Relational Visual Similarity

Thao Nguyen[1], Sicheng Mo[2], Krishna Kumar Singh[3], Yilin Wang[3], Jing Shi[3], Nicholas Kolkin[3]
Eli Shechtman[3], Yong Jae Lee[1,3,†], Yuheng Li[3,†]
[1]University of Wisconsin-Madison, [2]University of California, Los Angeles, [3]Adobe Research
https://thaoshibe.github.io/relsim

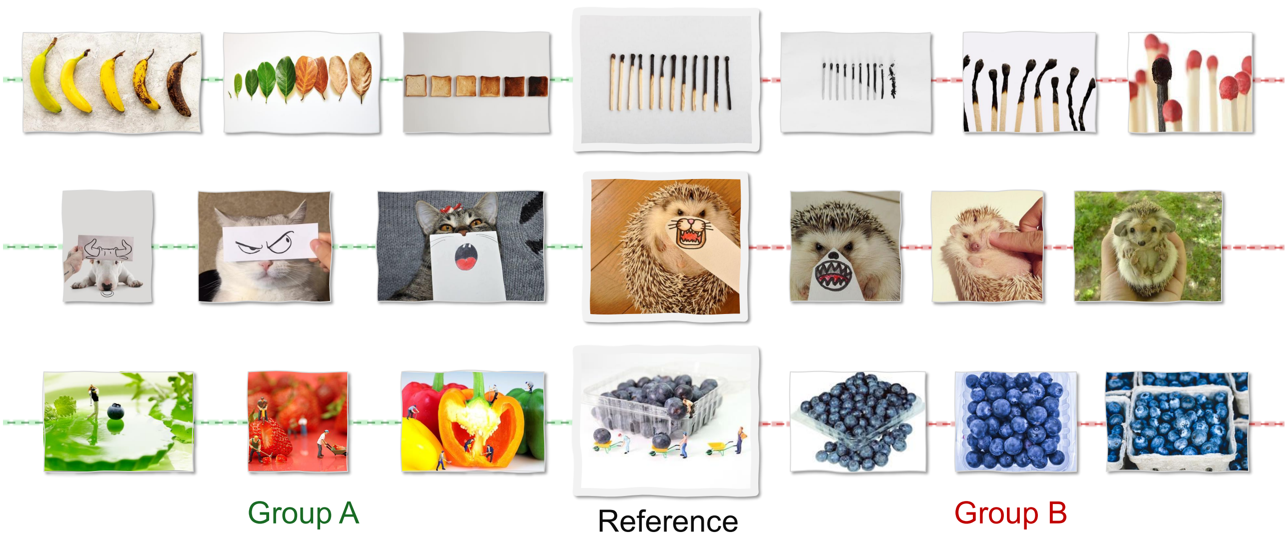


Figure 1. **Would you say images in Group A are similar to the Reference Image?** Current state-of-the-art image similarity models (e.g., LPIPS [1], CLIP [2]) would answer *no*. These models would say only **Group B** are similar to the reference image, as they equate similarity with a high degree of shared perceptual attribute features (i.e., color, shape, semantic class). However, as humans, we would confidently say *yes*—images in both groups are similar to the reference. While **Group B** is similar in perceptual attributes, **Group A** is similar in a more abstract, relational sense (e.g., "transformation of {subject} through time", first row). In this paper, we propose to model this missing dimension of visual similarity, or called *relational visual similarity*, capturing human-like reasoning over relational structures.

Abstract

Humans do not just see attribute similarity—we also see relational similarity. An apple is like a peach because both are reddish fruit, but the Earth is also like a peach: its crust, mantle, and core correspond to the peach's skin, flesh, and pit. This ability to perceive and recognize relational similarity, is arguable by cognitive scientist to be what distinguishes humans from other species. Yet, all widely used visual similarity metrics today (e.g., LPIPS, CLIP, DINO) focus solely on perceptual attribute similarity and fail to capture the rich, often surprising relational similarities that humans perceive. How can we go beyond the visible content of an image to capture its relational properties? How can we bring images with the same relational logic closer together in representation space? To answer these questions, we first formulate relational image similarity as a measurable problem: two images are relationally similar when their internal relations or functions among visual elements correspond, even if their visual attributes differ. We then curate 114k image–caption dataset in which the captions are anonymized—describing the underlying relational logic of the scene rather than its surface content. Using this dataset, we finetune a Vision–Language model to measure the relational similarity between images. This model serves as the first step toward connecting images by their underlying relational structure rather than their visible appearance. Our study shows that while relational similarity has a lot of real-world applications, existing image similarity models fail to capture it—revealing a critical gap in visual computing.

† denotes equal advising

1. Introduction

The ability to perceive and recognize visual similarity is arguably the most fundamental sense for any visual creature, including humans, to interact and make sense of the world [3, 4]. We process visual attributes to guide decisions: recognizing that a peach is red might signal that it is edible. We also notice similarities across different objects (e.g., shape, color, texture) to categorize, remember, and abstract them: an apple and a peach are both red and round, so they are likely both fruits. Beyond this, we can see relational similarity as well: we abstract familiar patterns to understand more complex or unseen phenomena. For example, we can anticipate the Earth is like a peach, as its layers—crust, mantle, and core—roughly correspond to the peach's skin, flesh, and pit, even though no one has directly observed it. In cognitive science, attribute similarity and relational similarity are often considered the two central pillars when it comes to understanding human perception of similarity [5, 6]. Attribute similarity underlies everyday activities (e.g., recognition [7], classification [8], memorization [9]), while relational similarity fuels reasoning and creativity (e.g., analogies [10], abstract thought [11]). Some researchers argue that relational similarity is even more central to human cognition, as it drives analogical learning and creativity—the traits that set humans apart from other intelligent species [12–14].

Unfortunately, current state-of-the-art visual similarity frameworks focus almost exclusively on attribute-level similarity. Traditionally, image similarity in computer vision has been framed as the task of comparing two images and deciding whether they are visually similar, typically at the pixel or feature level using handcrafted descriptors [15, 16]. In recent years, large-scale hierarchical datasets (e.g., ImageNet [17]) and cross-modal datasets (e.g., LAION-2B [18]) have enabled deep learning models to move beyond low-level visual details. Modern approaches (e.g., [2, 19–23]) can recognize different images of the same semantic class or images that match a rough textual description—for example, "a photo of matchsticks"—even if they differ in shape, color, or other low- to mid-level details (Fig. 1, Group B, first row).

However, by focusing primarily on surface-level features, these models struggle to capture *relational similarity* (see [24, 25], Sec. 4.2). For instance, they cannot easily recognize that the burning stages of a match resemble the ripening stages of a banana (Fig. 1, Group A, first row). Capturing this type of similarity requires a shift in perspective: instead of relying solely on visual features, we must reason about how different visual elements, interact, abstracting the underlying relationships. For example, both the match and the banana undergo a gradual transformation over time. The similarity lies not in their specific appearance but in the logic of change. This raises questions: which attributes should be preserved or ignored during comparison? How can we identify which relational patterns are relevant or useful?

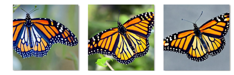
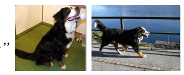
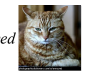
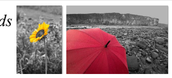
Dataset	Data Format	Example
BAPPS [1]	image triplet *(low-level perceptual)*	
NIGHTS [22]	image triplet *(mid-level perceptual)*	
ImageNet [17]	semantic class *(attribute-based)*	*"Bernese mountain dog"* 
LAION-2B [18]	{image, caption} *(attribute-based)*	*"Cats of Torcello, Italy"*  *"An annoyed cat"* 
Ours (*relsim*)	{image, anonymous caption} *(relational-based)*	*"A {object} stands out against gray background"* 

Table 1. **Survey of prominent datasets** used for training visual similarity metrics. All are organized based on attribute similarity, whereas ours focuses on relational similarity.

Insights from cognitive science, encouragingly, offer a spark for these questions. Works [10, 26] showed that humans process attribute similarity perceptually, but relational similarity requires conceptual abstraction, often supported by language or prior knowledge. This suggests that recognizing relational similarity first requires understanding the image, drawing on knowledge, and abstracting its underlying structure. Take the example of a photo of burning matches: we first observe how each match relates to the others—they burn sequentially from left to right. With prior knowledge, we understand that burning is a temporal transformation, a process that can occur in many other objects (e.g., a leaf aging, a banana ripening). If asked to write a caption capturing this logic rather than the specific objects, one might write "transformation of {subject} over time". We call such captions *anonymous captions*—they do not describe any particular visible object but instead capture the relational logic conveyed by the image. These captions act as the glue connecting images with similar underlying logic. In other words, a successful relational visual similarity model must understand, abstract, and use anonymous captions to bring logically similar images together.

To model relational similarity, we follow a path inspired by insights from cognitive science. Since no existing dataset captures relational visual similarity (see Tab. 1), we first filter a large image corpus, LAION-2B [18], to extract 114k images likely to contain transferable relational structures. This step improves dataset quality by removing low-quality, mislabeled, or relationally uninformative images, which are common in LAION-2B [27, 28]. We then train an anonymous captioning model to generate captions for these images, creating a set of {image, anonymous caption} pairs. Finally, we train a relational visual similarity model, *relsim*, on this dataset, optimizing it to bring together images whose captions encode similar relational abstractions. We demonstrate the utility of *relsim* for tasks such as relational image retrieval and analogical image generation.

In short, our contributions are as follows:

- A new notion of image similarity, *relational visual similarity*, which complements traditional attribute similarity.
- A novel relational dataset, consisting of 114k {image-anonymous captions} designed to capture the abstraction and logic in each image.
- A new tuned metric, *relsim*, that captures the relational visual similarity between two images.
- Analysis of the relationship between relational and attribute similarity, along with experiments demonstrating the limitations of current image similarity models.
- Demonstration of downstream applications in image retrieval and image generation.

2. Related Works

Similarity in Cognitive Science. The question of what makes two subjects similar has always been considered one of the most significant questions in cognitive science [3, 9, 29–31]. Similarity is fundamental to human cognition, as it affects how the mind organizes, categorizes, and reasons about the world. For decades, Tversky's theory of similarity [9], also called the contrast model, has been widely adopted and has inspired multiple domains [1, 22, 32]. Tversky frames similarity as a psychological comparison of matching individual properties or characteristics of objects (e.g., size, shape, color). For example, an apple and a banana are similar because they are both fruits. While powerful, Tversky's theory cannot account for similarities such as the one Stephen Hawking made when he said, "I regard the brain as a computer" [33]. There are no obvious visual features shared between a human brain and a computer. This kind of similarity, which cannot be fully accounted for by Tversky's model, was later formalized as relational similarity, alongside its counterpart, now called attribute similarity. These concepts emerged from Gentner's research on analogy, often referred as Structure-Mapping theory [10]. Relational similarity is a comparison based on the relationships between objects. Returning to the previous example, Stephen Hawking was making a relational comparison: he viewed the brain as a biological machine and the process of death as analogous to a computer breaking down. Substantial research shows that while both type of similarity are important, relational similarity (often associated with analogical reasoning) plays a distinct and often deeper role in human cognition (i.e., analogical learning and reasoning [3, 12–14]).

Image Similarity. Comparing similarity between two visual signals is a core concept in computer vision, as it underpins many tasks (e.g., object recognition, image retrieval, image matching). Before the deep learning era, most image similarities were computed directly via pixel-level metrics (e.g., L1, L2, MSE, RMSE, PSNR) or handcrafted features (e.g., SSIM [34], FSIM [35], SIFT [15]). With the rise of deep learning and neural networks (e.g., VGG [21], ResNet [23]), deep-feature-based image similarity metrics better align with human perceptual judgment (e.g., LPIPS [1], PieAPP [36], DISTS [37]). More recently, with the aid of Vision Transformers (ViT) [38] and Self-Supervised Learning (SSL), modern vision encoders (e.g., DINO [20], CLIP [2], dreamsim [22], SigLIP [39]) not only provide robust visual embeddings for image similarity, but also enable semantic comparisons that go beyond pixel-level matching. However, all of these approaches rely on the assumption that image similarity is based solely on attribute similarity, and thus cannot capture relational similarity, as we demonstrate in our experiments (Sec. 4.2). Here, we, for the first time, propose to consider *relational visual similarity*.

Mutimodal Large Language Models. Research on multimodal models (e.g., [40–48]) has become an increasingly attractive topic in recent years. In particular, progress in developing unified models that can both understand and generate visual and textual inputs/outputs has transformed how we interpret and interact with visual information. While traditional vision encoders (e.g., CLIP [2]) can mostly only "see" what is explicitly shown in an image (e.g., "a photo of a mother hugging a child"), integrating them with MLLMs allows us to capture what is not directly depicted (e.g., "the image representing a sense of parental care"). Since relational similarity often requires a deeper understanding of images that goes beyond mere perception, we choose to leverage MLLMs, particularly Vision Language Models (VLMs), as the backbone for image feature extraction.

3. Relational Visual Similarity

We formalize the problem of measuring the relational visual similarity as follows. Given two input images I_1 and I_2, we aim to train a visual feature extractor f_V such that the resulting features capture the *relational similarity* between the two images. Our core assumption is that if two images exhibit high relational similarity, then their corresponding anonymous captions, A_1 and A_2, should also be similar. Specifically, we define the relational similarity score s_{12} between the two images as:

$$s_{12} = f_V(I_1) \cdot f_V(I_2) \approx f_T(A_1) \cdot f_T(A_2),$$

where "$\cdot$" denotes the cosine similarity between the feature embeddings. Here, f_T represents a textual encoder that produces embeddings for the corresponding captions.

In Sec. 3.1, we describe how to construct the relational dataset, including how to sample image $\{I_i\}_{i=1}^N$ and generate their corresponding anonymous captions $\{A_i\}_{i=1}^N$. Then, in Sec. 3.2, we detail the training procedure for f_V.

3.1. Creating a Relational Dataset

Filtering interesting images $\{I_i\}_{i=1}^N$. Not all images are equally informative with deep logic for learning relational

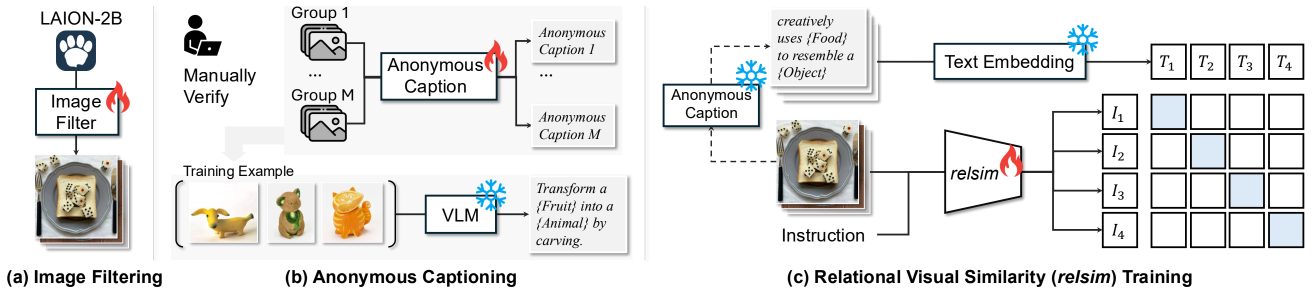


Figure 2. **Overall pipeline.** (a) We train an image filtering model to select high-quality relational images from LAION-2B [18]. (b) Anonymous captioning model is trained on groups of images that share the same underlying logic, pairing all images in each group with the same anonymous caption. (c) Training relational visual similarity (*relsim*) model involves a contrastive loss between image features and their corresponding anonymous captions.

Figure 3. Examples of relationally interesting vs. ordinary images.

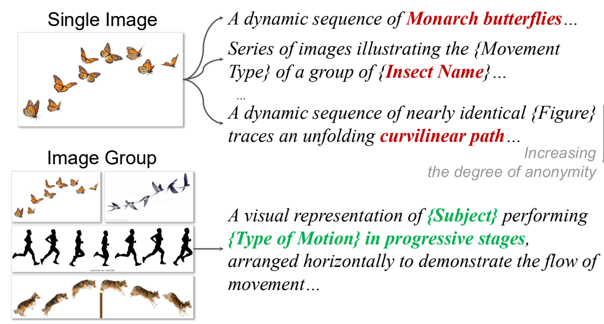


Figure 4. Writing an anonymous caption is hard from a single image, but easier with an image group where the pattern is clear.

structures. For instance, an image of a single sofa merely conveys surface-level object appearance, offering limited deep cues about relational organization. In contrast, a photo of "strawberry heart" expresses creatively compositional relations that can be abstracted and transferred to new visual content (e.g., "walnut brain", Fig. 3, second row).

Given the vast nature of LAION-2B, we first perform a filtering step to identify images potentially containing higher-order relational cues (which we refer to as *interesting* images). We fine-tune Qwen2.5-VL-7B-Instruct [41] to classify whether an image is relationally interesting, using 1.3k positive and 11k negative human-labeled examples (Fig. 2a). Annotators were instructed: "Can you see any relational pattern, logic, or structure in this image that could be useful for creating or linking to another image?". The fine-tuned model achieves 93% agreement with human judgments, and when applied to LAION-2B, it yields $N = 114k$ images identified as relationally interesting. Details of the prompt and model configuration are provided in the Supp.

Generating anonymous captions $\{A_i\}_{i=1}^N$. Writing a shared relational attribute from a single image is inherently challenging. For example, given only a sequence depicting a butterfly's flight stages (Fig. 4, first row), it is unclear which visual details are irrelevant and which constitute the underlying relational pattern. In contrast, when this image is shown alongside others expressing the same logic (Fig. 4, second row), the shared relational structure becomes immediately apparent, making it easy to articulate a caption that abstracts away object specifics.

Motivated by this observation, we manually curate $M = 532$ groups of images, where all images within a group exhibit the same underlying relational logic or pattern. Each group has N_g images (a minimum of 2 and a maximum of 10 images). We present each full group to an frozen VLM and prompt it to produce a single anonymous caption A^g—a relational description that avoids object-specific terms by replacing them with placeholders (e.g., {subject}). This caption is then human-verified and paired with every image in the group, yielding an anonymous training dataset (Fig. 2b):

$$\{(I_i^g, A^g) \mid i = 1, \dots, N_g\}_{g=1}^M$$

This procedure encourages the model to assign similar anonymous captions to images expressing the same relational pattern. We use Qwen2.5-VL-7B-Instruct [41] to train this captioning model. After training, we apply it to all "interesting" images identified in the previous step, yielding a dataset consisting of images annotated with anonymous relational captions, $\{I_i, A_i\}_{i=1}^N$, where $N = 114,881$ to be exact.

3.2. Modeling Relational Visual Similarity

Objective. Given the collection of relationally interesting images with their corresponding anonymous captions $\{(I_i, A_i)\}_{i=1}^N$, we train a visual extractor f_V with a frozen text encoder f_T to produce normalized embeddings:

$$v_i = \frac{f_V(I_i)}{\|f_V(I_i)\|}, \quad t_i = \frac{f_T(A_i)}{\|f_T(A_i)\|}.$$

We compute the similarity between an image and its anonymous caption using a dot product scaled by a learnable temperature parameter $\tau > 0$:

$$s_{ij} = \frac{v_i^\top t_j}{\tau}.$$

For a batch of size B, we use the InfoNCE training loss [2]:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \left[-\log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})} \right]$$

This training paradigm encourages the visual extractor to capture relationally meaningful features that align with the abstract concepts represented in the anonymous captions.

Model Selection. Traditional visual similarity methods rely on pure vision encoders (e.g., [2, 20, 22]), which derive representations solely from attribute-level features. We find these vision-only encoders insufficient for capturing *relational similarity*, even after tuned, as relational reasoning goes beyond mere visual recognition (See 4.2).

To address this, we leverage Vision Language Models (VLMs) for two reasons: (1) vision encoders emphasize visual attributes or semantics, which can conflict with relational understanding; and (2) relational reasoning often requires higher-level semantic knowledge—which can be found nowhere better than in a Large Language Model, where it was already trained with world knowledge. Accordingly, we employ a VLM as our visual extractor f_V (Fig. 2c). Optionally, the task–instruction can be paired with the image as a fixed, steering prompt (e.g., "Carefully analyze image to understand its underlying logic...").

4. Experiments

We now discuss our experimental settings, baselines, and evaluation protocol, followed by additional analyses.

4.1. Settings

Implementation. We adopt Qwen2.5-VL-7B-Instruct [41] as our visual feature extractor f_V. Specifically, we append a learnable query token to the end of the image as instruction token, and feed them together into the LLM. We use the query token's feature from the LLM's last layer as our visual relational feature. For the text embedding model f_T, we use all-MiniLM-L6-v2, a widely used and efficient pre-trained model from the Sentence-Transformers library [49]. We train Qwen2.5-VL-7B-Instruct with LoRA [50] for 15k iterations on a single node with 8$\times$A100 GPUs and a batch size of 64.

Data. To ensure complete separation between training and evaluation, we randomly split the dataset of 114k images into 100k for training and 14k for evaluation. For evaluation, we consider the image retrieval setting. Specifically, given a query image, we retrieve the most similar image from the database (excluding the query itself); ideally, the retrieved image should be *relationally similar* to the query. The database consists of the 14k images from the test set, combined with another 14k new images randomly sampled from LAION-2B [18] to better approximate a real-world database. From this database, 1000 images are randomly chosen from 14k test set to serve as query images.

Evaluation protocol. We employ GPT-4o [42] as an automated judge to evaluate retrieval results. For each query image and retrieved image pair, GPT-4o is prompted to assign a relational similarity score on a scale from 0 to 10, where 10 indicates highly relationally similar and 0 indicates no similarity (See Supp. for full prompt). Along with this automatic evaluation, we conduct a user study to capture human preferences. Participants are shown a query image along with two retrieved images: one from ours and one from a baseline method (randomly named as A or B)—and are asked to select which retrieved image is *relationally more similar* to the query (A, B, or Same). For each baseline, we randomly constructed 300 triplets, and each triplet was independently evaluated by at least three users, resulting in approximately 900 responses per baseline. This study allows us to quantify the proportion of cases in which users prefer our retrieval results over the baselines.

Baselines. We compare our approach with prominent image similarity metrics, including LPIPS [1], DINO [20], dreamsim [22], and CLIP-I [2] (image-to-image). These models can directly output similarity scores for a pair of images. We also consider baselines that operate via captions. In these settings, we first prompt Qwen [41] to generate an anonymous or abstract caption for each image, and then perform retrieval using this caption as the query feature. We evaluate two variants: (1) Apply CLIP-based text-to-image retrieval denoted as CLIP-T; and (2) Text-to-text retrieval denoted as Qwen-T. Note that in both of these caption-based baselines, we use the original Qwen model rather than our finetuned version. This allows us to show the performance of prompting a VLM to produce the anonymous caption from a single image (see Fig. 4) whereas finetuned model is our method which benefits from a group of images.

4.2. Evaluations

Can existing metrics capture relational similarity? Results are presented in Fig. 6, where higher values indicate better performance. As shown, LPIPS [1], which focuses purely on perceptual similarity, achieves the lowest score (4.56). DINO [20] performs only slightly better (5.14), likely because it is trained solely in a self-supervised manner on image data. CLIP-I [2] yields the strongest results among the baselines (5.91), presumably because some abstraction is sometimes present in image captions. However, CLIP-I still underperforms relative to our method, as achieving a better score may require the ability to reach even higher-level abstractions, such as those in anonymous captions. Our

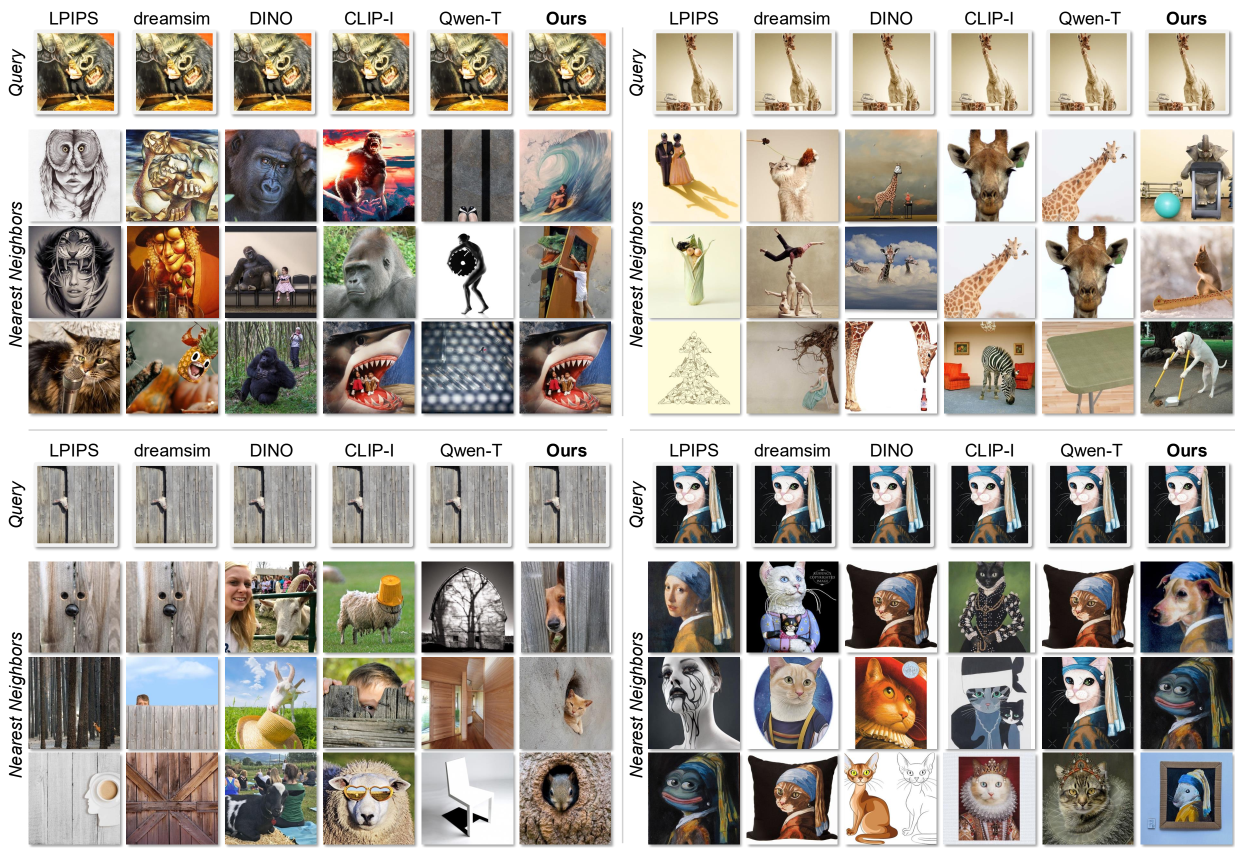


Figure 5. **Attributes vs. Relational Visual Image Retrieval.** Visualization of nearest neighbor using different visual similarity metrics. As can be seen, only ours understands and can detect the relational similarity.

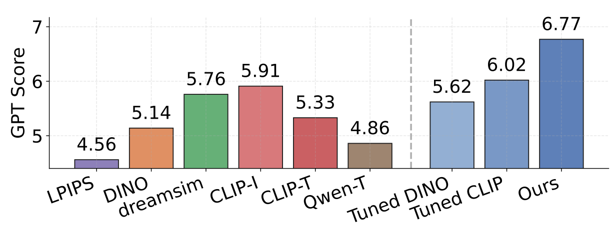


Figure 6. **Relational visual similarity performance.** All existing image similarity metrics fail to capture relational similarity, even after being tuned. Our final model (*relsim*) which leverages knowledge from VLMs, achieves the highest score (6.77).

vision encoder, being equipped with LLM knowledge and anonymous captions, yields the highest score (6.77).

Why generate anonymous captions from a group? As described in the approach section, our anonymous captions are generated from manually selected groups of similar images. Using a group makes it easier to identify the shared relational structure required for a high-quality anonymous caption. The CLIP-T and Qwen-T baselines further illustrate this point (Fig. 6): in both cases, anonymous captions are produced from a single image using the original Qwen2.5-VL-7B-Instruct [41]. We find that, under this setting, the model is hard to prompt and often leaks semantic or attribute information, causing retrieval to overly focus on semantics rather than relational similarity, thus yielding poor results (i.e., 5.33 and 4.86, compared with ours, 6.77).

Knowledge is essential for capturing relational similarity. Our argument is that relational similarity requires more than visual perception—it demands a deeper form of image understanding. Such knowledge is largely absent in vision-encoder-only models. To test this hypothesis, we conduct an ablation study in which we finetune pure vision encoders (CLIP [2] and DINO [20]) using the same anonymous captions training data and the same loss. The results (denoted as Tuned CLIP/DINO), shown in the right panel of Fig. 6, indicate that finetuning with anonymous captions does improve these models' ability to capture structural relationships. However, their performance still falls short of our model, which is equipped with a VLM. This gap is likely because VLMs, which integrate visual features with language-based world knowledge, are inherently necessary to understand and encode relational similarity.

Do humans agree with ours? The result of our user study, shown in Fig. 8, indicates that users consistently prefer our method across all baseline comparisons, with preference rates ranging from 42.5-60.7%. The gray bars indicate the tie rate. This is highly encouraging, as it demonstrates not only that our model, *relsim*, can successfully retrieve relationally

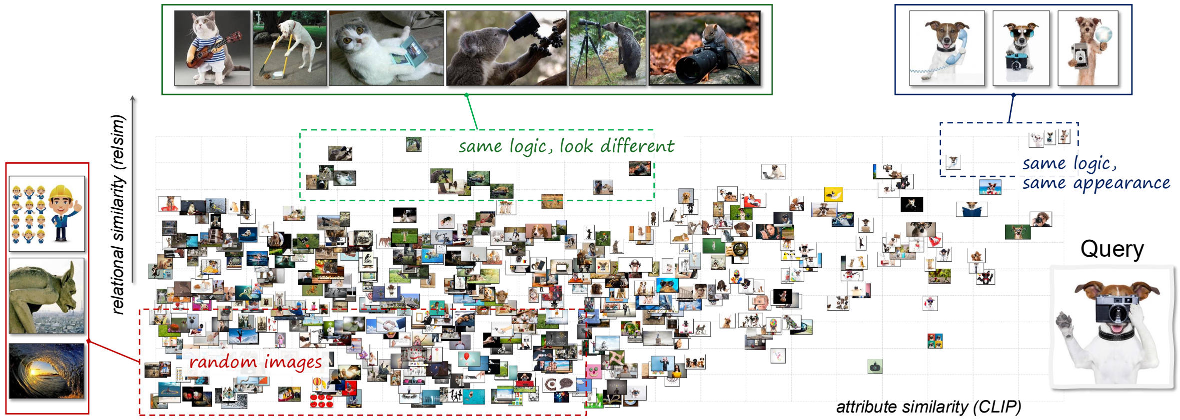


Figure 7. **Similarity space** showing different kinds of *visual similarity* in terms of degree of relational vs. attribute similarity.

similar images, but also, again, confirms that humans do perceive relational similarity—not just attribute similarity!

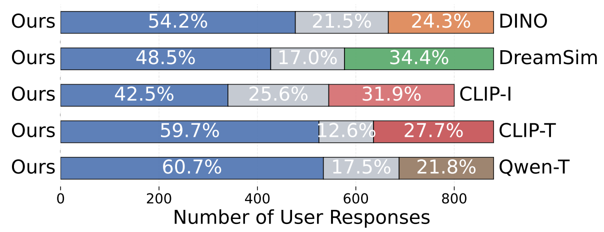


Figure 8. **User study.** AB testing shows that our model aligns significantly better with human perception of relational similarity compared to the baselines.

Relational similarity complements attribute similarity. At this point, a skeptical reader might ask: then, when to use relational, when to use attribute similarity? The answer is not straightforward. Relational and attribute similarities serve different but complementary roles: while they are often considered separate, combining them can reveal richer structures in visual data. Inspired by the similarity theory [12], we visualize visual similarity space using a query image "A dog holding a camera", and random 3000 images compared to it (Fig. 7). As shown, combining these two aspects of similarity allows us to discover interesting relationships: (1) same logic, same appearance: other photos of similar-looking dogs performing human-like activities; (2) same logic, look different: images of other {animal} performing human-like activities; and (3) random images: most other images fall into this category. This result shows that relational and attribute similarities are, perhaps, most powerful when used together rather than in isolation.

5. Applications

In this section, we illustrate scenarios where relational image similarity is useful for downstream applications, including, but not limited to, the examples below.

Relational image retrieval. Relational similarity im-

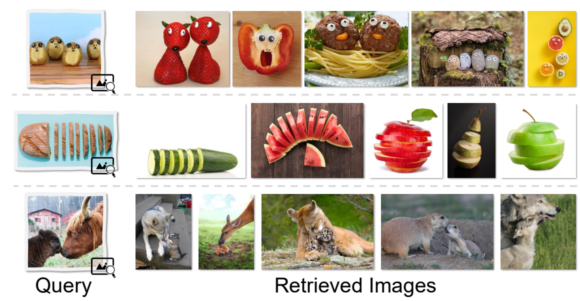


Figure 9. **Relational image retrieval**. We demonstrate that image can also be searched based on logic or abstraction (relational-based), not only perceptual or semantic similarity.

proves retrieval performance in scenarios where attribute-based matching fails, allowing users to search for images not only by semantics but also by higher-level interactions and functions between elements. This approach makes retrieval more aligned with human intuition, which is especially useful for inspiration or creativity. For example, a user might want to retrieve images showing a similarly creative way to decorate a food item with human eyes (Fig. 9, first row).

Analogical image generation. Relational similarity extends image manipulation beyond surface attributes, allowing the transfer of deeper relational structures and conceptual ideas rather than just shape or texture, unlike conventional image editing. For example, Fig. 11 (second row) shows a visual pun realized through typography (i.e., "ice-scream"); users may wish to generate new images conveying the same concept without predefined constraints on objects or attributes. Evaluating how well current image-editing or MLLM-based methods preserve such relational structures is challenging, but relational similarity provides a promising framework for addressing this gap.

To test this, we manually collected 200 image pairs sharing underlying ideas or logic, along with corresponding human-written text instructions, forming triplets: {"Input", Text Instruction", "Example Output"} (Fig. 10, first three columns). Each triplet reflects a setting where a user pro-

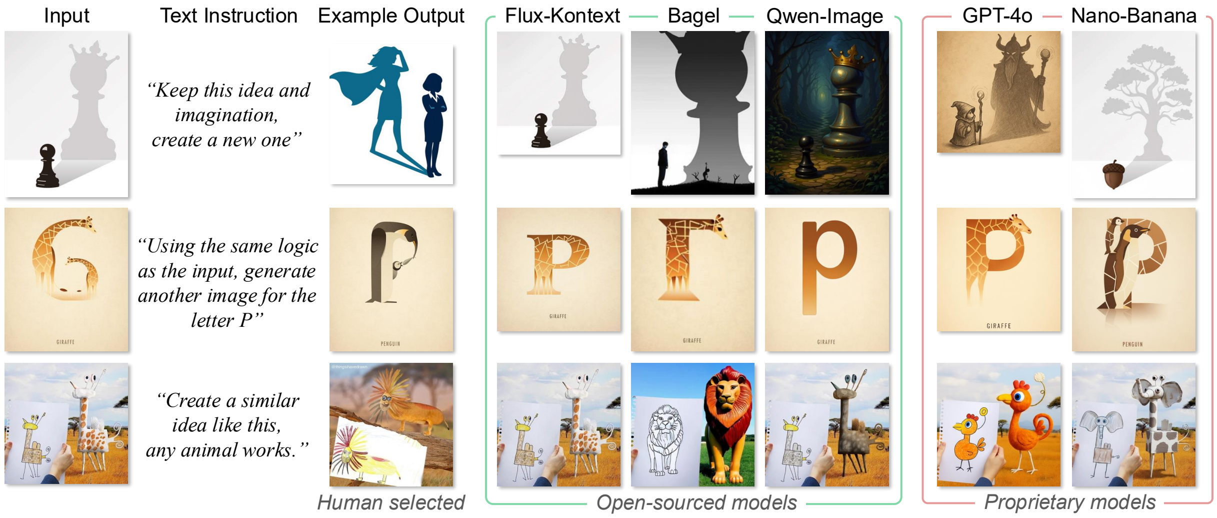


Figure 10. **Qualitative results for analogical image generation**. Proprietary models are generally better at understanding and performing sophisticated relational transformations, while open-sourced models still lag behind.

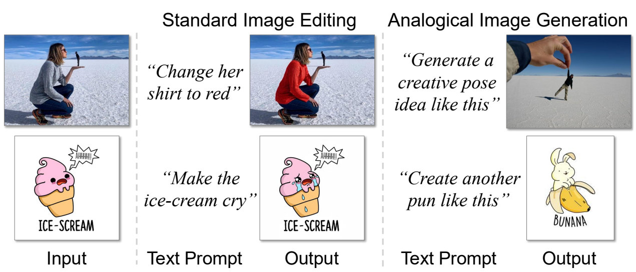


Figure 11. **Analogical image generation**. Unlike standard image editing, which modifies surface attributes, analogical generation transfers deeper relational structures and conceptual ideas.

Model	LPIPS (↓)	CLIP (↑)	relsim (↑)
Open-sourced model			
FLUX-Kontext [51]	**0.28 ± 0.22**	**0.87 ± 0.12**	0.71 ± 0.26
Bagel [43]	0.32 ± 0.19	0.79 ± 0.12	0.74 ± 0.21
Qwen-Image [52]	0.29 ± 0.21	0.86 ± 0.13	0.71 ± 0.22
Proprietary model			
GPT4o-Image [42]	0.47 ± 0.15	0.77 ± 0.10	0.82 ± 0.14
Nano-Banana [53]	0.41 ± 0.20	0.78 ± 0.11	0.84 ± 0.11
Example Output	0.60 ± 0.17	0.66 ± 0.11	**0.88 ± 0.11**

Table 2. **Quantitative benchmarking of analogical image generation**. LPIPS, CLIP and relsim measure perceptual, semantic, and relational similarity, respectively, between input and edited images.

vides an input image and a text instruction to generate a new image capturing the same underlying idea or logic. The results (Tab. 2) benchmark open-source and proprietary models using CLIP-I [2], LPIPS [1], and relsim scores to evaluate semantic, perceptual, and relational structure preservation. Key findings: (i) Example Outputs can be logically similar to the Input Image (highest relsim: 0.88) while visually differing or belonging to different semantic classes (lowest CLIP: 0.66, highest LPIPS: 0.60), showing that preserving the underlying idea can be more important than visual similarity. (ii) Open-source models tend to preserve visual similarity (i.e., CLIP: 0.8x) but often miss logical transformations compared to closed-source models (relsim: 0.7x vs. 0.8x) (see Fig. 10). These results highlight both the performance gap between proprietary and closed-source models; and the need for more challenging analogical image generation datasets to improve open-source model training.

6. Conclusion and Discussion

We have proposed *relsim*, a metric modeling *relational visual similarity*—an important aspect of visual understanding that has been largely overlooked. We show that relsim captures image logic and abstraction, which are not effectively measured by existing attribute-based similarity metrics. We further demonstrate several applications of relsim, including visual exploration (image similarity space), image retrieval, and analogical image generation.

That said, our paper is not without limitations. First, the anonymous captioning model is currently trained on 532 manually curated image groups, which may be imperfect, potentially biased, and not scalable. Developing an automated, scalable pipeline to expand these image groups, or relational logics, is an important direction for future research. Second, like other VLMs, the anonymous captioning model can exhibit biases or hallucinations, which can lead to some incorrect captions. Last but not least, we acknowledge that one image can embody multiple different relational structures, potentially leading to multiple valid relational mappings. Determining how to use text prompts to specify which relational structure a user intends remains an open question. Nevertheless, our paper highlights relational visual similarity—an overlooked aspect of image similarity—and we hope to open new avenues for future research in relational understanding and generation for vision systems.

Acknowledgment

This work was supported in part by NSF IIS2404180, NetApp Inc., and Institute of Information & communications Technology Planning& Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration, (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training), and (No. RS-2025-2543949. Environment-Aware and Domain-Adaptive Multimodal Embodied AI for Real-World Interaction).

References

[1] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1, 2, 3, 5, 8

[2] 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. In *ICML*, 2021. 1, 2, 3, 5, 6, 8

[3] Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Respects for similarity. *Psychological review*, 1993. 2, 3

[4] Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 2019. 2

[5] Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, 1990. 2

[6] Arthur B Markman and Dedre Gentner. Structural alignment during similarity comparisons. *Cognitive psychology*, 1993. 2

[7] Roger N Shepard. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, 1967. 2

[8] Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 1986. 2

[9] Amos Tversky. Features of similarity. *Psychological review*, 1977. 2, 3

[10] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 1983. 2, 3

[11] Dedre Gentner. Analogical learning. *Similarity and analogical reasoning*, 1989. 2

[12] Dedre Gentner and Arthur B Markman. Structure mapping in analogy and similarity. *American psychologist*, 1997. 2, 3, 7

[13] Keith J Holyoak and Paul Thagard. *Mental leaps: Analogy in creative thought*. MIT press, 1996.

[14] Dedre Gentner. Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive science*, 2010. 2, 3

[15] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 2, 3

[16] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeuRIPS*, 2022. 2, 4, 5

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[20] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 5, 6

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 3

[22] 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. In *NeurIPS*, 2023. 2, 3, 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3

[24] Amir Rosenfeld, Markus D. Solbach, and John K. Tsotsos. Totally looks like - how humans compare, compared to machines. In *CVPRw*, 2018. 2

[25] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007. 2

[26] Robert L Goldstone. The role of similarity in categorization: Providing a groundwork. *Cognition*, 1994. 2

[27] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv*, 2021. 2

[28] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv*, 2023. 2

[29] Robert L Goldstone and Ji Yun Son. Similarity. *The Oxford Handbook of Thinking and Reasoning*, 2012. 3

[30] Amos Tversky and Itamar Gati. Studies of similarity. In *Cognition and categorization*, 2024.

[31] Ulrike Hahn and Nick Chater. Concepts and similarity. In *Knowledge concepts and categories*, 2013. 3

[32] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, 2017. 3

[33] Ian Sample. Stephen hawking: 'there is no heaven; it's a fairy story'. *The Guardian*, 2011. Accessed: 2025-11-09. 3

[34] 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*, 2004. 3

[35] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 2011. 3

[36] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, 2018. 3

[37] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, 2020. 3

[38] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 3

[39] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *CVPR*, 2023. 3

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 3

[41] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv*, 2025. 4, 5, 6

[42] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv*, 2024. 5, 8

[43] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv*, 2025. 8

[44] Sicheng Mo, Thao Nguyen, Xun Huang, Siddharth Srinivasan Iyer, Yijun Li, Yuchen Liu, Abhishek Tandon, Eli Shechtman, Krishna Kumar Singh, Yong Jae Lee, et al. X-fusion: Introducing new modality to frozen large language models. In *ICCV*, 2025.

[45] Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, and Yuheng Li. Yo'chameleon: Personalized vision and language generation. In *CVPR*, 2025.

[46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv*, 2023.

[47] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. In *NeurIPS*, 2024.

[48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 3

[49] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv*, 2019. 5

[50] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 5

[51] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv*, 2025. 8

[52] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv*, 2025. 8

[53] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*, 2025. 8