A TEXT-PARAMETRIZED METRIC FOR IMAGE SIMILARITY

Anonymous authors
Paper under double-blind review

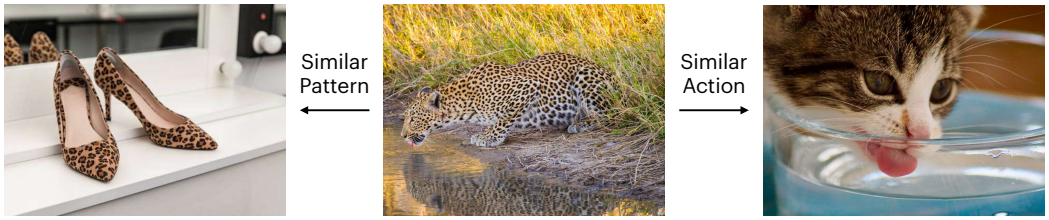


Figure 1: The same reference can be more similar to different images depending on the axis of comparison (e.g. pattern versus action).

ABSTRACT

Although assessing visual similarity comes naturally to humans, measuring such perceptions remains a challenge. State-of-the-art metrics typically only consider one approach or format for measuring similarity, and do not generalize well to new scenarios. While recent methods begin to compare images in ways beyond overall visual appearance, there are many kinds of similarities without an established metric. For example, if users want to compare images with respect to art style, noise patterns, or even the number of cats, would they need to construct a dataset and learn a metric from scratch? Instead, our goal is to learn a single model where the similarity comparison, in addition to the images, is also a model input. To this end, we learn a text-parametrized perceptual metric that predicts the conditional similarity specified in the text prompt. We construct a two alternative forced choice (2AFC) dataset to finetune a perceptual embedding where image representations are influenced by the comparison described in the text. This allows us to compare fine-grained and complex visual attributes with one model. We study our model's capability to measure visual similarity in diverse contexts and demonstrate applications on conditional image retrieval.

1 INTRODUCTION

Imagine seeing a leopard in the wild drinking water as in Figure 1. Its mannerisms might remind you of your own house cat, and its fur coat might have the same pattern as a fancy pair of shoes. When deciding whether one image looks like another, we might compare their subjects, color palettes, textures, camera angles, or even subtle details such as the curvature of a particular object. All of these attributes affect visual similarity, but their contribution depends on the kind of comparison we wish to make. Work in psychology has argued that similarity is highly subjective, difficult to define, and requires context to be meaningful Tversky (1977); Tversky & Gati (1978); Goodman (1972); Murphy & Medin (1985). This could explain the extensive variety of image similarity metrics, which range from general-purpose to extremely niche. Although recent metrics go beyond pixels and patches to examining perceptual forms of similarity, they are mostly restricted to the task they are designed for. For example, LPIPS Zhang et al. (2018) and DreamSim Fu et al. (2023), which measure global image similarity, yield mixed results when applied to new comparisons. Recent work has addressed other ways to measure similarity and produced specialized metrics considering attributes such as subject identity Chae et al. (2026) and relational similarity Nguyen et al. (2025). However, if we want to study similarity for new comparisons, such as the leopard's fur pattern,

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

pose, or action, this could mean developing a new model from scratch, requiring the effort of dataset collection and model training.

In this paper, we seek to learn a single model that accurately predicts multiple forms of image similarity. Prior work has tried to parametrize similarity without text Veit et al. (2017), or by aligning separate image and text subspaces Vaze et al. (2023); Agnolucci et al. (2025); Baldrati et al. (2023); Saito et al. (2023); Tang et al. (2024); Gu et al. (2024); Lim et al. (2026). To allow multiple different definitions of similarity to parametrize the same model, we turn to language. We observe that language can express numerous ways to compare images. Inspired by vision language models (VLMs) which respond differently depending on the text prompt, we seek to learn a model for visual similarity parametrized by text. To this end, we collect a dataset of visual similarity judgments for a wide variety of image comparisons which are guided by a text prompt. Similar to Fu et al. (2023), each example in our dataset is a two-alternative-forced-choice (2AFC) example that consists of a reference image, two distortions, and additionally includes a text prompt specifying the comparison axis. To generate diverse and interesting text conditions, we prompt a VLM to identify candidate comparison dimensions for a given reference image. We create image-editing instructions for the reference based on the candidates and use them to generate synthetic distortion images. Finally, we finetune a perceptual embedding on a dataset and learn image representations parametrized by the text dimension. For example, representations for the same image will change depending on whether the text condition is "color," "lighting," or "the pose of the cat." Trained on our dataset, we present a text-parametrized similarity metric that predicts visual similarity for a variety of creative contexts during inference.

In summary, our main contributions are:

1. A dataset of human visual similarity judgments for a variety of different text comparisons.
2. A single model that predicts image similarity parametrized by a text prompt.
3. Our analysis finds that our text-parametrized similarity metric can accurately predict similarity for a variety of comparisons. Our model also supports effective image retrievals and similarity judgments which consider multiple attributes.

2 Related Work

Image similarity Image comparisons have always been a necessary part of computer vision and graphics. Early metrics directly compare pixels (MSE, L1, PSNR), or use patch-based features (SSIM Wang et al. (2004), FSIM Zhang et al. (2011), SIFT Lowe (2004), HOG Dalal & Triggs (2005), HDR-VDP-2 Mantiuk et al. (2011)) mainly to assess global image quality. While these metrics are fast, they are often brittle and break under small transformations.

Perceptual metrics are typically built on features from deep neural networks Szegedy et al. (2016; 2015); Simonyan & Zisserman (2014); Krizhevsky et al. (2012); He et al. (2016); Deng et al. (2009). These feature comparisons have opened the door for image similarity measurements based on semantics, style transfer, and more Kornblith et al. (2020); Zhai et al. (2019); Simonyan et al. (2013); Gatys et al. (2016); Johnson et al. (2016). Recent methods use vision transformer (ViT) features Dosovitskiy et al. (2021) such as CLIP Radford et al. (2021); Cherti et al. (2023), SIGLIP Tschannen et al. (2025); Zhai et al. (2023), and DINO Caron et al. (2021); Oquab et al. (2024); Siméoni et al. (2025). ViT features can effectively compare images and even generalize to drawings and surface materials Chan et al. (2022); Vinker et al. (2022); Sharma et al. (2023). Modern similarity metrics typically collect human similarity judgments to finetune deep features Kotar et al. (2023); Zhang et al. (2018); Prashnani et al. (2018); Ding et al. (2020); Fu et al. (2023).

The state-of-the-art image similarity metrics are LPIPS Zhang et al. (2018) and DreamSim Fu et al. (2023) for low and mid-level visual similarities respectively. Similar to our method, both metrics finetune a vision model on a dataset of human similarity judgments (BAPPS and NIGHTS respectively). Table 1 provides a high-level dataset comparison. Similar to our approach, NIGHTS makes use of synthetic data for its distortions, and the DreamSim model produces embeddings which can be used for retrieval and other applications. However, both metrics are designed for one task: measuring similarity across entire images. Unlike our method, they cannot adapt to new kinds of comparisons.

Although most metrics only address one form of similarity, there has been some effort to consider multiple comparison axes within one model Veit et al. (2017) and more recently for retrieval tasks Lim et al. (2026); Vaze et al. (2023). More recent work focuses on creating metrics for specific kinds of visual comparisons such as relational similarity Nguyen et al. (2025) and subject identity Chae et al. (2026). Concurrent work TPIPS Wang et al. (2026) also uses text prompts to adjust the image similarity in context and was developed independently.

Conditional Image Retrieval Given an input image and text prompt, the goal of conditional image retrieval is to find new images which are similar to the input under the context of the prompt. Composed Image Retrieval (CIR) is a sub-task where the text prompt focuses on changing the input image. Similar to our method, many composed image retrieval methods train a fusion encoder to combine image and text representations Vaze et al. (2023); Agnolucci et al. (2025); Baldrati et al. (2023); Saito et al. (2023); Tang et al. (2024); Gu et al. (2024). Modern approaches still learn a combined representation space, but bridging the domain gap between the real image and text representations remains a challenge Wang et al. (2025); Li et al. (2025). Recent methods transform pretrained VLM features into conditional representations Lim et al. (2026); Hsieh et al. (2025); Liu et al. (2026). Although these approaches are fast and effective for retrieval tasks, they are not designed as perceptual metrics.

Vision Language Models Progress in multimodal large language models has enabled recent models to integrate visual and text information Liu et al. (2023a); Li et al. (2022); Radford et al. (2021). While this has led to better image understanding and evaluations using model logit scores Kumari et al. (2026); Luo et al. (2025); Chin et al. (2025) or "LLM-as-a-judge" Zheng et al. (2023); Wang et al. (2023); Liu et al. (2023b); Chiang & Lee (2023); Wu et al. (2024), many models are limited by the text-only output format, yielding unreliable scores which are highly sensitive to the model and text prompt. Furthermore, methods effective at measuring image-text alignment Lin et al. (2024); Bahng et al. (2025); Xu et al. (2024) are inflexible and break when applied to out-of-distribution text inputs. To mitigate these issues, for our base model we use the pretrained Qwen3-VL-Embedding model Li et al. (2026) which takes image and text inputs and outputs embeddings, instead of raw float scores. These models are based on Qwen3-VL Bai et al. (2025a) and have undergone multi-stage training with contrastive learning.

3 METHOD

Our goal is to learn a visual similarity metric S parametrized by text. Given arbitrary images x_0, x_1 and a text prompt t, our model should output a perceptually accurate similarity score $S^t(x_0, x_1)$ for the two images according to the text prompt, which specifies the axis of comparison. Existing metrics can only predict one form of visual similarity for two images and cannot be adjusted after training Ding et al. (2020); Fu et al. (2023); Prashnani et al. (2018); Chae et al. (2026); Nguyen et al. (2025). Instead, we learn a single model to predict different similarity scores between images depending on the text condition t. The goal of the text prompt is to restrict or focus the visual comparison towards certain characteristics or content in the images. This construction allows us to output visual similarity scores for different attributes such as color, lighting, and background from a single model. We first collect a dataset of perceptual similarity judgments for a variety of comparisons, and learn a perceptual embedding F^t that represents images for the text prompt t.

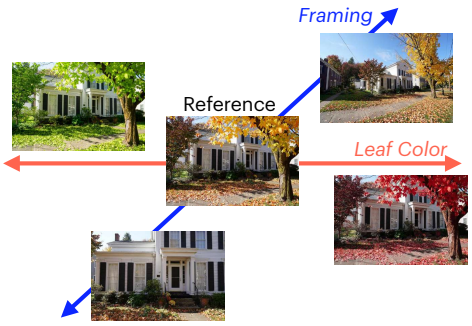


Figure 2: **Comparison Axes**. We show distortions created for Framing and Leaf color text conditions. We assemble 2AFC examples by selecting distortions along the same axis (Strat1), or one from each axis (Strat2).

3.1 DATASET

Our first step is to curate a dataset to gather human similarity judgments based on a text prompt. Similar to Zhang et al. (2018); Fu et al. (2023), we construct a dataset of two alternative forced

Dataset	Metric	Task	Inputs	# Training Examples
BAPPS	LPIPS	General Visual Similarity	Image patches	188k
CUTE	Kotar et al. (2023)	Object Re-ID	Images	18k
NIGHTS	DreamSim	General Visual Similarity	Image	20k
Subjects2k	ID-Sim	Subject Identity	Image	10k
Anonymous-Captions-114k	RelSim	Relational Similarity	Image	114k
Ours	Ours	Parametrized by Text Input	Image, Text	15k

Table 1: **Dataset characteristics.** We provide an overview of some existing datasets used for learning perceptual similarity metrics. Note our method is the only dataset that is flexible to the task, and our model has text inputs used to parametrize the similarity comparison.

choice (2AFC) examples. These examples are of the form $(x, \tilde{x_0}, \tilde{x_1}, t)$ where x is the reference image, $\tilde{x_0}, \tilde{x_1}$ are *distortions* of x, and t is the text prompt stating the comparison context.

Designing Comparisons For scale and diversity, we explore automatic methods to add unique comparisons for our dataset. To construct a 2AFC example, the first step is to assemble three images and a relevant text prompt. We find sampling images randomly or based on dataset labels leads to noisy examples where comparisons are poorly formed or limited by the dataset scope. Instead, we first sample the reference image alone independently. We then query a VLM to propose various ways to compare the reference with other images. We take several steps to encourage the VLM to propose meaningful and creative comparisons: **1)** We prompt the model to identify diverse and interesting parts of the image for comparison. These include both "general" comparisons that can be applied to generic images (e.g. color, lighting), and "specific" comparisons which highly depend on the image content. **2)** We source diverse comparisons naturally by sampling reference images from a wide pool of datasets covering diverse image domains as seen in Figure 3.

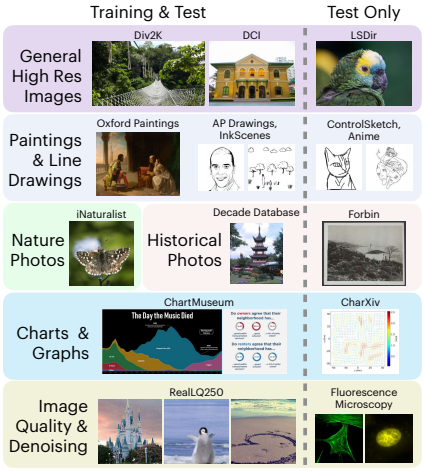


Figure 3: **Dataset Sources**

Generating Distortions After proposing candidate comparison axes for a given image, the next step is to find distortion images that are compatible with both the reference image and text prompt. Instead of searching for relevant distortions, we directly synthesize them. To this end, we ask a VLM to write an image editing instruction given both the reference image and the comparison dimension. We query the VLM to generate two separate instruction prompts, one for each distortion, that change the reference differently with respect to the text prompt. Both instruction prompts are output simultaneously and designed to be distinct from each other. We then input both instruction prompts and the reference image x to state-of-the-art image editing model Nano Banana 2 Gemini (2023) to generate distortions $\tilde{x_0}, \tilde{x_1}$.

To make for interesting and diverse comparisons, we propose two strategies for generating distortions as seen in Figure 2. Consider reference image x and candidate comparison dimension t. To generate distortions, we apply image-editing instructions to x which change t. Note that we can create both distortions using this strategy, as long as instructions specify different degrees of change. We can also create distortions by keeping t constant, but changing other parts of the image. We show dataset examples in Figure 4, and detail our distortion generation strategies as follows:

1. **Strategy 1**: Given reference image x and one relevant comparison dimension t, we generate distortions $\tilde{x_0}, \tilde{x_1}$ which differ from x along t in different ways, forming example $(x, \tilde{x_0}, \tilde{x_1}, t)$.
2. **Strategy 2**: Given reference image x and two distinct comparison dimensions t_0, t_1, we generate distortion $\tilde{x_0}$, which differs from x only along dimension t_0, and symmetrically we create distortion $\tilde{x_1}$ which differs from x only along dimension t_1. This strategy creates two different 2AFC examples for the dataset: $(x, \tilde{x_0}, \tilde{x_1}, t_0)$ and $(x, \tilde{x_0}, \tilde{x_1}, t_1)$.

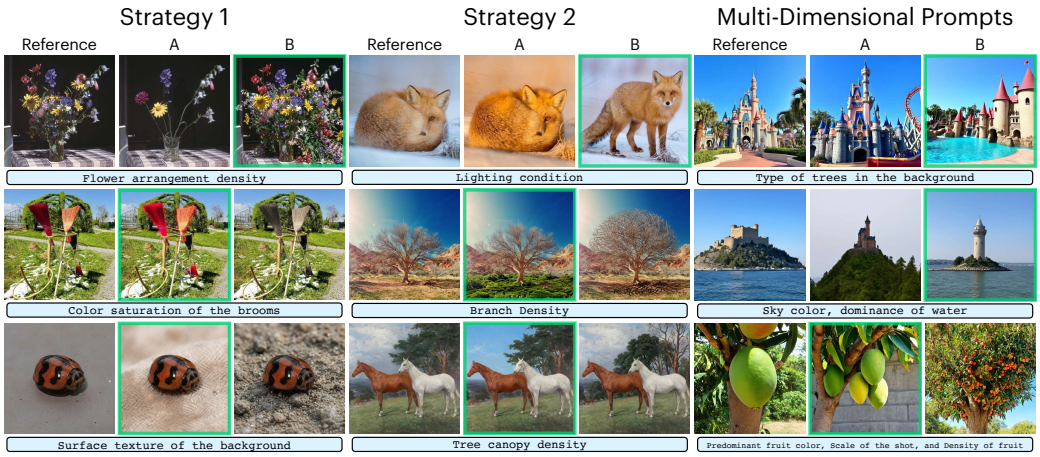


Figure 4: Image comparisons from our benchmark. The most similar image according to the human similarity judgments is outlined in Green. (**Left**) shows examples created with Strategy 1, (**Middle**) for Strategy 2, and (**Right**) for multi-dimensional prompts.

Annotations To annotate the dataset, we create a web-based survey to collect human visual similarity judgments, similar to NIGHTS Fu et al. (2023). Participants are shown the three 2AFC images $x, \tilde{x_0}, \tilde{x_1}$, and then asked to select which image is more similar to the reference according to text condition t. For training and test sets, we collect 3 and 5 judgments respectively. To remove noise from our evaluation set, we remove annotations with a 2-3 split and retain annotations with good ratios. We fully filter disagreements from the training set because of the smaller annotation pool. For details see Section 5.

Implementation Reference images span a wide range of image domains and qualities. We sample from Oxford Paintings Crowley & Zisserman (2014b;a), AP Drawings Yi et al. (2019), InkScenes Tang et al. (2025), Div2K Agustsson & Timofte (2017), DCI Urbanek et al. (2024), Decade Database Palermo et al. (2012), iNaturalist Van Horn et al. (2018), ChartMuseum Tang et al. (2026), and RealQ250 Ai et al. (2024). Text comparison dimensions and image-editing instructions are generated by Qwen3-VL-8B Bai et al. (2023; 2025b;a). Image distortions $\tilde{x_0}, \tilde{x_1}$ are created with Nano Banana 2 Gemini (2023). Our dataset contains 15k comparison examples after label denoising.

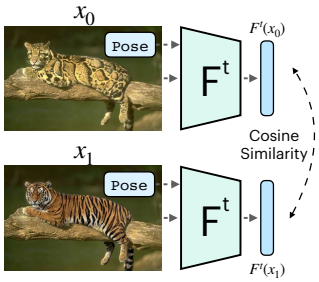


Figure 5: We compute similarity using cosine similarity between embeddings.

3.2 Learning a Text-Parametrized Perceptual Embedding

We explore a training setup similar to Fu et al. (2023) that finetunes a pretrained model to output an embedding for each image aligned with similarity judgments. To this end, we finetune the multimodal Qwen3-VL-Embedding Bai et al. (2025a); Li et al. (2026) model which outputs a single vector for multiple image and text inputs. Let F^t denote our embedding model parametrized by text prompt t. The model F^t then predicts separate embeddings for each image according to text prompt t. For training, we use a hinge loss Chechik et al. (2010) on the cosine distances according to the annotation y. To parametrize the model with respect to t, we use the system prompt "Represent the user's input in terms of t." For training settings, see Appendix B.

During inference, the model, parametrized by text prompt t, outputs a similarity score given a pair of images as seen in Figure 5. To compute image similarities with respect to t, we use the cosine similarity between the conditional embeddings:

$$S^t(x_0, x_1) = \text{CosSim}(F^t(x_0), F^t(x_1)) \tag{1}$$

For each example in the test set $(x, \tilde{x_0}, \tilde{x_1}, t)$ annotated by human perceptual judgment $y \in \{0, 1\}$, text-parametrized image similarity metric S^t predicts the similarity correctly if and only if $(y = 0 \text{ AND } S^t(x, x_0) > S^t(x, x_1))$ OR $(y = 1 \text{ AND } S^t(x, x_1) > S^t(x, x_0))$. In other words, the metric is correct on the test example when its prediction aligns with the annotation.

Method	*Strategy* 1	2	*Dataset* Seen	OOD	*Granularity* Generic	Specific	Overall
General Visual Similarity							
CLIP	80.95	68.02	72.03	73.40	72.50	72.91	72.53
LPIPS	79.03	67.46	72.21	70.27	71.28	73.65	71.49
DreamSim	85.33	70.96	75.46	76.85	75.76	78.11	75.97
Zero Shot VLM-as-Judge							
Claude-Sonnet-5	66.92	72.08	70.71	69.54	70.37	69.39	70.28
Gemini-Flash-3.8	60.56	75.21	70.80	68.91	70.36	67.53	70.10
GPT-6-Luna	62.89	71.18	68.51	67.89	68.61	65.04	68.28
Pretrained Multimodal Embedding							
Qwen3-VL-2B-Embedding	74.80	72.61	74.15	72.04	73.58	71.24	73.37
Qwen3-VL-8B-Embedding	79.56	75.32	77.22	72.54	77.16	76.17	76.80
Parametrized Visual Similarity							
Ours-8B	**86.00**	**88.84**	**88.29**	**87.11**	**88.38**	**82.56**	**87.85**

Table 2: **Visual Diversity Evaluation Set Results.** Our model consistently outperforms all baselines on all subsets of the benchmark.

3.3 BENCHMARK

We introduce a new 2AFC evaluation dataset to benchmark visual similarity metrics on diverse image domains and different comparisons in context. Our benchmark consists of two parts: 1) Evaluating on a wide variety of comparisons sourced from diverse image domains and 2) Evaluating complex text prompts which specify multiple similarity judgments.

Visual Diversity Evaluation We construct an evaluation set similar to Section 3.1 where we use Strategies 1 & 2 to automatically assemble examples, and source comparisons naturally from the diversity of the datasets as in Figure 4 (left, middle). To increase image diversity, we supplement our evaluation set with six new datasets as seen in Figure 3: ControlSketch Arar et al. (2025), Anime colorization dataset Kim (2018) (line drawings), LSDir Li et al. (2023)(general high resolution images), Forbin dataset (historical images) Chelali et al. (2026), CharXiv Wang et al. (2024), and Fluorescence Microscopy dataset Hagen et al. (2021) (image restoration).

Multi-dimensional Text Prompts To evaluate our model on complex text prompts, we create a dataset of examples which consider multiple visual dimensions simultaneously in a single similarity judgment as seen in Figure 4 (right). While our current strategies for generating distortions work well for making direct comparisons, they are often only suitable for comparisons based on 1 or 2 attributes. Instead, we need to find images which are both comparable and have enough natural variation to source multiple different comparisons. To find related images, we sample reconstruction triplets from CyclePrefDB Bahng et al. (2025) which are generated from different captions of the same image. We then propose five different comparison dimensions per image triplet. Often, one distortion is much more visually similar to the reference than the other, so we directly prompt the model to identify commonalities between each distortion and the reference. While generating the dimensions, we ask the VLM Team et al. (2024) to label whether each dimension is more relevant to A or B (i.e. is A or B more similar to the reference along the dimension). We then group the dimensions based on their relevance towards A and B, and concatenate them into a single text prompt to create comparisons based on multiple dimensions. We obtain human similarity judgments for this evaluation set as described in Section 3.1.

4 EXPERIMENTS

We evaluate our parametrizable visual similarity model, demonstrate several downstream applications, and examine its learned representation space.

Comparison methods Our main comparison baselines are the pretrained embedding models Qwen3-VL-Embedding-2B and Qwen3-VL-Embedding-8B Li et al. (2026).

Although these models have a default instruction prompt, we modify this prompt to match our training setting, and find our version yields more favorable results. For conditional image retrieval tasks, we also compare with CLAY Lim et al. (2026) which is a training-free method for transforming vision-language embeddings for text-conditional visual similarity comparisons.

We also compare to standard approaches for measuring visual similarity including LPIPS Zhang et al. (2018) and DreamSim Fu et al. (2023), which are both metrics for general perceptual similarity using pure vision backbones. We add a comparison using vision embeddings from CLIP Radford et al. (2021), which is popular for its ViT image features. Because using "LLM-as-a-judge" Zheng et al. (2023); Wang et al. (2023); Liu et al. (2023b); Chiang & Lee (2023); Wu et al. (2024) is popular for subjective evaluations, we explore the abilities of large foundation models to produce zero-shot similarity scores. We prompt each model with two images, a text condition, and text outlining the scoring system, and the models output their reasoning (if any) and a scalar similarity score. We test the following vision-language foundation models: Qwen3-VL Bai et al. (2025a), Gemini-3.8-Flash Gemini (2023), Claude Sonnet-5 Anthropic (2025), and GPT-6-Luna Singh et al. (2025).

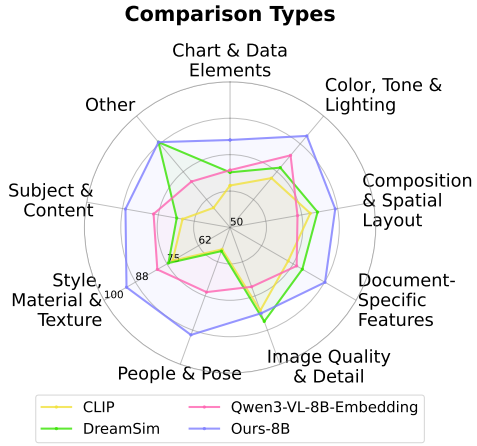


Figure 6: Model comparison on measuring similarity for different forms of image comparisons. Our model (blue) performs well and consistently across comparison types.

4.1 Visual Similarity

We evaluate how well our model predicts similarity on our benchmark detailed in Section 3.3.

Visual Diversity Table 2 reports results on our visually diverse evaluation set. We report overall accuracy, as well as accuracy on several analysis categories. The results are broken down into accuracy for examples created by Strategies 1 & 2, examples in distribution with the training set (Seen) or out (OOD), and by how specific the comparison is to the image inputs (Generic or Specific). To classify whether a comparison is generic or not, we use an LLM (Claude-Sonnet-5 Anthropic (2025)) to group common text prompts by the nature of their comparison. We classify every dataset example into each category, and then label categories which only pertain to one or two subsets of images as "specific" while the rest are labeled "generic." See Appendix A.2 for details. Our model achieves the highest accuracy overall and across all categories. Note that Strategy 1 examples are much easier for general visual similarity models, especially DreamSim, than Strategy 2 where they struggle. On the other hand, Strategy 2 examples are much easier for proprietary models, whereas they struggle with Strategy 1 examples. The embedding models, including our text-parametrized model, have more consistent performance across data types. Our model performance drops slightly when applied to new datasets/comparisons, but this is also true for most methods. However, our model attains the highest accuracies across the board.

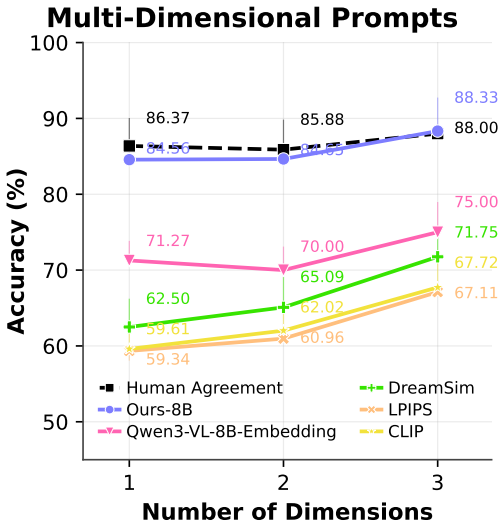


Figure 7: **Multi-dimensional text prompts results.** We plot the accuracy for the number dimensions considered in the text prompt. Our model obtains the best results and close performance to the average human annotation agreement rate. For all models, the similarity predictions generally become more accurate as the number of dimensions is increased from 1 to 3.

Method	Stanford-40			CLAY-Human			CLAY-Object		
	Action	Location	Mood	Age	Action	Background	Object	Subcategory	Color
Conditional Image Retrieval									
CLIP	41.55	48.72	54.24	48.72	50.63	47.24	66.95	38.92	12.79
DreamSim	47.89	51.06	53.14	50.82	50.77	46.84	70.00	43.71	32.31
CLAY (SIGLIP)	66.21	59.39	58.64	61.00	80.40	**89.40**	**94.30**	81.90	63.30
Pretrained Multimodal Embedding									
Qwen3-VL-Embedding-2B	78.95	55.51	67.78	74.61	75.18	62.68	86.54	**91.79**	88.78
Qwen3-VL-Embedding-8B	75.62	**62.07**	**70.78**	74.67	64.34	65.63	84.47	90.08	65.36
Parametrized Visual Similarity									
Ours-8B	**89.00**	58.20	69.09	**75.55**	**97.49**	88.74	91.10	89.06	**92.31**

Table 3: **Conditional Image Retrieval Benchmark Results.** Our model performs well on the CIR task on both Stanford40 and the CLAY-EVAL subsets. We obtain high performance for all categories, especially for the "action" and "color" queries.

Method	CLAY-Human				CLAY-Object	
	Age, Action	Age, Background	Action, Background	All	Category, Color	Subcategory, Color
Conditional Image Retrieval						
CLIP	31.0	29.9	39.8	32.3	11.1	9.5
DreamSim	32.4	29.9	37.3	30.9	36.3	37.7
CLAY (SIGLIP)	49.0	**55.6**	**81.5**	**52.0**	44.7	55.0
Pretrained Multimodal Embedding						
Qwen3-VL-Embedding-2B	42.2 ± 0.3	30.1 ± 1.2	68.8 ± 1.7	40.8 ± 2.0	35.8 ± 0.9	54.8 ± 27.3
Qwen3-VL-Embedding-8B	50.0 ± 1.0	38.5 ± 4.5	67.1 ± 0.5	47.7 ± 2.3	43.6 ± 10.2	69.5 ± 8.0
Parametrized Visual Similarity						
Ours-8B	**61.9** ± 2.8	39.9 ± 24.0	72.7 ± 9.0	34.7 ± 16.1	**74.6** ± 1.2	**79.3** ± 1.0

Table 4: **Multi-Conditional Image Retrieval CLAY Benchmark Results.** For methods with outputs dependent on a text prompt, we report the mean and standard deviation in gray over different orderings of the category names. Our model performs well on the multi-conditional image retrieval task, outperforming all other models on CLAY-Object, but experiences some sensitivity towards the prompts in the CLAY-Human subset.

Figure 6 examines model performance for common comparison groups in our benchmark. We use an LLM (Claude-Sonnet-5 Anthropic (2025)) to assemble text comparisons into common groups (see Appendix A.2). While other methods either specialize in a few kinds of comparisons or have low performance, our model performs consistently well across many different kinds of image comparisons. One exception is DreamSim which outperforms our model on Image Quality & Detail comparisons.

Text complexity Figure 7 shows visual similarity results for text prompts considering multiple dimensions. For text prompts using 1, 2, and 3 dimensions, we show both the average agreement for the human annotations, and each model's agreement with the majority annotation vote. Note that our model performance is very close to the overall human agreement rate for all numbers of dimensions. Most models follow the same trend as the human agreement. For all numbers of dimensions, our model achieves the best performance. Generally, for all models, the accuracy increases as the number of comparison dimensions increases, although there is a slight dip in performance when two dimensions are considered. However, all models achieve the highest performance on text prompts with three dimensions.

4.2 Conditional Image Retrieval

Conditional Image Retrieval We evaluate our model on two conditional image retrieval datasets. Stanford-40 Yao et al. (2011) contains natural images with human annotated labels for action, location, and mood categories from Kwon et al. (2024). However, note there are high correlations between the labels (e.g. "biking" actions typically take place outdoors). CLAY-EVAL Lim et al. (2026) is a conditional retrieval benchmark with synthetic images. It consists of two sub-datasets: CLAY-Human, which has labels for subject age, action, and background, and CLAY-Object, which has labels for the object, object subcategory, and color. For CLAY-EVAL, we also report results using multiple conditions. We report mean Average Precision (mAP) on all tasks.

Table 3 displays conditional retrieval results. Overall, our model performs well in all categories. Note that the pretrained Qwen3-VL-Embedding-8B achieves the best performance for the location

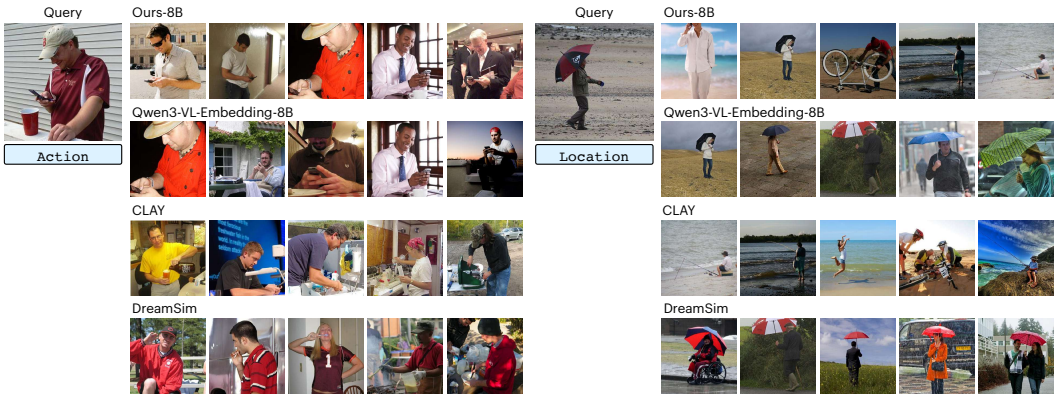


Figure 8: **Conditional Image Retrieval Top-5 NN**. We show each query's top-5 nearest neighbors, from to left right, for each method. Our model retrieves images that are relevant to the query and ignores distractions which affect other methods, such as the cup and umbrella in the left and right examples respectively.

and mood categories in the Stanford-40 benchmark, which are more semantic attributes that are not typically used for visual comparisons.

Table 4 shows results on the CLAY-EVAL benchmark for multiple text conditions. For methods with outputs that depend on a text prompt, we permute the order of the text conditions and report the mean mAP and standard deviation. Our model performs well on most conditions and even outperforms CLAY and the baseline Qwen3-VL-Embedding model on the (Age, Action), (Category, Color), and (Subcategory, Color) conditions. Exceptions include results on (Action, Background) and (All) where the standard deviations are quite high. This shows one limitation of our model; the prompt wording and order can vary output similarity scores. However, this does not affect all prompts, as the CLAY-Object multi-condition results are more robust to permutations.

Figure 8 shows qualitative conditional image retrieval results. Our model is able to effectively focus on the comparison given in the text prompt, whereas other metrics can be sensitive to distractions. For example, in the left example CLAY focuses on the cup, which is unrelated to the subject's action, whereas our model correctly interprets the scene semantics. In the right example, our model is able to separate the location from the user's action and correctly finds images with sand instead of retrieving images of subjects holding umbrellas.

5 Discussion

Our approach learns a text-parametrized image similarity metric that compares images in diverse and creative ways with a single trained model. To learn this model, we first construct a 2AFC dataset where the images vary along a specified comparison dimension given by a text prompt. The model outputs perceptual embeddings, parametrized by text, which represent each image according to the text prompt. We demonstrate that this embedding serves as an effective conditional image similarity metric and finds relevant images for conditional retrieval tasks.

Limitations Despite our model's performance, there are several limitations. Firstly, our dataset is constructed with synthetic text prompts and generated image distortions which inherit biases and failure modes from the underlying models. For example, sometimes the image generation process does not follow the instruction prompt well, making for a confusing distortion image (often with no visible change). Distortions which are meant to corrupt the image are a common failure case. Another challenge is automatically generating text prompts with interesting and creative comparisons. Often the model defaults to proposing generic text conditions that can be applied to most images. Our model is sensitive to the wording in the text prompt, and can yield different results depending on the prompt design, especially for multiple conditions.

AI USE STATEMENT

Tasks with Required disclosure We used generative AI tools to generate synthetic data sets.

We use a VLM (Qwen3-VL-8B Bai et al. (2025a)) to propose comparison dimensions given an image to obtain relevant text conditions for our dataset examples. We use the same model to then generate image-editing instructions for this image with respect to the candidate text dimensions. We also use an LLM (Claude Sonnet 4.6, 5 Anthropic (2025)) to help design several prompts for the model seen in Section A.1. For example, we also use an LLM (Qwen3-8B Yang et al. (2025)) to filter out poor quality editing instructions (e.g. suggest irrelevant changes or no change). We use a diffusion model to execute the editing instructions (Nano Banana 2 Gemini (2023)) and generate synthetic distortion images.

We have **not** used generative AI tools to: help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

Other tasks We have also used generative AI tools for modifying scientific plot formats, creating and editing software code, creating websites for visualization, setting up web infrastructure for our perceptual experiment, and writing suggestions. For example, we used generative AI tools to assist with the HTML/CSS and backend implementation of our web survey for collecting human annotations. We designed the experimental methodology and study ourselves, and AI assistance was limited to implementation.

We have reviewed and tested all AI generated and assisted code for correctness. We have checked the correctness of our plots, visualizations, and other materials. We take full accountability for the content in this paper.

ETHICS STATEMENT

Collecting Human Similarity Judgments For our dataset, we designed a human experiment to gather similarity judgments. Our study was performed with IRB approval, recruited adult participants on the Prolific platform, and collected data anonymously. Similar to Fu et al. (2023), we included sentinel examples to check the quality of annotations. Participants were compensated at approximately $12 per hour (USD). We selected our references images from reputable public datasets to avoid showing participants images with sensitive content.

REPRODUCIBILITY STATEMENT

We report text prompts used to generate, filter, and categorize our dataset in Appendix A.1 and hyperparameter and training settings in Appendix B. Our dataset, model, and code will be released on publication.

REFERENCES

Lorenzo Agnolucci, Alberto Baldrati, Alberto Del Bimbo, and Marco Bertini. isearle: Improving textual inversion for zero-shot composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017.

Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024.

Anthropic. Claude sonnet 5. `https://www.anthropic.com`, 2025. Large language model; accessed May 2026.

Ellie Arar, Yarden Frenkel, Daniel Cohen-Or, Ariel Shamir, and Yael Vinker. Swiftsketch: A diffusion model for image-to-vector sketch generation. In *Proceedings of the special interest group on computer graphics and interactive techniques conference conference papers*, pp. 1–12, 2025.

Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22934–22946, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15338–15347, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Julia Chae, Nicholas Kolkin, Jui-Hsien Wang, Richard Zhang, Sara Beery, and Cusuh Ham. Id-sim: An identity-focused similarity metric. *arXiv preprint arXiv:2604.05039*, 2026.

Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7915–7925, 2022.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010.

Mohamed Chelali, Sylvain-Karl Gosselet, Florence Cloppet, Camille Kurtz, Isabelle Bloch, and Daniel Foliard. The forbin dataset: A collection of historical photographs with archival metadata. *Journal of Open Humanities Data*, 12(1):56–56, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Cheng-Han Chiang and Hung-yi Lee. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*, 2023.

Hsiao-Yuan Chin, I-Chao Shen, Yi-Ting Chiu, Ariel Shamir, and Bing-Yu Chen. Autosketch: Vlm-assisted style-aware vector sketch completion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Elliot Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *Proceedings of the British Machine Vision Conference 2014*. British Machine Vision Association, 2014a.

Elliot J Crowley and Andrew Zisserman. In search of art. In *European conference on computer vision*, pp. 54–70. Springer, 2014b.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 886–893. Ieee, 2005.

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

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

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.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Gemini. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Nelson Goodman. Seven strictures on similarity. *Problems and projects*, 1972.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13225–13234, 2024.

Guy M Hagen, Justin Bendesky, Rosa Machado, Tram-Anh Nguyen, Tanmay Kumar, and Jonathan Ventura. Fluorescence microscopy datasets for training deep neural networks. *GigaScience*, 10 (5):giab032, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Cheng-Yu Hsieh, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Chun-Liang Li, Ranjay Krishna, Oncel Tuzel, and Hadi Pouransari. Focallens: Instruction tuning enables zero-shot conditional image representations. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL https://openreview.net/forum?id=uDfDMrrULe.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Taebum Kim. Anime sketch colorization pair. https://www.kaggle.com/ktaebum/anime-sketch-colorization-pair, 2018.

Simon Kornblith, Honglak Lee, Ting Chen, and Mohammad Norouzi. What's in a loss function for image classification. *arXiv preprint arXiv:2010.16402*, 16, 2020.

Klemen Kotar, Stephen Tian, Hong-Xing Yu, Dan Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. *Advances in Neural Information Processing Systems*, 36:40853–40871, 2023.

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

Nupur Kumari, Sheng-Yu Wang, Nanxuan Zhao, Yotam Nitzan, Yuheng Li, Krishna Kumar Singh, Richard Zhang, Eli Shechtman, Jun-Yan Zhu, and Xun Huang. Learning an image editing model without image editing pairs. In *ICLR*, 2026.

Sehyun Kwon, Jaden Park, Minkyu Kim, Jaewoong Cho, Ernest K Ryu, and Kangwook Lee. Image clustering conditioned on text criteria. In *International Conference on Learning Representations*, volume 2024, pp. 1393–1432, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3984–3993, 2025.

Sohwi Lim, Lee Hyoseok, Jungjoon Park, and Tae-Hyun Oh. Clay: Conditional visual similarity modulation in vision-language embedding space. *arXiv preprint arXiv:2604.11539*, 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.

Honglin Liu, Chao Sun, Peng Hu, Yunfan Li, and Xi Peng. Conditional representation learning for customized tasks. *Advances in Neural Information Processing Systems*, 38:31706–31737, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023b.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Grace Luo, Jonathan Granskog, Aleksander Holynski, and Trevor Darrell. Dual-process image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17972–17983. IEEE, 2025.

Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

Gregory L Murphy and Douglas L Medin. The role of theories in conceptual coherence. *Psychological review*, 92(3):289, 1985.

Thao Nguyen, Sicheng Mo, Krishna Kumar Singh, Yilin Wang, Jing Shi, Nicholas Kolkin, Eli Shechtman, Yong Jae Lee, and Yuheng Li. Relational visual similarity, 2025. URL https://arxiv.org/abs/2512.07833.

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

Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *European Conference on Computer Vision*, pp. 499–512. Springer, 2012.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817. IEEE, 2018.

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*, pp. 8748–8763. PMLR, 2021.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19305–19314, 2023.

Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. Materialistic: Selecting similar materials in images. *ACM Transactions on Graphics*, 42(4), 2023.

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

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

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Liu, Zayne Sprague, et al. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Mia Tang, Yael Vinker, Chuan Yan, Lvmin Zhang, and Maneesh Agrawala. Instance segmentation of scene sketches using natural image priors. SIGGRAPH Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715402. doi: 10.1145/3721238.3730606. URL https://doi.org/10.1145/3721238.3730606.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5180–5188, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Amos Tversky and Itamar Gati. Studies of similarity. *Cognition and categorization*, 1(1978):79–98, 1978.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26700–26709, 2024.

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

Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6862–6872, 2023.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1781–1789. IEEE, 2017.

Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pp. 1–11, 2023.

Lan Wang, Wei Ao, Vishnu Naresh Boddeti, and Ser-Nam Lim. Generative zero-shot composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29690–29700, 2025.

Sheng-Yu Wang, Yotam Nitzan, Aaron Hertzmann, Jun-Yan Zhu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. The many senses of visual similarity: A text-prompted image perceptual metric. *arXiv preprint arXiv:2607.18237*, 2026.

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.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54015–54029, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas J Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, volume 2, pp. 7. Barcelona, Spain, 2011.

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10735–10744. IEEE, 2019.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

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.

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*, pp. 586–595, 2018.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

APPENDIX

A DATASET

A.1 INSTRUCTION PROMPTS

We experimented and designed instruction prompts seen below.

Dataset Generation Prompts The following dataset generation prompts were designed with minor assistance from Claude Anthropic (2025).

Generating Interesting and Specific Text Dimensions

Task: As a researcher in fine-grained visual details, your task is to examine an image and then suggest some dimensions for comparison with other images which may share some similarities. You must examine the image and determine number dimensions for meaningful comparison between the input image and other similar images.

Important considerations:

1. The dimensions you suggest should be diverse and identify interesting parts of the images that would make for a good comparison. They can be either simple or complex, global to the entire image or local to certain parts, as long as they examine unique attributes.
2. You should consider both low-level and high-level characteristics, depending on the content of the image. For example, low-level characteristics could examine proportions, color, shape, texture, blur, noise, image quality, sharpness, line thickness etc. High level characteristics could examine semantic content in the scene, for example considering the appearance of objects or regions, positions, actions performed, the context behind the photograph etc.
3. Avoid comparisons that mention the "Presence of" or inclusion of a particular element in the image.
4. Please make your dimension as clear and concise using as few words as possible to accurately describe the comparison axis. Do not add any extra words. Do not put any parenthesis or extra explanations your outputs, including no e.g. examples. Just state the dimension.

Here are some possible examples of inputs and comparison dimensions:
Example 1:
Input Image: [closeup image of a flower with pointed edges]
Dimension: The shape of the petals
Example 2:
Input Image: [image of a large dog sitting on top of a pillow]
Dimension: Size of the subject in the image
Example 3:
Input Image: [image of two birds on a branch in front of a green bokeh, blurred background]
Dimension: How blurred the background is
Example 4:
Input Image: [Slightly blurred landscape photograph of a small wooden shack in a yellow field with an older car]
Dimension: What year the photograph was taken
Example 5:
Input Image: [a degraded picture of a lemon which has several artifacts corrupting the image]
Dimension: The magnitude of image corruption

You should the dimensions in JSON format as indicated below.
Follow this format exactly: { "dim": ["dim1", ... "dim{number}"] }

Generating Strategy 1 Image-Editing Instructions

Task: As a researcher in fine-grained visual details, your task is to first examine the image along the following dimension: dim. The text dimension is provided as a way to generate comparisons with other images. You will then generate two different instruction prompts for an image-editing model to change the image along this dimension in two distinct ways.
Important considerations:

1. Depending on the selected dimension, the instruction prompt can be local to certain parts of the images, specific to certain objects, or global transformations. Changes can be as creative as you want, including re-imagining the same image in different styles, perspectives, shapes, colors etc - as long as they are visually noticeable and within the capabilities of modern image editing tools.

2. The instruction prompt is allowed to change the image in creative ways, or be congruent with the visual integrity of the original image. For instance, changes can either improve the visual quality or decrease it.
3. Try to be imaginative and detailed in your prompts. The text-to-image model may try to reuse parts and aspects of the input image in undesirable ways, so focus on describing the new parts of the image in your prompt.
4. Prompts should be detailed, exact, and clear for an text-to-image model to interpret without confusion, and generate a new image successfully. You should be specific about the changes you want to see, and leave little for the text-to-image model to decide on its own.
5. Ideally, the two instruction prompts should change the image along the given text dimension in two different magnitudes. You must design both instruction prompts so that they both clearly change the image in some way. You cannot indicate "no change" or changes that are very small and would be hard to recognize.

Here are some possible examples of final instruction prompts:
Example 1:
Input Image: [closeup image of a flower]
Dimension: The shape of the petals
Prompt 1: Change the petals of the flower so that the edges appear sharper and more serrated
Prompt 2: Change the petals of the flower so that the edges appear more rounded and circular.
Example 2:
Input Image: [image of a large dog sitting on top of a pillow]
Dimension: Size of the subject in the image
Prompt 1: Transform this image so the subject is about 20% of its original size relative to its environment while keeping the lighting, background, and style exactly the same.
Prompt 2: Transform this image so the subject appears gigantic relative to its environment while keeping the lighting, background, and style exactly the same.
Example 3:
Input Image: [image of two birds on a branch in front of a green slightly blurred background]
Dimension: How blurred the background is
Prompt 1: Change the background so that it is more blurry and has some bokeh effects, while keeping the lighting, style, and foreground of the image exactly the same.
Prompt 2: Sharpen the background so that it is detailed and in focus, while keeping the lighting, style, and foreground of the image exactly the same. Example 4:
Input Image: [image of a beautiful watercolor painting of stained glass window with a yellow flower]
Dimension: The artistic style of the image
Prompt 1: Transform this image into the style of a classical oil painting (visible brushstrokes, rich color, painterly texture) while keeping the subject, composition, lighting, and background exactly the same.
Prompt 2: Re-imagine this image as a cute pixel art stained glass window as if it is in a cozy indie video game while keeping the subject, composition, lighting, and background exactly the same.
Example 5:
Input Image: [Landscape photograph of a small wooden shack in a yellow field with a orange-pink sunset]
Dimension: The viewpoint of the image
Prompt 1: Transform this image to a wide-angle view (showing much more of the surroundings, subject appears smaller in the scene) while keeping the lighting, background, and style exactly the same.
Prompt 2: Transform this image to an over-the-shoulder perspective while keeping the subject, lighting, background, and style exactly the same.

Example 6:
Input Image: [a small orange fish peeking out of a small decorated pond on a cloudy overcast night]
Dimension: The lighting
Prompt 1: Transform this image so the lighting looks like golden hour (warm, low-angle sunlight with long soft shadows) while keeping the subject, composition, background, and style exactly the same.
Prompt 2: Transform this image so the scene is lit only by moonlight (cool blue tones, dim ambient light, nighttime atmosphere) while keeping the subject, composition, background, and style exactly the same.
Example 7:
Input Image: [a person running down the grocery store aisle]
Dimension: The amount of motion in the image
Prompt 1: Transform this image so the person is standing still with their hands behind their back and looking for a particular item in the aisle, while keeping the subject, composition, background, and style exactly the same.
Prompt 2: Transform this image so the person is walking slowly browsing while keeping the subject, composition, background, and style exactly the same.

You should output the final instruction prompts in JSON format as indicated below.
Follow this format exactly: { "prompts": [["prompt1", "prompt2"]] }

Generating Strategy 2 Image-Editing Instructions

Task: As a researcher in fine-grained visual details, your task is to first examine the image along the two given text dimensions: dim1 and dim2. The text dimensions are provided as a way to generate comparisons with other images. For each dimension, you will generate one instruction prompt for an image-editing model to change the image along this dimension.
Important considerations:

1. The text prompts you write should describe a new image which is the same as the input image along the given dimension. You can assume that the text-to-image model will see the original image, and you can refer to it as the "input image" in your prompt.
2. You should write the prompts so that the two new generated images are different along both dimensions you have identified. For example, if dimension 1 is the background and dimension 2 is the shape of flower petals, prompt 1 should generate a new image which has changed the shape of the flower petals from the input image, while keeping the background. Symmetrically, prompt 2 should generate another new image with a different background from the input, but the shape of the flower petals should match the input.
3. Try to be imaginative and detailed in your prompts. The text-to-image model may try to reuse parts and aspects of the input image in undesirable ways, so focus on describing the new parts of the image in your prompt.
4. The instruction prompt is allowed to change the image in creative ways, or be congruent with the visual integrity of the original image. For instance, changes can either improve the visual quality or decrease it.
5. Prompts should be detailed, exact, and clear for an text-to-image model to interpret without confusion, and generate a new image successfully. You should be specific about the changes you want to see, and leave little for the text-to-image model to decide on its own.
6. Add "Keep the aspect ratio of the image the same." to the end of the prompt if this (or a similar sentence) is not already there.
7. Make sure to only generate TWO prompts total - one for each dimension!

Here are some possible examples of inputs, dimensions, prompts, and questions.
Example 1:
Input Image: [Image of a large white balloon in a deep blue sky]
Dimension 1: The background of the image
Prompt 1: Change this image so that the balloon is a giant passenger airplane taking off. Keep the background the exact same and the aspect ratio of this image the same.
——
Dimension 2: The identity of the main subject
Prompt 2: Change this image so that the background sky is foggy and visibility is poor. Keep the balloon the same and keep the aspect ratio of the image.

Example 2:
Input Image: [Image of a brown clay sculpture of a bunny rabbit against a black background]
Dimension 1: The material of the main subject
Prompt 1: Change the bunny in the image into a statue of a cat made of the same material. The cat statue should show the cat majestically leaping up into the air. Make the background a green decorated garden. Keep the aspect ratio of the image the same.
——
Dimension 2: The type of animal in the image
Prompt 2: Give me an image with the same main subject as this image, but the bunny is a real life black and white fluffy bunny eating a dandelion on a fluffy purple bed.

Example 3:
Input Image: [Image of a cat wearing a noticeable, whimsical purple and orange pointy hat]
Dimension 1: The hat the main subject is wearing
Prompt 1: Make a closeup image of a green scaly gecko wearing the same hat as in the input image. The gecko should be positioned sideways and it is positioned on all fours on a gray rock, and it is wearing a hat identical to the one in the photo, except it is appropriately sized for the gecko. The gecko should take up most of the image. Keep the aspect ratio of the image the same.
—— Dimension 2: The size of the main subject
Prompt 2: Generate an image of capybara, where the capybara is the same size (bounding box) as the cat in the input image. The capybara should be wearing a white baseball hat. The capybara should be on the right of the image, sniffing an orange next to a pond. Keep the aspect ratio of the image the same.

You should output the final instruction prompts in JSON format as indicated below.
Follow this format exactly: { “prompts”: [[“prompt1”, “prompt2”]], }

Generating Multi-Dimensional Text Comparisons

Task: As a researcher in fine-grained visual details, your task is to examine the three input images. The first image is the reference, the second is image A, and the third is image B. Your job is to identify num different dimensions which highlight visual characteristics that offering meaningful comparisons when comparing the reference image to images A and B. Specifically, you should create a balanced list of dimensions so that across the different comparisons, sometimes image A is more similar to the reference and sometimes image B will be more similar.
Then for each dimension you choose, please annotate whether image A or image B is more similar to the reference in terms of that dimension.
Important considerations:

1. Given the three images, the dimensions should highlight relevant characteristics across all three of the images. Even if two of the images are much more similar to each other than the third, the dimensions you select should be inclusive and relevant to ALL three of the images.
2. The dimensions you select should all be unique and different from each other.
3. The dimensions you suggest should be diverse and examine interesting aspects of the image. They can be either simple or complex, global to the entire image or local to certain parts.
4. You should both consider dimensions which are both specific to the input image (e.g. water level if the image shows some liquid, the color of a hat if the subject is wearing a hat, the number of stripes on a frog's back, etc) and general to most images (e.g. color, lighting, perspective, etc).
5. You should consider both low-level and high-level characteristics, depending on the content of the image.
6. Please make your dimension as clear and concise using as few words as possible to accurately describe the comparison axis. The dimension should be easily understood from the text alone without viewing the image. Do not add any extra words (such as "Overall"), parentheses, or explanations in your outputs, including no e.g. examples. Just state the dimension.
7. Avoid comparisons that mention the "Presence of" or "composition" or inclusion of a particular element in the image. Instead, use other terms to describe the comparison.

Here are some possible examples of inputs and comparison dimensions:
Example 1:
Input Images: [closeup image of a blue flower with pointed edges, landscape photo of a sunflower showing the sky, a sunflower sitting in an open field facing the sky]
Example Dimensions: The shape of the petals, the type of flower, the background in the image, the color of the flower,
Example 2:
Input Image: [images of dogs in various locations and backgrounds]
Dimension: Size of the dog in the image, positioning of the dog, the color of the dog, how fuzzy the dog is, the texture of the image, etc
Example 3:
Input Image: [a degraded pictures of a lemon which has several artifacts corrupting the image]
Dimension: The magnitude of image corruption, camera perspective, the texture of the lemon, the position of the lemon, the color of the background, etc

You should the dimensions in JSON format as indicated below. For the annotations strings please choose one of "A" or "B" to designate if image A or image B is more similar in terms of the corresponding dimension.
Follow this format exactly: { "dim": ["dim1", ... , "dim{num}], "annotation": ["annotation1", ..., "annotation{num}"] }

LLM filtering To improve dataset quality, we use an LLM (Qwen3 Yang et al. (2025)) to identify instruction prompts which do suggest no change, or do not pertain to the comparison dimension they are supposed to consider. We relied heavily on Claude for designing this prompt. We use the following text prompt to filter out examples before generating distortions.

LLM filtering out low quality instructions

You are auditing a dataset of image-editing prompt pairs for a consistency error.
Each entry has a ”dim” field describing the single visual attribute that is supposed to be varied between Image A and Image B (relative to a reference image). The fields ”promptA” and ”promptB” are text instructions that were used to generate Image A and Image B respectively. Sometimes the LLM that generated these prompts loses track of the dimension and writes instructions about a completely different attribute instead.
Your job: for each of promptA and promptB, determine whether the prompt actually describes a change/comparison relevant to ”dim”. A prompt is a MATCH if it centers on manipulating or preserving the attribute named in ”dim” (even if worded differently). A prompt is a MISMATCH if it is primarily about a different attribute, even if it incidentally mentions the dim's attribute as a side constraint.
Note: it is common and fine for only ONE of promptA/promptB to match — that still produces a usable comparison. Only flag the entry if NEITHER prompt relates to the dimension at all.
Output only this JSON object, nothing else:
{ “promptA_match”: true or false, “promptB_match”: true or false, “entry_flagged”: true or false }
—
Example 1 (entry_flagged: false):
dim: ”Texture of the tree bark”
promptA: ”Change the texture of the tree bark in the image to appear smooth and polished, resembling the surface of marble, while keeping the white insect, the hand, and the background bokeh exactly the same. Keep the aspect ratio of the image the same.”
promptB: ”Change the perspective angle of the image so that it is now viewed from directly above, looking down at the insect and hand holding it against the tree bark, while keeping the tree bark's original rough, natural texture and the background bokeh exactly the same. Keep the aspect ratio of the image the same.”

Output:
{”promptA_match”: true, ”promptB_match”: false, ”entry_flagged”: false}

—
Example 2 (entry_flagged: true):
dim: ”Color saturation of the background”
promptA: ”Reduce the size of the insect to make it appear much smaller, almost like a tiny larva or nymph clinging to the tree bark, while the hand and background remain unchanged in composition and lighting. Keep the aspect ratio of the image the same.”
promptB: ”Make the insect in the image appear significantly larger, as if it were a giant moth with wings spread wide, resting on the same tree bark and being gently held by the same hand. The background should remain exactly the same, with the blurred green foliage and bokeh lights. Keep the aspect ratio of the image the same.”

Output: {”promptA_match”: false, ”promptB_match”: false, ”entry_flagged”: true}
—
Now evaluate this entry:
dim: {dim} promptA: {promptA} promptB: {promptB}

A.2 Categorizing the Evaluation Set

To generate the generic/specific splits and group common comparison types for our dataset, we use the following prompts.

Granularity We use the following prompt to generate a list of common categories for comparison prompts. We generate the initial list using LLM Claude-Sonnet-4.6 Anthropic (2025).

Generating common categories from text prompts

You are given the following python list of strings, which lists potential dimensions for visual comparisons between two images. Your task is to come up with a list of categories that group similar kinds of visual comparisons together. For the categories, try to avoid specific "semantics" (i.e. if they mention specific types of objects in the picture), and instead group the dimensions by the kind of comparison involved (e.g. color palette, size, position, lighting, type of main subject, pose, location, counting, layout, material, perspective, camera angle, etc). The category names should be specific and descriptive of the dimensions in the category, but also concise. For example, if you name a category Lighting, then examples another category formed from a different dataset named Lighting should also fit the category. Please output your final list of categories as a python list of tuples, where each tuple has the category name and then a brief description of the category. Please try your best to list around 15 categories.

Then for each dataset example, we individually classify each example into one of the categories using a local LLM Qwen3 Yang et al. (2025) with the following prompt.

Classifying examples by text prompt into categories

Given the following list of categories and one dimension, please classify which category the dimension best fits. Please try your best to classify the dimension into exactly ONE best fit category according to how related it is to the category.

Your output should be exactly the category which best fits the dimension, exactly how it appears in the category list. No extra characters, no extra text, and no explanations.

categories = {cat} dimension = {dim}
Category:

Categorizing comparisons Since the number of categories generated is usually quite large (15-20), we decide to merge categories based on their text similarities. We first obtain the category classification for each dataset example, and then an LLM (Claude-Sonnet-4.6 Anthropic (2025)) to join categories based on how similar their text prompts are using the following prompt:

Merging comparison categories

You are an expert in analyzing visual comparisons.

You are given the following python dictionary. In the dictionary, the keys represent a certain type of image which has been manually labeled. The keys then map to a value which list of categories which encompass various types of visual comparisons that were made using the images of the corresponding type/key. The list is a list of lists, where the inner list has the format: ["Category name", "brief category explanation"].

Your job is to analyze this dictionary, the keys and the comparison types and explanations, and then summarize the kinds of comparisons made in each of the keys in the dictionary. Your final output should be a python list of tuples like so:

categories = [("category 1", "explanation 1"), . . . , ("category n", "explanation n")]

where the categories are the similar kinds of comparisons made across different keys, and the explanations give some examples of the kinds of comparisons that fall under the corresponding category.

We then use the final list of categories for the comparison types in Figure!6.

Method	NIGHTS	BAPPS
General Visual Similarity		
CLIP	82.08	64.29
LPIPS	70.24	68.2
DreamSim	**95.52**	**68.3**
Pretrained Multimodal Embedding		
Qwen3-VL-Embedding-2B	86.46	63.61
Qwen3-VL-Embedding-8B	88.07	66.45
Parametrizable Visual Similarity		
Ours-8B	90.47	65.61

Table 5: **Results for general visual similarity benchmarks.**

B Implementation Details

Training We fully finetune for 2 epochs with lr $5e-6$ on 2 H200 GPUs with with lr warm up and batch size 32.

Hinge Loss For cosine similarities given by $s_0 = S(x, x_0, t)$ and $s_1 = S(x, x_1, t)$, the hinge loss is formulated as: $L = \max(0, m - (s_1 - s_0)y)$ where m is the margin hyperparameter that controls the tolerance around the decision boundary. In practice we set $m = 0.05$. The final embedding size is 4096.

C Additional Results

C.1 General Visual Similarity Benchmarks

We also report results for other visual similarity benchmarks. NIGHTS and BAPPS are both evaluations introduced by DreamSim Fu et al. (2023) and LPIPS Zhang et al. (2018) for overall mid and low-level visual similarities respectively. Note that NIGHTS and BAPPS do not incorporate specific text prompts or conditions unlike our evaluation sets.

Table 5 shows results on the NIGHTS, and BAPPS evaluation sets. Note that unlike DreamSim, and LPIPS, our model is **not** trained explicitly to predict overall visual similarity. Unsurprisingly, DreamSim and LPIPS perform the best on the general similarity benchmarks. However, our model improves over the pretrained base models on the general similarity task.