361
86. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In: European Conference on Computer Vision. pp. 195–211. Springer (2024)