ID-Sim: An Identity-Focused Similarity Metric

Julia Chae[1,†] Nicholas Kolkin[2] Jui-Hsien Wang[2] Richard Zhang[2] Sara Beery[1,*] Cusuh Ham[2,*]

[1]MIT CSAIL [2]Adobe Research

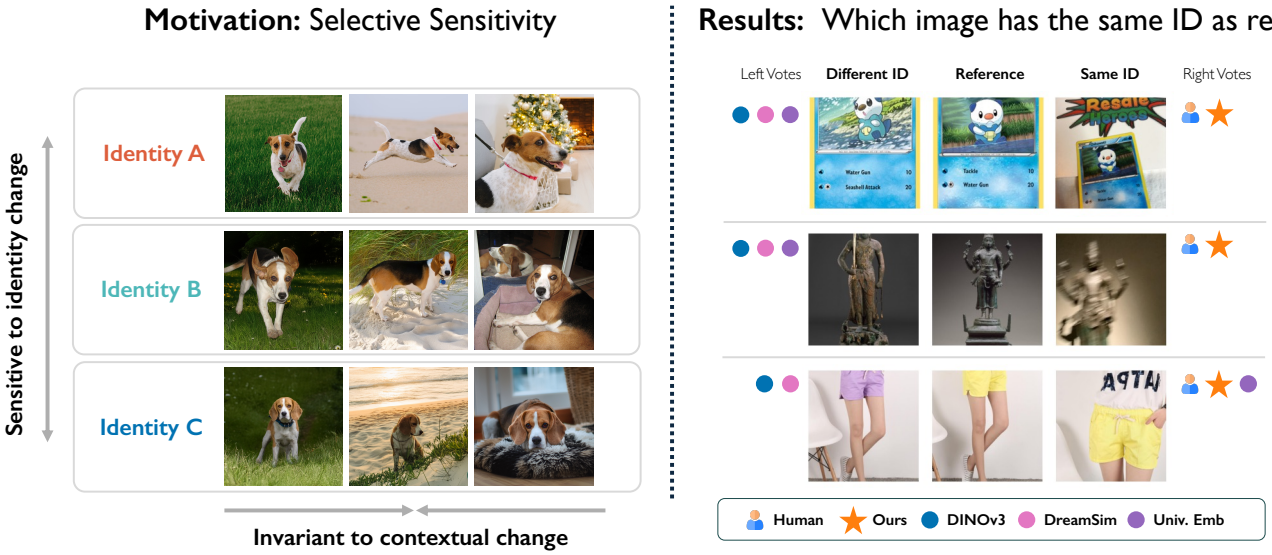

Figure 1. **ID-Sim motivation & results.** (Left) An identity-focused metric should exhibit *selective sensitivity:* invariant to contextual changes (e.g. background, pose, lighting), yet sensitive to subtle identity-altering changes. (Right) We present ID-SIM, which captures this property more effectively than existing metrics, and achieves strong improvements across a diverse set of identity-focused tasks.

Abstract

Humans have remarkable selective sensitivity to identities–easily distinguishing between highly similar identities, even across significantly different contexts such as diverse viewpoints or lighting. Vision models have struggled to match this capability, and progress towards identity-focused tasks such as personalized image generation is slowed by a lack of identity-focused evaluation metrics. To help facilitate progress, we propose **ID-SIM**, *a feed-forward metric designed to faithfully reflect human selective sensitivity. To build ID-Sim, we curate a high-quality training set of images spanning diverse real-world domains, augmented with generative synthetic data that provides controlled, fine-grained identity and contextual variations. We evaluate our metric on a new unified evaluation benchmark for assessing consistency with human annotations across identity-focused recognition, retrieval, and generative tasks. Our project page id here.*

1. Introduction

Humans readily recognize the same individual or object across large variations in viewpoint, illumination, pose, and context while remaining highly sensitive to subtle differences that signal identity changes [3, 9, 40, 49]. This balance, which we term *selective sensitivity*, enables both robust generalization and fine-grained discrimination–we recognize a familiar character from an unusual angle, identify a personal item under new lighting, or pick our own pet out of a crowd [3, 48, 55]. From a cognitive perspective, this corresponds to learning representations in which diverse appearances of the same identity cluster tightly while distinct identities remain well separated [9, 40].

Existing "identity"- or "instance"-focused works in computer vision employ widely varying definitions of what this means, from broad semantic categories (e.g., cities or product types) to unique physical objects. To reduce ambiguity, we adopt a specific, property-based definition for this work. We first define the concept of *visual identity*, and then use it to define an *instance*.

*Equal advising, randomly ordered.
†Work done while at Adobe as an intern.

Visual identity: An object's unique set of intrinsic visual properties (e.g., shape, texture, color).
Instance: Objects sharing the same visual identity.

Despite remarkable progress in visual representation learning, vision systems still struggle with identity-focused tasks. Even foundation models trained on massive datasets [29, 54, 68] fail to recognize the same object under moderate transformations (e.g., changes in viewpoint or illumination) and confuse identities that share superficial visual features like the background (see examples in Figure 1). Specialized systems for instance retrieval [64], re-identification [1, 78, 100], or personalized evaluation [12, 51], address aspects of this challenge, but typically in narrow, domain-specific contexts. None provide a general measure of identity consistency that captures when a transformation preserves, versus alters, an identity.

Historically, advances in perceptual metrics have catalyzed progress in computer vision. The shift from signal-based measures (PSNR and SSIM [84]) to learned perceptual metrics like LPIPS [98] and DISTS [10] transformed how visual similarity is quantified, enabling models that better align with human judgments of appearance. However, these metrics are focused on *appearance similarity*, not *identity*. To catalyze progress on identity-focused tasks, we propose a new perceptual metric that explicitly prioritizes selective sensitivity. We curate a diverse, instance-level training dataset that unifies and extends existing benchmarks across domains, augmented with a generative editing pipeline for controlled identity-preserving and identity-altering transformations. We train our model using complementary global and local contrastive objectives to balance invariance and discrimination, and evaluate and analyze our metric across diverse identity-focused tasks.
Our main contributions are:

- **A new identity-focused perceptual metric ID-Sim,** trained to mimic human selective sensitivity via curated real and synthetic instance-level data.
- A **comprehensive benchmark for identity perception**, combining existing instance-level tasks across domains with a new human-annotated generative evaluation dataset (SUBJECTS2K).
- A **systematic sensitivity analysis** using controlled generative edits, revealing the influence of viewpoint, lighting, and contextual changes on perceived identity consistency.

2. Related Works

2.1. Identity-focused tasks

Re-identification (Re-ID) aims to identify the same individual across contexts [60, 100]. Deep metric models for Re-ID are: (i) highly specialized to specific domains (e.g., animals [1, 47, 61], humans [8, 23, 38, 62, 70, 72, 81]), with models trained on domain X failing on domain Y [32, 47, 63, 93], (ii) require extensive domain-specific fine-grained annotations, and (iii) optimize for discrimination (maximizing inter-class margins) rather than perceptual alignment (matching human similarity judgments).

Instance retrieval entails finding matches to an example object from within a large candidate pool [6, 101]. Recent works like UnED [95] and GPR-1200 [59] have pushed towards generalizing instance retrieval across categories, from products to landmarks. Many prominent models train on data that conflate fine-grained classification with instance identity, which may limit their ability to differentiate two visually similar but distinct objects, as observed in Figure 1. Related work [89] explores training an instance-retrieval representation using generative edited data. While the approach is promising, the model is evaluated only on retrieval benchmarks and not on a broader range of identity-focused tasks.

Personalized vision works [30, 58, 73] adapt large models to a user-specified concept for tasks like subject-driven generation [18, 21, 24, 36, 56, 76, 92] or personalized segmentation [99]. Personalized generation faces a core challenge with identity fidelity, as models often struggle to faithfully preserve a subject's unique features. This failure of preservation also makes robust evaluation hard, creating a clear need for an approach that can reliably measure fine-grained identity similarity. Tasks like personalized segmentation (e.g., PerSAM [99]) pursue different goals, such as producing a pixel-level mask for the target subject, rather than quantifying its identity consistency.

2.2. Visual similarity metrics

Perceptual metrics. SSIM [84], PSNR [26], and other classical perceptual metrics [57, 83, 97] are hand-designed, and often fail to capture the complex nuances of human perceptual similarity [98]. Alternatively, learning-based methods (e.g., LPIPS [98], PieAPP [52], DreamSim [17], DISTS [10]) show that embeddings from deep networks [35, 67] can be calibrated or trained on perceptual judgments, and even align well with human perceptual judgments [98]. This observation extends to other modalities, such as stereo [75] and audio [41]. DiffSim [71] has also found that diffusion model features align well with human judgments of perceptual similarity. Since these metrics optimize for *overall* similarity rather than identity consistency, they are influenced by contextual changes that are irrelevant for identity-focused tasks.

Contrastive representations. The distance between contrastive representations is often used to quantify visual similarity. Vision models trained with self-supervised contrastive objectives [5, 20, 22, 25, 43, 77, 90] learn by attracting representations of augmented views of the same image

and repelling those of different images. Thus, the representations capture the broad semantics of an image while ignoring the effects of transformations used as positive augmentations. For example, SimCLR [5] and MoCo [22] use cropping, color jittering, and blurring, encouraging invariance to low-level global changes. Similarly, the DINO model family [4, 45, 68] and CLIP [54] apply contrastive learning at scale, with CLIP aligning images to text—often compressing fine-grained visual differences in favor of higher-level semantic similarity.

Applications of visual similarity metrics. Metrics aligned with human perception have been shown to benefit downstream tasks like segmentation and instance retrieval [74]. As mentioned above, another primary application is evaluating subject-driven generation, where identity fidelity is crucial. However, general perceptual metrics are often insufficient, as they can confuse high visual similarity (e.g., two similar purses in the same pose) with true identity preservation (e.g., the same purse in a different pose). MLLMs (e.g., GPT-4V [44, 50]) are also used and align well with human judgments, but they face issues with prompt sensitivity, stochasticity, and scalability [65]. This necessitates an efficient metric focused on instance-level identity. Concurrent work also proposed specialized metrics for detecting generative inconsistencies [12], but may falter under occlusion or lighting changes.

3. Methods

3.1. Characterizing our definition of an instance

Under our definitions, two images depict the *same instance* when they show visually indistinguishable objects, such as two factory-identical screwdrivers, even when these objects are transformed by *extrinsic* variations (e.g., pose, viewpoint, or lighting). Conversely, two images depict *different instances* if their visual identities differ, including clearly different objects, significant temporal changes (e.g., a kitten aging to a cat), and physical alterations (e.g., a repainted chair).

3.2. Training data curation

To train a metric that mimics selective sensitivity, we need data with three complementary signals:

- *Context diversity* supporting invariance to different backgrounds, lighting, and viewpoints
- *Visual identity diversity* enabling sensitivity to subtle appearance differences
- *Domain diversity* ensuring generalization beyond specific categories

No existing datasets provide all three simultaneously, so we curate a training set using: **(Subset 1)** existing real instance-level datasets, and **(Subset 2)** synthetic data with:

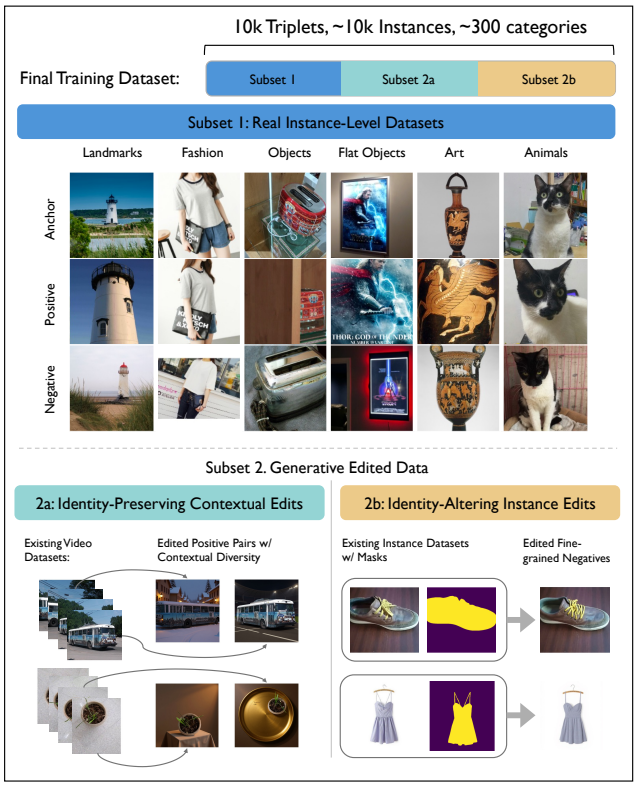

Figure 2. **Dataset curation pipeline.** We highlight the different real and synthetic data subsets that enable ID-Sim training. Together, they provide high context, domain, and visual identity diversities.

Dataset	Type	Objects	#Cat	Included in	#Inst
ILIAS [33]	Img	General	N/A	S1	281
FORB [87]	Img	Flat Obj	7	S1	761
MET [94]	Img	Artworks	1	S1	226
GLDv2 [85]	Img	Landmarks	12	S1	769
Dogs [42]	Img	Animal	1	S1	494
Cats [80]	Img	Animal	1	S1	140
DF2 [19]	Img	Fashion	13	S1, S2b	2466
UCO3D [39]	Vid	General	146	S2a, S2b	3884
LASOT [15]	Vid	General	34	S2a	101
YouTubeVIS [91]	Vid	General	35	S2a	414
GOT10k [28]	Vid	General	72	S2a	604

Table 1. **Overview of datasets used for training set curation.** Cat and Inst refer to number of categories and instances respectively. Colored text indicates different subsets that the dataset images appear in: S1, S2a, S2b, which can be matched to Figure 2

(a) *contextual edits* that diversify the contexts in which instances appear, and **(b)** *identity edits* that perturb visual identity (see Figure 2). These generative edits (S2a and S2b) expand the training pool, alleviating the limited diversity of real-world data, which is difficult to collect and annotate at scale.

We formulate our training data using triplets (an anchor image, a positive ID match, and a negative non-match), a standard structure for learning similarity metrics. Pos-

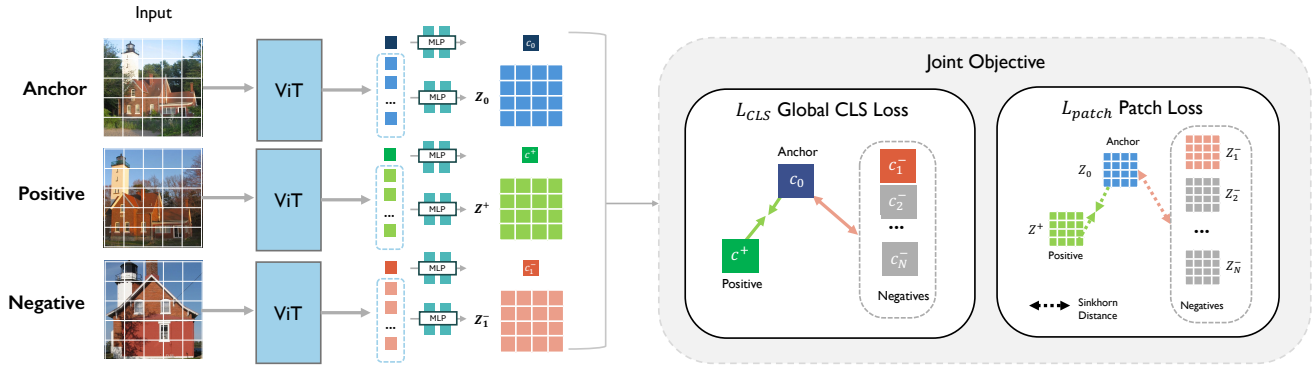

Figure 3. **ID-Sim training pipeline.** We train our metric with dual contrastive supervision. At the global level, CLS-token projections for anchor–positive pairs are contrasted against one hard negative and additional batch negatives using InfoNCE. At the patch level, projected patch tokens are compared using Sinkhorn distance for the same instance pairs.

itives come from real instance images (S1) or identity-preserving contextual edits (S2a), while negatives come from different real instances (S1) or identity-altering edits (S2b). Our training set contains 10k triplets (30k images) spanning ∼10k instances across 10 datasets, with an even split between triplets containing only real images, generative identity-preserving positives with real negatives, and real positives with identity-altering negatives. Table 1 provides an overview of the dataset composition. We analyze the effects of dataset scale and composition in Section 4.4, and include additional experiments and full details on the source datasets, splits, and editing pipelines in the Supplemental.

3.3. ID-Sim Training

Data formulation for contrastive learning. As seen in Figure 3, we follow the supervised contrastive learning framework [31], training our metric with positive (identity-preserving) pairs and negative (identity-breaking) pairs. We build our training batches from the dataset $\mathcal{D}$ introduced in Section 4.2 comprised of M instances $\{X_j\}_j^M$, and use these instances to curate triplets $(x_0, x^+, \{x_i^-\}_{i=1}^N)$ for training. Two images from the same instance are sampled as the anchor x_0 and positive x^+. We then sample a hard negative x_1^-, which may be either an identity-altering edit from S2b or a mined real negative from S1. For real negatives, we mine visually similar yet distinct instances using the nearest neighbors in the pretrained DINOv3 embedding space [68]. The remaining $N-1$ negatives $\{x_i^-\}_{i=2}^N$ are sampled from other instances within the batch.

Joint objective. We build upon a vision transformer (ViT) [11] backbone f_θ, following recent works [29, 46, 53]. Each image is passed through f_θ to obtain the global CLS token c' and a set of patch tokens Z'. Since these representations capture complementary global and local information, we project them into separate embedding spaces using a dual-headed MLP: $c = \mathrm{MLP}_{\mathrm{CLS}}(c')$ and $Z = \mathrm{MLP}_{\mathrm{Patch}}(z')$. We train using a joint supervised contrastive objective that combines global and local terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathrm{CLS}}(c) \; + \; \lambda \, \mathcal{L}_{\mathrm{Patch}}(Z), \tag{1}$$

We opt for this joint objective instead of supervising only on the global token since patch embeddings provide complementary spatial signals for dense downstream tasks.

1. Global CLS Loss. $\mathcal{L}_{\mathrm{CLS}}$ is the standard InfoNCE objective [79] applied to the projected CLS tokens:

$$\mathcal{L}_{\mathrm{CLS}} = -\log \frac{e^{\,s^+}}{e^{\,s^+} + \sum_{i=1}^{N} e^{\,s_i^-}}, \qquad s^+ = \mathrm{sim}(c_0, c^+)/\tau, \quad s_i^- = \mathrm{sim}(c_0, c_i^-)/\tau \tag{2}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity, τ is a temperature parameter, and c_0, c^+, c_i^- are the projected CLS tokens for the anchor, positive, and i-th negative, respectively.

2. Local patch loss. Patch tokens encode fine-grained local cues, but the spatial layouts of instances across images are often misaligned due to viewpoint or context changes, making direct position-wise comparisons unreliable. We therefore treat patch tokens between two images as an unordered set of local descriptors and measure their similarity via soft alignment. Given projected patch embeddings $A, B \in \mathbb{R}^{P\times D}$, we define their similarity as the negative entropically regularized optimal transport (OT) distance:

$$\mathrm{sim}_{\mathrm{patch}}(A, B) = -\,\mathcal{S}_\varepsilon(A, B), \tag{3}$$

where $\mathcal{S}_\varepsilon$ is the Sinkhorn distance computed with uniform weights over patches, using GeomLoss [16].

Unlike DenseCL [14], which builds hard nearest-neighbor correspondences using augmented views of the same image, our objective operates across different images of the same instance, learning correspondences implicitly through a soft global OT plan. $\mathcal{L}_{\mathrm{Patch}}$ is obtained by substi-

tuting $\text{sim}_{\text{patch}}(\cdot,\cdot)$ into the InfoNCE objective.

Using representations as an image similarity metric. Using a trained f_θ, similarity between images x and y can be measured as:

$$D(x, y; f_\theta) = 1 - \text{sim}\big(f_\theta(x),\ f_\theta(y)\big), \tag{4}$$

where the choice of feature representation $f_\theta(\cdot)$ and similarity function $sim(\cdot,\cdot)$ can vary. Since ID-Sim is a ViT-based metric, common features include the global CLS token or various patch token representations (e.g., aggregated or localized sets). The similarity function is typically cosine similarity. We explore alternative combinations of feature and similarity functions, which can enable different types of downstream tasks, in Section 4.4.

4. Experiments

We evaluate ID-Sim against 7 baselines across instance *recognition*, *retrieval*, and *preservation* tasks on 7 datasets, all disjoint from the training set.

4.1. Experimental setup

Network architecture and training details. We select DINOv3 ViT-L [68] at 448×448 resolution as the backbone f_θ, chosen for strong instance-level performance on our validation set (described below). We freeze the backbone and finetune only: (i) lightweight 2-layer dual MLP projection heads, and (ii) rank 16 LoRA adapters [27] on attention and feedforward MLP layers. Training uses standard augmentations (color jitter, Gaussian noise, random cropping).

Hyperparameter tuning and checkpoint selection are performed on a held-out validation set drawn from the training data domains. We also construct an "identity ablation set", a small Flux-generated [37] synthetic dataset of 5 instances with identity-preserving and identity-altering edits. Full training details and ablation studies are in the Supplemental.

Baselines. We test 7 baselines in three categories: (1) *perceptual metrics* (DreamSim [17], LPIPS [98], DiffSim [71]), (2) *foundation models* (DINOv3 [68], CLIP [54], OpenCLIP [29]), and (3) an *image retrieval model* – the 1st-place solution [64] from Google's Universal Embedding (UNED) challenge [95]. All models use the ViT-L architecture except for [64] (larger ViT-H), DreamSim (ViT-B), and DiffSim (U-Net).

4.2. Benchmarks

1. Concept preservation evaluation aims to quantify how well a model is able to generate images of a reference instance while preserving its visual appearance. We evaluate this using two benchmarks.

First, we report Spearman's ρ correlation against human judgments on DreamBench++ [51], a public benchmark for subject-driven generation. However, we found its human preference labels to be noisy, stemming from sparse annotations (see Supplemental). Thus, we introduce SUBJECTS2K, a new human-annotated subset of Subjects200k [102]. We collected new binary (same/different instance) human annotations to improve and evaluate the original dataset's GPT-4v [44] labels. On SUBJECTS2K, we report average precision (AP).

Figure 5. **Newly annotated Subjects2k.** We release a 2k high-quality human annotations with a subset of Subjects200k to serve as a new challenging concept preservation eval benchmark.

2. Instance retrieval tests the ability to find images of a given reference object from a pool of distractors. We report mean AP (mAP), averaged across each instance in the datasets on: (a) PODS [73], a dataset of household objects for instance-level retrieval and recognition under fixed distribution shifts, and (b) DeepFashion2 [19], a fashion dataset designed to match in-store clothing items to in-the-wild consumer images.

3. Re-identification (Re-ID) / instance classification assesses whether individuals can be consistently recognized across viewpoints and conditions. We evaluate using: (a) mAP on PetFace [66], a multi-species pet re-ID dataset, (b) mAP on AerialCattle [2], consisting of 23 individual cattle captured from aerial viewpoints, and, following the protocol from DiffSim, (c) accuracy on CUTE [34], where the model must identify which instance out of a pair of candidates matches an anchor object.

We also evaluate results on additional metrics (e.g. AUROC or NDCG for ranking [82]) for all tasks in the Supplemental.

4.3. Results

Improved identity-alignment across tasks. We evaluate ID-Sim across diverse domains and task types (Figure 4), using the global CLS token for similarity computation across all ViT-based methods. Across 49 evaluation setups, ID-Sim outperforms prior work in 48 cases.

The strongest gains emerge along two axes of selective sensitivity: (1) *Recognizing instances across contextual changes*, and (2) *discriminating small visual identity changes*. This challenge of (1) is prominent in datasets such as PODS and DeepFashion2, where in addition to requiring fine-grained discrimination, positive instances are explicitly observed in different contexts (background, pose, and distractors in PODS; in-store vs in-the-wild for Deep-

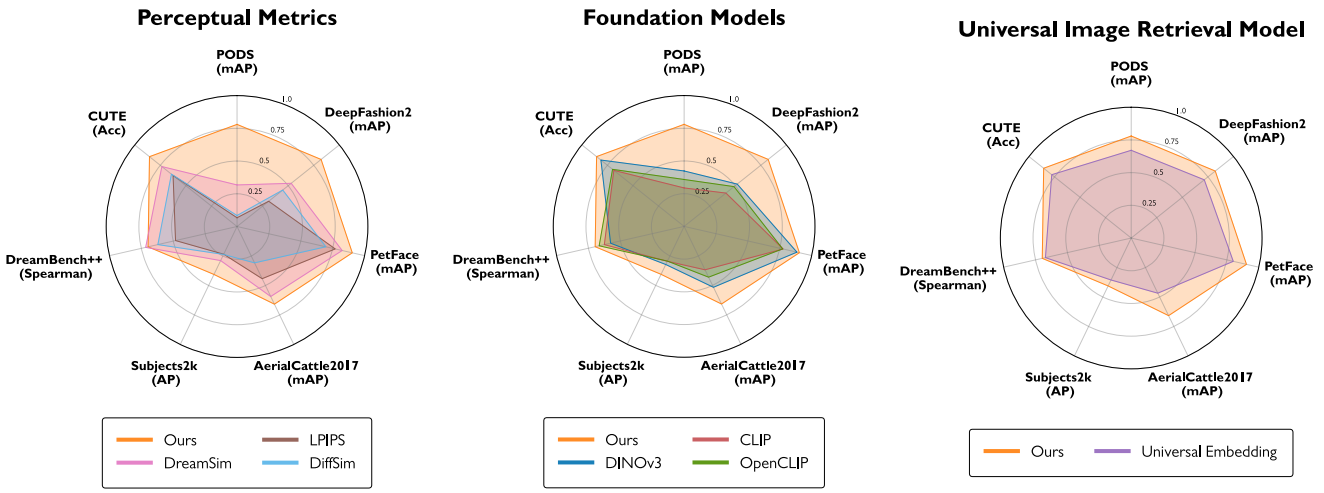

Figure 4. **Performance of ID-Sim vs. baseline models**. We compare ID-Sim against standard perceptual metrics, large-scale vision foundation models, and a supervised "Universal Embedding" model (the top entry in Google's universal embedding challenge). Across tasks – instance retrieval, concept preservation, and re-identification – ID-Sim consistently outperforms all baselines, including the instance-retrieval-focused model, despite using over 100× less labeled data and a smaller backbone (our VIT-L vs. VIT-H). Full results and seed variance are reported in the Supplemental.

Method	Model	Subjects2k (AP)	DreamBench
Ours	ViT-L	**0.4063**	0.697
MLLM*	GPT-4o	0.2901	**0.748**
MLLM	GPT-5	0.3159	0.3554
MLLM	Gemini	0.3354	0.70

(a) **Comparison with MLLMs on concept preservation.** MLLM* uses the original Subjects200K and DreamBench++ prompts and models respectively; MLLM rows use a controlled identity-preservation prompt for both datasets.

Dataset	Metric	DINOv3	Ours (no patch)	Ours
DF2	mAP	0.4071	0.4765	**0.7967**
AC2017	mAP	0.4516	0.5471	**0.6245**
CUTE	Acc	0.6561	0.6439	**0.8189**
DB++	Spearman	0.5479	0.5913	**0.6834**
PetFace	mAP	0.7849	0.8377	**0.8446**
PODS	mAP	0.5825	**0.8181**	0.7907
S2k	AP	0.2314	0.2348	**0.3674**

(b) **Patch-level Performance** of ID-Sim, ID-Sim without patch supervision, and DINOv3 across tasks

Method	mAP	F1
PerSAM + DINOv3	0.153	0.18
PerSAM + Ours w/o patch sup	0.214	0.235
PerSAM + Ours	**0.436**	**0.409**

(c) **Personalized segmentation (PerSAM)** performance on PODS with varying metrics.

Table 2. **Overview of results.** (Left): comparison with MLLMs on concept preservation. (Middle): performance across recognition and retrieval datasets. (Right): transfer to personalized segmentation with PerSAM.

Fashion2). With ID-Sim, we see some of the strongest relative improvements in these cases, with $+0.11$ and $+0.30$ gains in mAP over the second-best and the third-best models for both cases. For (2), the Subjects2k benchmark presents some of the most challenging examples of fine-grained identity variation across datasets, with hundreds of visually similar negative instance pairs distinguished only by subtle details. On this benchmark, ID-Sim outperforms the second-best metric by $+0.05$ mAP.

Comparing metrics. Across baselines, clear trends emerge in the strengths and limitations. Perceptual metrics generally underperform on identity-focused tasks, as they capture perceptual similarity rather than identity discrimination (though DreamSim performs best on DreamBench++, consistent with its human-aligned objective). Foundation models like DINOv3 perform well on datasets like CUTE and PetFace that primarily test identity similarity under lighting variations, but struggle to maintain identity similarity under other context shifts such as background variation, and also struggle with retrieval tasks. The Universal Embedding model achieves the second-strongest overall performance, but benefits from a larger backbone (ViT-H) and millions of labelled instance-level and fine-grained examples. ID-Sim delivers consistently strong performance across all datasets, indicating broader generalization and a more unified notion of identity-alignment.

Comparison to MLLMs for concept preservation. Multimodal LLMs (MLLMs) have shown strong potential for identity-based evaluation, often aligning more closely with humans than DINO or CLIP [51]. Therefore, we compare ID-Sim against MLLMs using structured evaluation protocols consistent with prior work as shown in Table 2a. As shown, ID-Sim performs competitively and even surpasses MLLMs on Subjects2k, our more fine-grained concept-

Dataset Group	Bal.	Pos. Edit	Neg. Edit	Ratio	Val Score
All datasets	✗	✗	✗	–	0.693
All datasets	✓	✗	✗	–	0.752
Filtered datasets	✓	✗	✗	–	0.890
Filtered datasets	✓	✓	✗	1:1	0.937
Filtered datasets	✓	✓	✓	1:1:1	**0.965**

Table 3. **Ablation of dataset composition and editing strategies.** Balancing and targeted editing of positive and negative samples improve performance.

preservation benchmark. Notably, MLLM performance is sensitive to prompt and model choice: DreamBench++ accuracy drops substantially when its original rubric-guided prompts are replaced with controlled identity-preservation prompts, whereas ID-Sim remains stable across evaluations. MLLMs also introduce practical limitations, including stochastic outputs and reliance on pairwise comparisons that increase cost at scale, which is challenging for tasks like retrieval. In contrast, ID-Sim provides deterministic, feed-forward evaluations that match or exceed MLLM performance with significantly lower computational overhead. Full prompting details and MLLM evaluation settings are provided in the Supplemental.

Beyond global similarity: Patch-level embeddings and localization power. While the global CLS token used in Figure 4 captures a holistic representation, ViT patch tokens offer complementary, spatially localized features essential for fine-grained correspondence and region-level discrimination. We compare ID-Sim's patch embeddings against DINOv3 [68], the strongest baseline with well-established patch embeddings, and ablate patch-level supervision to assess its contribution.

Table 2b shows performance across tasks when similarity is computed using patch embeddings. ID-Sim significantly outperforms DINOv3 across all datasets, indicating that it learns stronger and more discriminative local representations. While the variant trained only with CLS supervision improves performance by 13% over DINOv3, explicit patch-level supervision substantially amplifies these gains, yielding a 40% relative improvement.

To further assess whether our patch embeddings encode spatially meaningful information, we evaluate ID-Sim within the state-of-the-art personalized segmentation framework, PerSAM [99], which uses patch-token similarity to localize SAM point prompts and score segmentation predictions. As shown in Table 2c, our patch features improve segmentation mAP significantly from 0.153 to 0.436 and F1 from 0.18 to 0.409 over DINOv3. Even without explicit patch supervision, ID-Sim features improve over DINOv3 (0.214 mAP, 0.235 F1). Our patch embeddings capture both aggregated and spatially coherent information for precise localization and discrimination of identities.

4.4. Analysis

What makes for the best training data? While developing ID-Sim, we systematically explored different strategies for curating and prioritizing high-value, identity-focused training data. Results are shown in Table 3, demonstrating that these choices significantly impact metric performance. We find that *balanced composition is crucial*. Ensuring balanced positive and negative samples prevents overfitting to dominant instances and leads to more stable convergence. Additionally, *dataset quality matters:* filtering out noisy or inconsistent instance-level samples significantly improves generalization. This matches prior literature–high-quality data is particularly vital for fine-grained tasks [7]. Finally, we find that *synthetic data boosts performance*: incorporating edited samples enhances both diversity and robustness—positive edits improve intra-instance consistency and edited negatives sharpen inter-instance discrimination.

Exploring sensitivity to visual variation. In order to isolate the visual factors that metrics are most sensitive to, we conduct a systematic sensitivity analysis measuring how similarity scores change with respect to four dimensions of variation: identity, background, viewpoint, and lighting. We use 100 diverse objects from MVImgNet [96], a multi-view dataset not used in training or evaluation, which provides 180 views per object on a clean surface with natural viewpoint variation. For the other dimensions, we apply generative edits: identity changes are simulated by editing the foreground with Qwen-Edit-Inpainting [86] (varying noise strengths), background changes via inpainting with 14 scene prompts, and lighting variations using Qwen-Edit [86] with nine illumination prompts. For each reference, we construct an edit grid varying jointly along identity and one other factor and compute the similarity of each image back to its original anchor. Sensitivity scores are then estimated by fitting a regression model to quantify the similarity decrease per unit change in each dimension. Final scores are averaged across instances, with uncertainty estimated via bootstrapped confidence intervals.

Figure 6 summarizes our sensitivity analysis across these four factors and shows that ID-Sim achieves the most desirable balance: high identity sensitivity and low contextual sensitivity. Performance of other metrics varies across these challenges. DreamSim exhibits moderate identity sensitivity but remains similarly sensitive to background and lighting variation. In contrast, the Universal Embedding model and DINOv3 show greater invariance to viewpoint and lighting but are more sensitive to background changes. CLIP, OpenCLIP, and LPIPS show the weakest identity sensitivity, measuring semantic or image-level similarity rather than identity similarity. Examining the similarity scores in the bottom row of Figure 6 offers a complementary perspective: compared to other metrics, ID-Sim exhibits the largest similarity drop in response to identity changes while main-

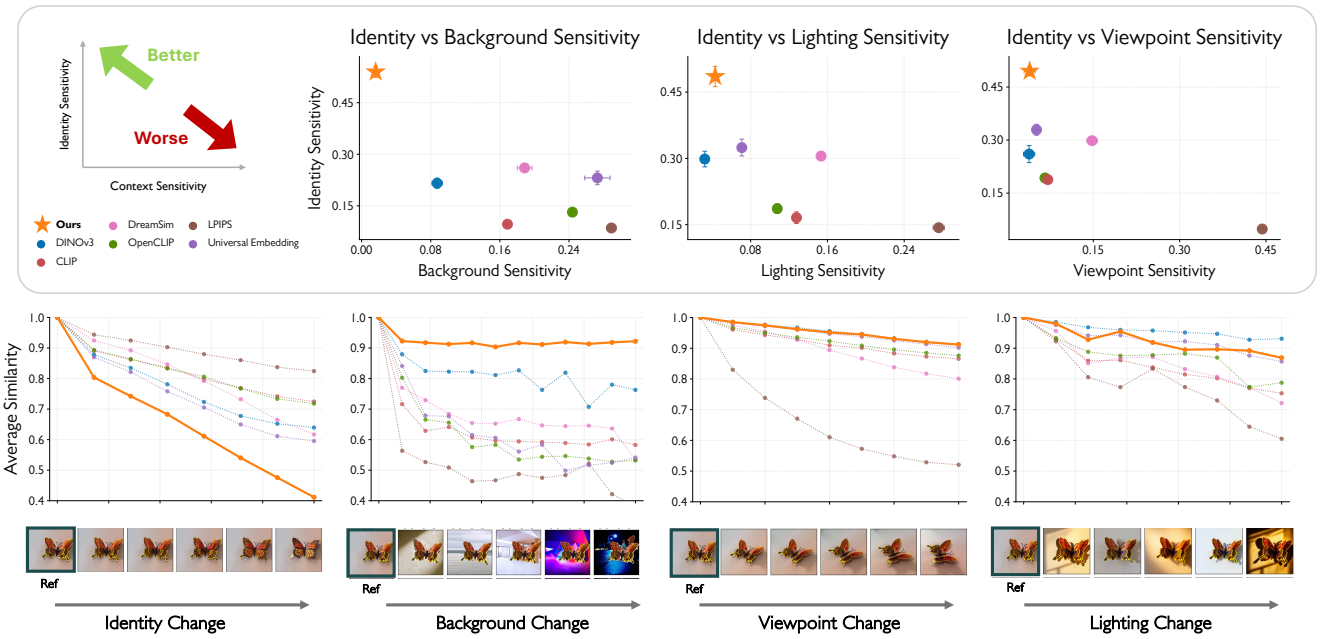

Figure 6. **Selective sensitivity analysis.** We evaluate model sensitivity across four axes of visual change: identity, background, viewpoint, and lighting. For 100 anchor instances, we generate controlled variations and compute both sensitivity scores and similarity trends. **(Top row.)** Compared with baseline methods, our model is notably more sensitive to identity differences while remaining stable under background, viewpoint, and lighting changes. **(Bottom row.)** When systematically increasing variations across each dimension, we see that, as desired, only identity changes significantly reduce similarity measured by ID-Sim.

taining invariance to other factors, supporting its stronger identity sensitivity. ID-Sim is slightly less robust to lighting variation than DINOv3, reflecting a tradeoff to preserve fine-grained color cues for identity.

5. Limitations, Future Work, and Conclusions

Limitations. Our instance definition relies on consistent visual identity and therefore does not fully capture broader notions of identity that may require user-specified invariances (e.g., aging, accessories, or stylistic changes). Also, ID-Sim is a global prompt-free metric and does not resolve the identity to target in multi-entity scenes; doing so requires external conditioning, either using spatial cues (e.g., masks) or text prompts, as explored by concurrent work Omni-Attribute [13]. We show in the Supplemental that our localized patch embeddings provide a natural foundation for more flexible, spatially-conditioned identity specification.

Future work. Recent work personalization works [69, 88] has used synthetic data to bootstrap training, improving generalization and reducing overfitting. However, automating this has been difficult and error-prone, lacking the general, selectively sensitive identity embeddings that our work (ID-Sim) introduces. We believe leveraging ID-Sim for this task is a promising direction. In addition, conditioning signals can be incorporated for selective identity specification.

Conclusions. Our results demonstrate that by combining a carefully curated dataset (Section 4.2) and training formulation (Section 3.3), it is possible to train a general purpose identity-focused similarity metric with state of the art performance across a wide variety of tasks, all at a fraction of the inference costs, training costs, and data requirements of MLLM foundation models. ID-Sim produces both global and local embeddings that can be easily plugged into any application that requires identity sensitivity and robustness to contextual changes (e.g., pose, background, lighting).

Acknowledgements This work was supported by an NSERC PGS-D, a Schmidt Science AI2050 Early Career Fellowship, NSF CAREER Award No. 2441060, the NSF and NSERC AI and Biodiversity Change Global Center (NSF Award No. 2330423 and NSERC Award No. 585136), the MIT Generative AI Consortium, and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Lukáš Adam, Vojtěch Čermák, Kostas Papafitsoros, and Lukas Picek. Wildlifereid-10k: Wildlife re-identification dataset with 10k individual animals. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2090–2100. IEEE, 2025. 2

[2] William Andrew, Jing Gao, Siobhan Mullan, Neill Campbell, Andrew W. Dowsey, and Tilo Burghardt. Visual identification of individual holstein-friesian cattle via deep metric learning. *Computers and Electronics in Agriculture*, 185: 106133, 2021. 5

[3] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94 (2):115, 1987. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[5] 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. 2, 3

[6] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7270–7292, 2022. 2

[7] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14755–14764, 2022. 7

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[9] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007. 1

[10] 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. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4

[12] Abdelrahman Eldesokey, Aleksandar Cvejic, Bernard Ghanem, and Peter Wonka. Mind-the-glitch: Visual correspondence for detecting inconsistencies in subject-driven generation, 2025. 2, 3

[13] T.S. Chen et al. Omni-attribute: Open-vocabulary attribute encoder for visual concept personalization, 2025. 8

[14] X. Wang et al. Dense contrastive learning for self-supervised visual pre-training, 2021. 4

[15] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, Yong Xu, Chunyuan Liao, Lin Yuan, and Haibin Ling. Lasot: A high-quality large-scale single object tracking benchmark, 2020. 3

[16] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019. 4

[17] 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. *arXiv preprint arXiv:2306.09344*, 2023. 2, 5

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[19] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345, 2019. 3, 5

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[21] Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhang, and Tobias Hinz. Personalized residuals for concept-driven text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8186–8195, 2024. 2

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3

[23] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person reidentification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8450–8459, 2019. 2

[24] Qiyuan He and Angela Yao. Conceptrol: Concept control of zero-shot personalized image generation. *arXiv preprint arXiv:2503.06568*, 2025. 2

[25] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua

Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[26] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 2

[27] 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, 2021. 5

[28] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. 3

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 2, 4, 5

[30] Yuxin Jiang, Yuchao Gu, Yiren Song, Ivor Tsang, and Mike Zheng Shou. Personalized vision via visual in-context learning, 2025. 2

[31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 4

[32] Inès Hyeonsu Kim, JoungBin Lee, Woojeong Jin, Soowon Son, Kyusun Cho, Junyoung Seo, Min-Seop Kwak, Seokju Cho, JeongYeol Baek, Byeongwon Lee, et al. Pose-dive: Pose-diversified augmentation with diffusion model for person re-identification. *arXiv preprint arXiv:2406.16042*, 2024. 2

[33] Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Suma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, and Giorgos Tolias. Ilias: Instance-level image retrieval at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14777–14787, 2025. 3

[34] Klemen Kotar, Stephen Tian, Hong-Xing Yu, Daniel L.K. Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. 2023. 5

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[36] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2

[37] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5

[38] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2

[39] Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. Uncommon objects in 3d. In *arXiv*, 2024. 3

[40] Nikos K Logothetis and Jon Pauls. Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral cortex*, 5(3):270–288, 1995. 1

[41] Pranay Manocha, Adam Finkelstein, Richard Zhang, Nicholas J Bryan, Gautham J Mysore, and Zeyu Jin. A differentiable perceptual audio metric learned from just noticeable differences. *arXiv preprint arXiv:2001.04460*, 2020. 2

[42] Guillaume Mougeot, Dewei Li, and Shuai Jia. A deep learning approach for dog face verification and recognition. In *PRICAI 2019: Trends in Artificial Intelligence*, pages 418–430, Cham, 2019. Springer International Publishing. 3

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[44] OpenAI. Gpt-4v (vision): Multimodal gpt-4 with image and text input. https://openai.com/research/gpt-4v-system-card, 2023. Accessed: 2025-11-13. 3, 5

[45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 4

[47] Lasha Otarashvili, Tamilselvan Subramanian, Jason Holmberg, JJ Levenson, and Charles V Stewart. Multispecies animal re-id using a large community-curated dataset. *arXiv preprint arXiv:2412.05602*, 2024. 2

[48] Thomas J Palmeri and Celina Blalock. The role of background knowledge in speeded perceptual categorization. *Cognition*, 77(2):B45–B57, 2000. 1

[49] Thomas J Palmeri and Isabel Gauthier. Visual object understanding. *Nature Reviews Neuroscience*, 5(4):291–303, 2004. 1

[50] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 3

[51] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In

The Thirteenth International Conference on Learning Representations, 2025. 2, 5, 6

[52] 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. 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4

[54] 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. 2, 3, 5

[55] Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104 (3):192, 1975. 1

[56] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 2

[57] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009. 2

[58] Dvir Samuel, Rami Ben-Ari, Matan Levy, Nir Darshan, and Gal Chechik. Where's waldo: Diffusion features for personalized segmentation and retrieval, 2024. 2

[59] Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Klaus Jung. Gpr1200: A benchmark for general-purpose content-based image retrieval. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*, page 205–216, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[60] Stefan Schneider, Graham W Taylor, Stefan Linquist, and Stefan C Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470, 2019. 2

[61] Stefan Schneider, Graham W Taylor, and Stefan C Kremer. Similarity learning networks for animal individual re-identification: an ecological perspective. *Mammalian Biology*, 102(3):899–914, 2022. 2

[62] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[63] Tom Shaked, Yuval Goldman, and Oran Shayer. Minimizing embedding distortion for robust out-of-distribution performance. *arXiv preprint arXiv:2409.07582*, 2024. 2

[64] Shihao Shao and Qinghua Cui. 1st solution in google universal image embedding. https://www.kaggle.com/datasets/louieshao/guieweights0732, 2023. 2, 5

[65] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*, 2024. 3

[66] Risa Shinoda and Kaede Shiohara. Petface: A large-scale dataset and benchmark for animal identification, 2024. 5

[67] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[68] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 2, 3, 4, 5, 7

[69] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 8

[70] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 719–728, 2019. 2

[71] 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. 2, 5

[72] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 2

[73] Shobhita Sundaram, Julia Chae, Yonglong Tian, Sara Beery, and Phillip Isola. Personalized representation from personalized generation, 2024. 2, 5

[74] Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Y. Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations?, 2024. 3

[75] Netanel Tamir, Shir Amir, Ranel Itzhaky, Noam Atia, Shobhita Sundaram, Stephanie Fu, Ron Sokolovsky, Phillip Isola, Tali Dekel, Richard Zhang, et al. What makes for a good stereoscopic image? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 261–272, 2025. 2

[76] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 14940–14950, 2025. 2

[77] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 2

[78] Tobias Trein and Luan Fonseca Garcia. Siamese networks for cat re-identification: Exploring neural models for cat instance recognition. *arXiv preprint arXiv:2501.02112*, 2025. 2

[79] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 4

[80] Vojtěch Čermák, Lukáš Picek, Lukáš Adam, and Kostas Papafitsoros. WildlifeDatasets: An Open-Source Toolkit for Animal Re-Identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5953–5963, 2024. 3

[81] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2

[82] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures, 2013. 5

[83] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, pages 1398–1402. Ieee, 2003. 2

[84] 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. 2

[85] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020. 3

[86] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 7

[87] Pengxiang Wu, Siman Wang, Kevin Dela Rosa, and Derek Hu. Forb: a flat object retrieval benchmark for universal image embedding. *Advances in Neural Information Processing Systems*, 36:25448–25460, 2023. 3

[88] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 8

[89] Yankun Wu, Zakaria Laskar, Giorgos Kordopatis-Zilos, Noa Garcia, and Giorgos Tolias. Instance-level generation for representation learning, 2025. 2

[90] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[91] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation, 2019. 3

[92] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[93] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 2

[94] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021. 3

[95] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovskỳ, Pelin Dogan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ondřej Chum, and André Araujo. Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 11290–11301, 2023. 2, 5

[96] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 7

[97] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 2

[98] 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. 2, 5

[99] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot, 2023. 2, 7

[100] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2

[101] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions*

on pattern analysis and machine intelligence, 40(5):1224–1244, 2017. 2

[102] Xingyi Yang Qiaochu Xue Zhenxiong Tan, Songhua Liu and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 5