Show Me Examples: Inferring Visual Concepts from Image Sets

Nick Stracke*[12], Kolja Bauer*[12], Stefan Andreas Baumann[12], Miguel Ángel Bautista[3], Josh Susskind[3], and Björn Ommer[12]

[1] CompVis @ LMU
[2] Munich Center for Machine Learning
[3] Apple

Abstract. Vision-language models (VLMs) can follow complex textual instructions, yet they struggle to reason from purely visual context. In particular, current models fail to infer shared concepts from sets of example images and apply them to new inputs. We introduce Visual Concept Inference from Sets (VICIS), a task that evaluates this capability. Given a small context set of images sharing a concept and a query image, the model must generate new images that preserve the context-defined concept while remaining consistent with the query. We show that state-of-the-art VLMs perform poorly on this task, often ignoring the visual context or defaulting to biased generations. To address this gap, we propose a training framework and architecture that learn to infer visual concepts from image sets and extract concept-specific embeddings from queries. Experiments on synthetic data and large-scale ImageNet/WordNet data show that our model generates more accurate and diverse outputs and generalizes to unseen concepts and modalities such as sketches.

Keywords: Visual Reasoning · Concept Inference · VLMs

1 Introduction

Generative vision-language models (VLMs) have advanced rapidly, enabling strong performance in image captioning, visual question answering, and text-conditioned image synthesis [2, 12, 31]. A key capability underlying the success of large language models (LLMs) is in-context learning: the ability to adapt behavior at test time by conditioning on a small set of examples, without parameter updates [17]. If visual models are to exhibit a similar form of generalization, they must be able to infer what matters directly from visual context, even when the underlying concept is not easily verbalized.

Despite impressive progress, *visual* in-context learning remains limited. Recent proprietary image models [7, 19, 40] can follow many natural language instructions, yet they often fail when the instruction is specified implicitly through images

* Equal Contribution

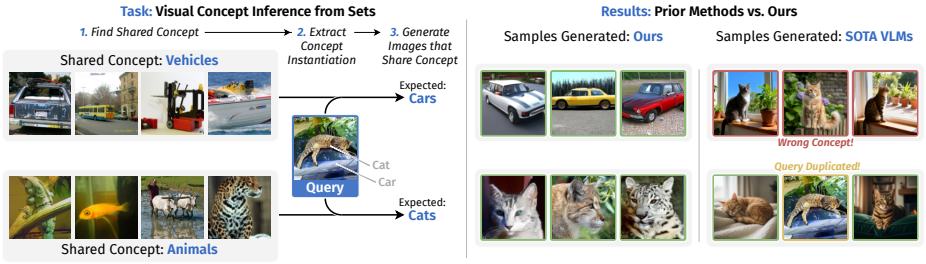


Fig. 1: Visual Context Inference from Sets (VICIS). *(left)* We introduce VICIS, a task aimed at evaluating the visual reasoning capabilities of vision models. Inspired by human capabilities, the model needs to infer a shared visual concept from a small set of example images and extract it from a query image to generate new images that preserve only this concept. *(right)* Our model adapts to different contexts for the same query (e.g., *vehicles* vs. *animals*), demonstrating visual reasoning over context and query. In contrast, Nano Banana largely ignores the visual context and instead defaults to reproducing a dominant concept from the query.

(Figure 2). In these cases, models frequently ignore the visual context, collapse to reproducing one of the inputs, or default to biased generations, suggesting that they do not reliably reason over sets of images as contextual evidence.

In this work, we isolate a concrete capability that current models struggle with, and that is central to visual reasoning: inferring a shared concept from a small set of example images and relating it to a new image. Humans routinely perform this type of reasoning: given a few observations, we can infer which aspects are shared and use that understanding to interpret new situations. Successfully solving this task requires identifying common structure across a set of images, separating relevant from irrelevant variation, and transferring the concept to a novel image.

Concretely, we introduce the task of *Visual Concept Inference from Sets (VICIS)* (Figure 1) to capture this form of visual reasoning. The model is given (i) a small *context set* of images that share a common concept and (ii) a *query* image. The goal is to generate new images that reflect the concept specified by the context while remaining consistent with the relevant aspect of the query. The query plays a crucial role: it provides a concrete instantiation of the concept, allowing us to assess whether the model has correctly inferred the shared structure of the context set. Without it, a model could solve the task by generating another image that resembles the context set, turning it into a task of summarizing examples rather than reasoning.

We chose image generation as the task interface for two reasons. First, it provides a visual readout of whether the model correctly uses the context set: success requires producing outputs consistent with the context-defined concept while avoiding trivial solutions, such as copying the query. Second, it allows direct, apples-to-apples comparison to image models and VLMs that also operate in the image space, revealing a gap between general-purpose instruction following and true visual in-context reasoning (Figure 2).

To address this gap, we design a training framework and architecture that enable us to specifically train and test this visual reasoning capability on the

VICIS task. Training such a model requires constructing context sets and query-target relations that reflect shared but unlabeled concepts. We achieve this using a weak supervisory structure derived from auxiliary groupings, such as the WordNet hierarchy of ImageNet, which provides a scalable scaffold for assembling training episodes without per-concept annotation. At test time, the model operates purely on visual input: it must infer from the context set which variation is relevant and apply that inferred structure when conditioning generation on the query.

In addition to image generation, our model also predicts a compact embedding of the inferred concept. This representation provides a more direct readout of the model's internal reasoning about the context set and allows us to analyze the structure of the inferred concept space.

We evaluate our approach on controlled synthetic data and on a large-scale hierarchical dataset constructed from ImageNet/WordNet. Across qualitative and quantitative experiments, our model generates outputs that are both more accurate and more diverse than strong baselines, and it generalizes to challenging settings such as sketches and previously unseen classes. These results indicate that models can be trained to use visual sets as nonverbal specifications of concepts, enabling us to express intent through examples rather than language alone.

We summarize our contributions as follows:

- We introduce *Visual Concept Inference from Sets (VICIS)*, a task that probes nonverbal visual in-context learning: inferring a concept from a small set of images and applying it to a query without text, labels, or fine-tuning.
- We propose a training framework that uses weak supervisory signals from auxiliary groupings to construct scalable learning episodes for this task.
- We present a training framework and architecture that demonstrates improved accuracy and diversity in context-guided generation, including generalization to sketches and unseen classes.

2 Related Work

Visual In-Context Learning Visual in-context learning refers to the ability of visual models to adapt to new tasks using only examples provided during inference, without additional training. Although extensively studied in textual domains with large language models (LLMs) [9, 17], visual in-context learning remains relatively under-explored. Recent VLMs have begun addressing this challenge, yet only a handful support both unified understanding and image generation, meaning they can reason over multimodal text-image inputs and generate images. BAGEL [15] and ILLUME+ [25] offer this functionality and currently represent the state of the art in unified image understanding and generation among open-source models. Very recently, the largest proprietary models have showcased surprising multi-modal capabilities. Yet, they still fail at correctly solving our proposed VICIS task, as shown in Tab. 4. Other approaches that explicitly target visual in-context learning, such as Visual Prompting via Image Inpainting [6], Improv [41], sequential prompting [4], and Hummingbird [5], adapt via visual cues or prompts but require explicit in-context task demonstrations.

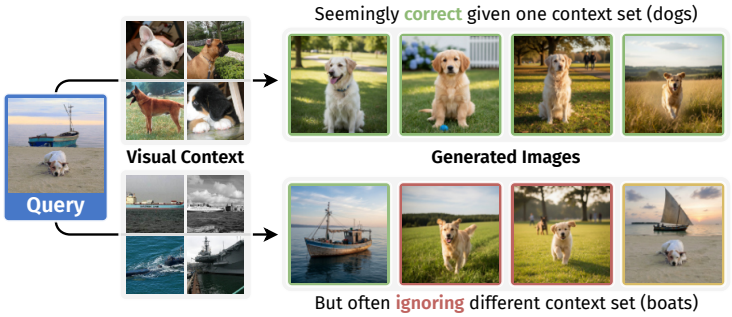


Fig. 2: Typical VLM failure cases on our VICIS task: The model largely ignores the visual context and fixates on a single concept of the query image, in this case the dog. While the first row may suggest that it has understood the task, the second row reveals a bias toward specific concepts rather than genuine task comprehension.

Our approach differs by removing the reliance on explicit in-context examples. Instead, our model implicitly discovers relevant concepts through provided image sets, closely resembling how humans learn from contextual examples. This framing provides a more natural and flexible mechanism for visually specifying complex, subtle, or otherwise difficult-to-describe concepts, significantly extending the scope and practicality of visual in-context learning.

Inferring Latent Factors and Representations A large body of work studies how to uncover latent factors of variation that explain visual data at train time, often aiming for representations in which different dimensions correspond to semantically meaningful properties such as pose, lighting, or identity. Unsupervised latent variable models such as β-VAE [24] and InfoGAN [11] introduce inductive biases, e.g., KL regularization or mutual-information maximization, to encourage interpretable latent structure, but are known to be non-identifiable in general [28, 33]. More recent work addresses this limitation by establishing conditions under which latent factors become identifiable, for example, by leveraging auxiliary observed variables such as iVAE [28], or by studying related assumptions in causal representation learning [43]. In contrast to these approaches, the goal of solving the VICIS task is not to recover a universal factorization of images or identifiable latent sources of the full data distribution. Instead, the relevant concept needs to be inferred from a small context set of images at inference time, with the goal of extracting it from the query image.

Controllable generation and concept editing. Another related line of work focuses on controlling generative models along semantic directions. They analyze the latent space of a pretrained generator and identify directions that induce consistent semantic changes in the output [22, 27]. Analyses of GAN latent spaces have shown that simple traversals or learned directions can induce meaningful changes in generated images [22, 27]. In diffusion models, recent work similarly finds semantic directions for more controlled generation by leveraging

the conditioning mechanism [8], or by learning concept-specific adapters for precise control [20, 21]. These methods typically assume that the target concept is specified explicitly (e.g., via text, curated example pairs, or editing objectives) and often require additional optimization or training to enable control for a given concept [8, 20, 21]. In contrast, VICIS studies the complementary problem of *inferring* the concept itself from a small set of visual examples without language or labels, and transferring it to a new query instance, without additional concept-specific training at inference time.

Measuring Visual Intelligence Recent benchmarks like the Abstract Reasoning Corpus (ARC) [13], CLEVR [26], and physics-based reasoning tests [35] assess model intelligence by evaluating reasoning abilities on visually grounded challenges. However, these approaches typically focus on explicit input-output mappings within well-defined constraints. Our proposed VICIS task is uniquely flexible, simulating human cognitive tests by implicitly defining visual concepts through examples. It evaluates a model's capability to generalize from abstract concept definitions to specific instantiations without explicit input-output pairs, aligning more closely with human visual intelligence assessment.

3 Method

Our goal is two-fold. We first want to assess the visual reasoning capabilities of current vision models. Concretely, we design a task called *Visual Concept Inference from Sets* (VICIS), where the model learns to identify shared visual concepts within image sets, as outlined in Sec. 3.1. Secondly, we propose an architecture to directly solve the VICIS task, thereby providing an intuitive and effective mechanism for explicitly instructing a visual model about concepts of interest in Sec. 3.2.

3.1 Task Definition: Visual Concept Inference from Sets

We frame our learning problem as a visual in-context task, where a model observes a small set of related images and is expected to infer the underlying visual concept they share. This setup simulates a natural form of visual learning, in which a user conveys a concept by providing example images, without relying on language or labels.

Formally, let an image $\mathbf{x} \in \mathbf{X}$ be associated with a set of visual concepts $C_{\mathbf{x}} \subseteq C$, where $C = \{c_1, c_2, \ldots, c_n\}$ denotes the complete space of possible visual concepts that can be present in an image. Each concept c_i can have multiple possible instantiations, such as *car type: sedan*; *surface finish: metallic*; or *color: blue*.

To define our VICIS task, we construct a *context set* $\mathbf{X}_{\text{set}}$ consisting of several images that share a common (but unlabeled) concept. For example, the context set might contain images of various dog breeds, objects in bright lighting, or items made of wood. This shared structure encourages the model to infer a latent

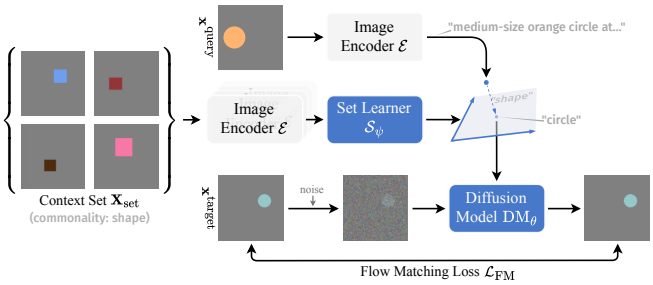


Fig. 3: Model Overview. Given a set of images that share a concept (here, *shape*), the set learner predicts a concept space that captures all possible instantiations of that visual concept. We project the embedded query image in that concept space to obtain its concrete instantiation (here, *circle*) and remove all other concepts that cannot be represented in that space. Finally, the diffusion model is conditioned on the projected query to denoise the target image.

concept that captures the variation across the set. The goal is for the model to identify and internalize the relevant visual dimension that unifies these examples.

In order to verify whether the model truly understands and precisely identifies the intended concept, we introduce two images: a query image $\mathbf{x}^{\text{query}}$ and a target image $\mathbf{x}^{\text{target}}$. Both the query and target images share the same specific instantiation of the concept (e.g., the same breed of dog). The role of the query image is to specify an instantiation of the relevant concept defined by the set, as illustrated in Figure 6. The target image then acts as the correct match or answer, sharing this exact instantiation.

To enable learning of this behavior, we require a mechanism for constructing context sets and corresponding query-target pairs that reflect shared concepts. This relies on auxiliary information that enables us to group images based on shared semantics. While this introduces a form of indirect supervision, the key idea is to teach the model how to infer structure from sets of images alone. At inference time, the model is presented with only a context set and a query image, and must reason about the relevant visual concept based purely on visual input.

3.2 Solving the VICIS Task

Our proposed method to solve the VICIS task comprises three components: (1) the *Set Learner*, (2) the *Instantiation Stage*, and (3) the *Diffusion Model*. See Figure 3 for an illustration. Each component plays a specialized role in our framework to discover, encode, and utilize meaningful visual concepts.

Set Learner The Set Learner identifies concept-specific directions in the feature space. Each image in the input set $\mathbf{X}_{set}$ is first encoded individually using a pretrained Vision Transformer $\mathcal{E}$ [10, 36, 37], producing feature token embeddings. The Set Learner itself is a separate Vision Transformer (ViT) that takes the combined embeddings of the entire set of images as input. It reasons across these embeddings jointly to identify the shared concept and subsequently predicts a set

of direction vectors $\mathbf{D}_c = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k\}$. These directions define the embedding space corresponding to the shared concept identified within the set.

Instantiation Module Given a query image $\mathbf{x}^{\text{query}}$, we encode it individually using the same pretrained encoder $\mathcal{E}$, but utilize only the CLS token embedding $\mathbf{e}^{\text{query}}$. To isolate the concept-specific information identified by the Set Learner, we project this embedding onto the embedding space spanned by the concept directions $\mathbf{d}_i$ through dot products:

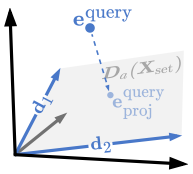


Fig. 4: Instantiation Illustration.

$$s_i = \langle \mathbf{e}^{\text{query}}, \mathbf{d}_i \rangle, \quad \forall \mathbf{d}_i \in \mathbf{D}_c \tag{1}$$

These scalars $\mathbf{s} = \{s_1, s_2, \ldots, s_k\}$ represent the query's concept instantiation within the broader concept space. Multiplying each scalar s_i by the respective direction $\mathbf{d}_i$ and aggregating gives a single concept-conditioned token:

$$\mathbf{e}^{\text{query}}_{proj} = \sum_{i=1}^{k} s_i \mathbf{d}_i \tag{2}$$

This resulting token $\mathbf{e}^{\text{query}}_{proj}$ explicitly captures only the concept-specific information from the query, discarding unrelated visual details (e.g., background or unrelated objects).

Diffusion Model The diffusion model DM_θ is conditioned on $\mathbf{e}^{\text{query}}_{proj}$ to generate target images $\mathbf{x}^{\text{target}}$. Specifically, $\mathbf{e}^{\text{query}}_{proj}$ is added directly to the timestep embedding of the diffusion model, ensuring strong concept-conditioned generation.

To train the entire setup, we use the rectified flow objective [3, 30, 32], which frames generative modeling as matching a continuous vector field. Specifically, given a data distribution $p(\mathbf{x})$ and an initial noise distribution, the diffusion model DM learns a time-dependent vector field $\mathbf{v}(\mathbf{x}, t)$ that transports samples smoothly from noise toward the data manifold. Formally, this involves minimizing the Flow Matching loss:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, \mathbf{x}_t^{\text{target}}\sim p_t, \mathbf{e}^{\text{query}}_{proj}} \left[\left\| \text{DM}_\theta(\mathbf{x}_t^{\text{target}}, t, \mathbf{e}^{\text{query}}_{proj}) - \mathbf{u}(\mathbf{x}_t^{\text{target}}, t) \right\|_2^2 \right] \tag{3}$$

This end-to-end training with a flow matching loss provides a stable objective that is easy to optimize and can work well with smaller batch sizes, which sets it apart from other methods such as contrastive learning that typically requires larger batch sizes to work successfully.

4 Experiments

We implement the Visual Concept Inference from Sets (VICIS) task for both controlled synthetic data and real-world data. We benchmark our model against

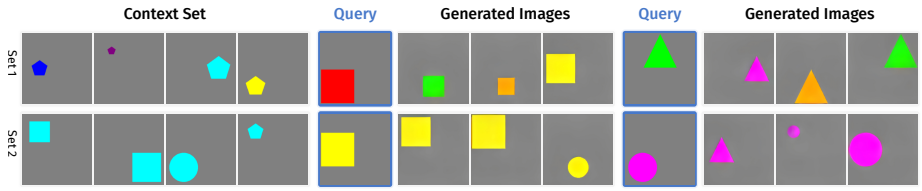


Fig. 5: Visualization of our toy dataset and results. The left third shows the context sets used for specifying a concept, here *shape* (first row) and *color* (second row). The middle and right part show three samples for two different query images (left column), respectively. The model learns to correctly identify the concept specified by the context set and extract its instantiation from the query image. Note that only the specified concept's instantiation is preserved while other concepts can vary freely, indicating that the model correctly disregards irrelevant information.

both task-specific and general-purpose baselines, including the Visual Prompting model [6] and recent vision-language models [15, 19, 25]. Our goal is to assess whether models can infer a shared visual concept from a context set and correctly extract its instantiation from a query image to generate appropriate targets. We first study the task in a synthetic setting where the data distribution is fully controlled (Sec. 4.1). We then scale the task construction to real-world data. To that end, we propose a generic scheme to assemble context sets and query-target pairs based on auxiliary groupings from the WordNet hierarchy [34]. Finally, we analyze the concept-conditional embeddings learned by our model to gain insights into the internal representations used to solve VICIS (Sec. 4.5). Implementation Details in Supp. Sec. C.

4.1 VICIS on Synthetic Data

We first study the VICIS task in a synthetic setting that provides full control over the visual concepts present in each image. Recent advances in generative models have made large-scale generation of high-quality synthetic visual data increasingly feasible [7, 19, 42]. In this work, we construct a simple synthetic dataset designed to isolate the core challenge of the VICIS task. This controlled environment allows us to precisely define shared concepts within context sets while varying irrelevant concepts. We illustrate the construction of such a VICIS task in the following paragraph and visualize the setup in Figure 5.

Dataset Construction To assess whether our model is generally able to reason about sets of images and find the shared concept in a context set, we construct a synthetic dataset that allows us to precisely obtain images with desired concepts for the context set and the query-target pair. Each image consists of a single object with concepts $C_{\mathbf{x}} = \{\text{shape}, \text{size}, \text{location}, \text{color}\}$. We construct context set, query and target such that one concept's instantiation is shared within the context set. Query and target share a potentially different instantiation of that concept (see Figure 5).

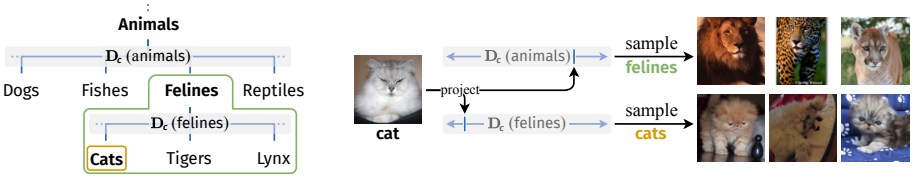


Fig. 6: Hierarchical Concept Spaces. A simplified visualization of our ImageNet hierarchy, which is loosely based on the WordNet hierarchy. We define a concept space with respect to the next hierarchy level. The $\mathbf{D}_c$(animals) concept space differentiates roughly between different types of animals whereas $\mathbf{D}_c$(felines) is more granular and operates on a lower level of the hierarchy.

4.2 VICIS on Real-World Data

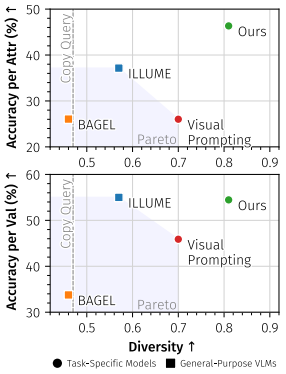
While the synthetic setting allows precise control over visual concepts, constructing VICIS tasks for real-world images is significantly more challenging. To scale the task beyond controlled environments, we propose a scheme for assembling context sets and query-target pairs using weakly labeled semantic structure derived from WordNet [34].

Dataset Construction To construct meaningful concepts, we employ a hierarchical structure derived from WordNet [34]. Each node in this hierarchy corresponds to a visual concept (e.g., *animal*), with branches representing increasingly specific concept instantiations (e.g., *mammal*→*dog*→*bulldog*). Sets are constructed at different hierarchy levels to define varying degrees of specificity for concept instantiation. For instance, if the concept *animal* is chosen, a set might contain mammals, birds, and fish. Given a query image depicting a mammal, the diffusion model should generate other mammals, since that is the next level in the hierarchy. Conversely, selecting a more specific concept (e.g., *dog*) restricts generation to matching exact breeds. This hierarchical approach addresses a fundamental challenge in concept inference: constructing training sets with meaningful relationships between context set, query, and target. While extensive annotations for concept instantiations would be ideal, such datasets are rare and challenging to create. Our hierarchical structure provides a flexible yet structured framework for visual concept inference, where the precise definition of sets directly shapes the model's learned capabilities. During training, context sets, queries and targets are constructed from the ImageNet training split, during evaluation they are assembled from the validation split.

Quantitative Evaluation We analyze our model's performance on the VICIS task and benchmark it against both task-specific and general-purpose VLMs. If the model correctly infers the relevant concept from the context set, it should generate targets that preserve the corresponding concept instantiation from the query. Thus, we analyze the images generated by our model to determine if the Set Learner has correctly inferred the relevant concept. In the hierarchical setting,

Table 1: Quantitative comparison of models on VICIS tasks. Correctly solving the VICIS task entails generating correct *and* diverse images. We investigate the models' ability to generalize to sketches from the ImageNet Sketch dataset [39] and to unseen classes from ImageNet21k [16]. We additionally visualize the results for the hierarchy dataset on the right to showcase the trade-off between accuracy and diversity that our model strikes best, significantly outperforming the pareto front.

Method	Dataset	Accuracy (%) ↑ per con.	Accuracy (%) ↑ per inst.	Diversity ↑
Baseline: Copy Query Image		–	–	0.47
ILLUME+ 3B [15]	Hierarchy	37.15	**55.00**	0.57
BAGEL 7B-MoT [25]		26.05	33.76	0.46
Visual Prompting [6]		26.02	45.89	0.70
Ours		**46.34**	54.46	**0.81**
Visual Prompting [6]	Sketch Queries	23.76	44.57	0.68
Ours		**32.60**	**47.11**	**0.72**
Visual Prompting [6]	Sketch Context	27.67	47.04	0.71
Ours		**41.50**	**52.24**	**0.79**
Visual Prompting [6]	Sketch C+Q	25.69	44.45	0.68
Ours		**31.90**	**46.47**	**0.71**
Visual Prompting [6]	21k Context	29.07	49.95	0.76
Ours		**39.63**	**51.40**	**0.79**



a correct solution entails understanding the hierarchy level and branch from the context set, identifying the query's position within the hierarchy, and generating samples from the appropriate branch. We compare our model against two state-of-the-art open-source VLMs, [15, 25]. Since these general-purpose models have to understand the task in-context, we design three different prompting schemes shown in Supp. Tab. E and ablate their performance in Supp. Tab. D. The VLM results in Tab. 1 use the best performing prompting scheme. Additionally, we compare against a Visual Prompting model [6] which we specifically train for our task, as it is the closest existing method to our approach. We train the Visual Prompting model from scratch on our proposed hierarchical dataset to ensure a fair comparison. Implementation details on how we adapt their method to the proposed VICIS task are provided in Supp. Sec. I. To quantify performance, we focus on the *animal* subtree of our hierarchy. We choose this subtree for its intuitive structure, as it follows the biological taxonomy, providing a clear hierarchy for evaluation. To quantify both the correctness and diversity of generated images, Tab. 1 reports results for both *accuracy* and *diversity*:

Accuracy: We evaluate the model's performance on the VICIS task by measuring the correctness of the generated images. To assess whether a generated image is in the correct part of the hierarchy, we use a classifier to obtain the class label and locate the predicted class within the hierarchy tree. We provide the mean accuracy averaged over concepts (*per con.*) and averaged over possible concept instantiations in the query (*per inst.*). We observe that our model significantly outperforms the Visual Prompting model as well as the VLMs on unseen validation samples from the same hierarchy used during training, as shown in the upper section of Tab. 1.

Diversity: We want the model to demonstrate understanding of the VICIS task at hand by not only producing correct but also diverse images. A trivial solution to correctly retrieve a concept's instantiation from the query is to simply retrieve all concepts' instantiations, merely copying the query. As this would not be penalized by a lower accuracy, we introduce a diversity score as a second decisive metric. This score captures the model's ability to generate diverse but correct outputs and penalizes trivial solutions. We define the diversity score as the mean ratio of entropies between generated distributions for a single concept instantiation q_{inst} projected onto two concept directions of different abstraction levels:

$$\text{Div} = \frac{1}{K \cdot Q} \sum_{q_{\text{inst}} \in Q} \sum_{k=1}^{K} \frac{\text{H}(q_{\text{inst}}, \text{concept}_k^0)}{\text{H}(q_{\text{inst}}, \text{concept}_k^1) + 1} \tag{4}$$

Q is the number of concept instantiations for the query image. K denotes the number of concept pairs where concept_k^1 is a sub-concept of concept_k^0, i.e. a direct subclass in the hierarchy (e.g. concept_k^1 = fish and concept_k^0 = animal). The entropy

$$\text{H}(q_{\text{inst}}, \text{attr}) = -\sum_{c_i \in M_{\text{concept}}} p(c_i \mid q_{\text{inst}}, \text{concept}) \, \log\big(p(c_i \mid q_{\text{inst}}, \text{concept})\big) \tag{5}$$

is calculated over all concept instantiations $c_i \in M_{\text{concept}}$ of the higher-level concept's subtree that the concept instantiation lies in. $p(c_i|q_{\text{inst}}, \text{concept})$ denotes the probability that given a context set for concept *concept* and a query with instantiation q_{inst}, the generated image's concept *concept* is instantiated as c_i. An illustration of the model's desired behaviour is provided in Figure 6: If we project a *cat* onto the concept direction *feline*, we want the model to generate diverse cats. If we project that same *cat* onto the concept direction *animal*, we want the model to generate diverse felines, not only cats. By doing so, the model demonstrates that it has correctly inferred the context set's hierarchy level.

To build intuition for our proposed diversity metric, we evaluate two naive baselines. Replicating the query image results in a diversity score of 0.47 and a near-perfect accuracy, only upper-bounded by the classifier's accuracy. Generating completely random images achieves a diversity score of 0.77, comparable to the values reported in Tab. 1, but leads to near-zero accuracy. The diversity score's theoretical upper-bound equals 1.87 in our proposed hierarchy setup.

Robustness We analyze our model's robustness with respect to context set quality and size. When varying the number of context images from two to seven, accuracy improves steadily with additional context, particularly from two to four examples, and then plateaus. This indicates that even small sets provide useful signal, while larger sets increase reliability for more nuanced concept inference. Diversity also increases with context size. Results are shown in Tab. 2. To test sensitivity to context noise, we replace one or two images out of five with random images from the dataset. As noise increases, accuracy decreases progressively, indicating that the model effectively uses relevant context while remaining robust

Table 2: Performance across context set sizes. Accuracy improves with additional context, especially in the low-data regime, and diversity increases with size.

Set Size	Accuracy (%) ↑		Div. ↑
	per con.	per inst.	
2	44.74	48.92	0.719
3	45.26	53.51	0.801
5	46.34	54.46	0.811
7	46.56	54.74	0.824

Table 3: Performance with noisy context sets. One or two of five context images are replaced with random images from the dataset.

Set Type	Accuracy (%) ↑		Div. ↑
	per con.	per inst.	
clean	46.34	54.46	0.811
noisy (1/5 rand.)	37.62	46.85	0.812
noisy (2/5 rand.)	26.01	36.42	0.906

to partial corruption. Diversity remains stable or slightly higher, suggesting no collapse under less informative inputs. Results are shown in Tab. 3. Overall, the model degrades gracefully under noise and benefits from larger context sets, which is consistent with the Fisher ratio trend discussed later in Figure 9, suggesting that larger sets reduce ambiguity. An extended analysis of performance degradation per concept is provided in Supp. Sec. B.

Generalization Capabilities We evaluate our model's ability to generalize beyond its training distribution by testing it under two challenging settings: (1) using sketch images as context and/or query inputs, and (2) providing context images from previously unseen ImageNet21k classes.

To assess whether the model has overfit to appearance cues or has instead learned to capture semantic content, we use the ImageNet-Sketch dataset [39]. We construct evaluation setups where the context set, the query, or both are drawn from sketch images. An illustration of this setup is shown in Figure 7. Quantitative results for all combinations of real and sketch images in the context and query are reported in Tab. 1. While performance decreases relative to settings using only real images, our model maintains strong accuracy and consistently outperforms the Visual Prompting baseline across all configurations.

Additionally, we test if the model can correctly infer relevant concepts given novel concept instantiations not seen during training. To that end, we construct context sets from ImageNet21k and report performance in the bottom section of Tab. 1. We observe only slight performance degradation, indicating that our model is able to generalize to unseen concept instantiations.

4.3 Performance of SOTA VLMs

Since many current SOTA VLMs are closed-source and only accessible via API providers, we perform a small-scale evaluation using the FAL.ai API [1] on our VICIS task. Since these models are unaware of the hierarchy, we probe them with an ambiguous query that shows two concepts and evaluate only whether they instantiate the correct concept, as specified by the context set, similar to Figure 1. Since the expected output of the VICIS task is images, we can directly compare

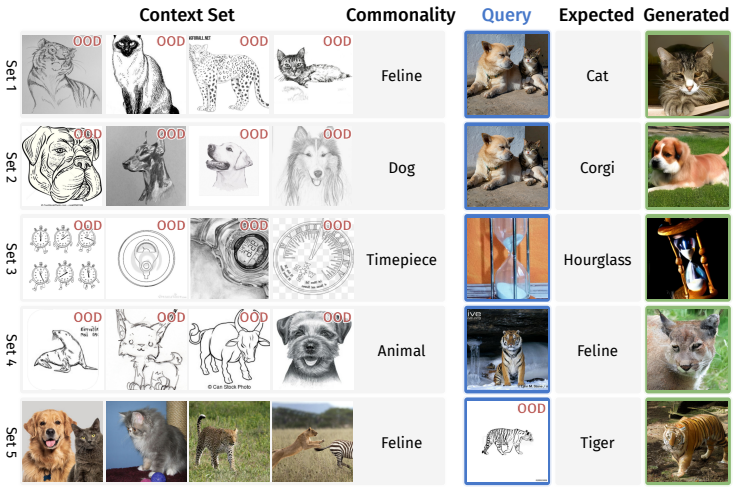


Fig. 7: When a query does not fully match the concept space defined by the context set, it aligns with the closest point on the learned manifold. Although trained only on ImageNet, the model generalizes to sketch inputs by extrapolating semantically. This applies to both query and context images. In the last two rows, we show how different hierarchy levels affect generation: projecting a tiger onto the *feline* space yields tiger-like samples, while projection onto the broader *animal* space results in more generic felines.

Table 4: Performance of current VLMs. We evaluate the performance of current SOTA VLMs on a simplified version of our VICIS task and report accuracy. Even the largest closed-source VLMs are not able to solve our task reliably. Qwen and Flux do not support multiple image inputs. Tiling multiple images into one does not lead to successful task completion.

Model	Nano Banana [19]	Gemini 2.5 Flash Image [19]	Qwen Image Edit [40]	FLUX.1 Kontext Pro [7]	**Ours**
Accuracy	46%	40%	not supported	not supported	**93%**

them with our model. The generated samples are evaluated and categorized by expert annotators to ensure correct labeling of all results and prevent bias from synthetic methods. We find that even current SOTA VLMs are not able to solve this task reliably, showing the need for a more principled approach (Tab. 4). More details can be found in Supp. Sec. D.

4.4 VICIS for Learning Transformation

While the previous experiments focused on concept inference, the VICIS task is a general formulation for visual reasoning. This is helpful because many useful visual instructions are easier to demonstrate than to name. In Figure 8, this is shown for arbitrary image transformation pairs, i.e., visual analogy from context examples $A \rightarrow A'$ to a new query. The model is trained on Relation252k [23] with additional positional encoding for any two images that define a transformation.

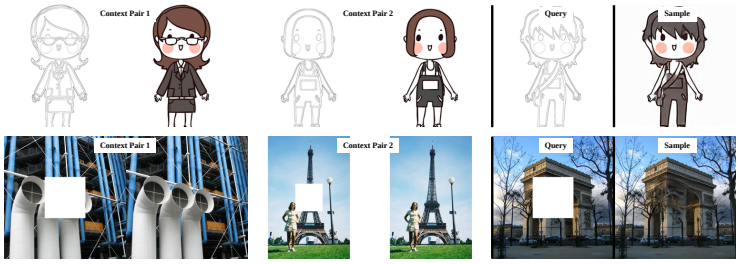


Fig. 8: We show how the VICIS task can be used for more applications, such as learning transformations. Here, the context set gives multiple examples of the desired transformation, which can then be applied to the query image. We show generations for held-out query images.

4.5 Concept-Specific Embedding Analysis

Since our model predicts an embedding of the instantiated query concept, we can analyze the structure and expressiveness of that embedding space to better understand how the model solves the task. As described in Sec. 3.2, our method predicts concept directions inferred by the Set Learner. This approach contrasts with traditional dimensionality reduction methods such as PCA and LDA, which derive directions from data statistics rather than explicitly learning concept-specific directions. Specifically, we evaluate how the number of examples in the context set influences performance, as larger sets should provide stronger signals for identifying the shared concept.

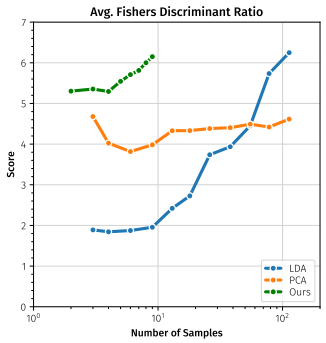


Fig. 9: Class Separation ↑ Our predicted directions better separate concepts with significantly fewer samples.

To quantify the structure of the predicted concept-specific embedding space, we use the Fisher Discriminant Ratio, which assesses class separability by comparing intra-class and inter-class variance. Formally, we compute the between-class scatter matrix $\mathbf{S}_B$ and the within-class scatter matrix $\mathbf{S}_W$, where N_c is the number of elements in class c. We then compute the trace ratio J_{trace} which provides a single scalar measure of how well-separated the concept representations are:

$$J_{\text{trace}} = \frac{\text{tr}(\mathbf{S_B})}{\text{tr}(\mathbf{S_W})}, \quad \mathbf{S}_B = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top, \quad \mathbf{S}_W = \sum_{c=1}^{C} \sum_{\mathbf{x}_i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top$$

To compare our approach with commonly used dimensionality reduction techniques, we evaluate the expressiveness of our learned concept directions against directions found by Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). To ensure a fair comparison, we average results across multiple context sets and concept spaces. We encode images using the same pretrained encoder $\mathcal{E}$ and linear projection layer as our method before

applying PCA or LDA. While PCA captures general variance across samples, LDA explicitly optimizes for inter-class separability using class labels.

For evaluation, we use the validation split of ImageNet. We randomly sample levels from our hierarchical structure to construct context sets, which in turn define concept directions. As mentioned earlier, we average results across multiple directions to enable a fair comparison. We present a quantitative comparison in Figure 9, showing that our method produces a structured concept space comparable to LDA but requires an order of magnitude fewer samples.

5 Conclusion

In this work, we introduce Visual Concept Inference from Sets (VICIS), a task designed to assess models' visual reasoning capabilities. Models are tasked to infer a shared concept from a small set of example images and relate it to a provided query image. Our experiments reveal that current VLMs perform surprisingly poorly on this task, highlighting a significant gap in their ability to reason over visual examples. To address this lack of capability, we present two schemes for constructing and training on VICIS tasks at scale using either synthetic data or weakly labeled real-world data. Additionally, we design an architecture that enables learning to infer these concepts from context images and extract concept-specific embeddings from query images. We demonstrate that our model successfully infers and instantiates relevant concepts and generates diverse outputs that reflect the extracted semantics. Our model outperforms both task-specific methods and general-purpose VLMs in accuracy and diversity of generated images. We validate the robustness and flexibility of our method, showing its ability to generalize to unseen concepts and modalities such as sketches.

Acknowledgments

This project has been supported by Apple, the Horizon Europe project ELLIOT (GA No. 101214398), the project "GeniusRobot" (01IS24083) funded by the Federal Ministry of Research, Technology and Space (BMFTR), the BMWE ZIM-project (No. KK5785001LO4) "conIDitional LoRA", the German Federal Ministry for Economic Affairs and Energy within the project "NXT GEN AI METHODS - Generative Methoden für Perzeption, Prädiktion und Planung", and the bidt project KLIMA-MEMES. The authors gratefully acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS/JUPITER at JSC and the HPC resources supplied by the NHR@FAU Erlangen.

Further, we would like to thank Owen Vincent for continuous technical support.

References

1. fal.ai. https://fal.ai/ (2025), accessed: 2025-09-25 12, 2
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 1
3. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023) 7, 2, 5
4. Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A.L., Darrell, T., Malik, J., Efros, A.A.: Sequential modeling enables scalable learning for large vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22861–22872 (2024) 3
5. Balazevic, I., Steiner, D., Parthasarathy, N., Arandjelović, R., Henaff, O.: Towards in-context scene understanding. Advances in Neural Information Processing Systems **36**, 63758–63778 (2023) 3
6. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. Advances in Neural Information Processing Systems **35**, 25005–25017 (2022) 3, 8, 10, 5
7. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025) 1, 8, 13
8. Baumann, S.A., Krause, F., Neumayr, M., Stracke, N., Sevi, M., Hu, V.T., Ommer, B.: Continuous, subject-specific attribute control in t2i models by identifying semantic directions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13231–13241 (June 2025) 5
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 3
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 6, 2
11. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. Advances in neural information processing systems **31** (2018) 4
12. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025) 1
13. Chollet, F.: On the measure of intelligence. arXiv preprint arXiv:1911.01547 (2019) 5
14. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=2dnO3LLiJ1 2, 6
15. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 3, 8, 10
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and

Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 10

17. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., Sui, Z.: A survey on in-context learning. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1107–1128. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.64, https://aclanthology.org/2024.emnlp-main.64/ 1, 3
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy 2
19. Fortin, A., Vernade, G., Kampf, K., Reshi, A.: Introducing gemini 2.5 flash image, our state-of-the-art image model (Aug 2025), https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, accessed: 2025-09-25 1, 8, 13, 3
20. Gandikota, R., Materzyńska, J., Zhou, T., Torralba, A., Bau, D.: Concept sliders: Lora adaptors for precise control in diffusion models. In: European Conference on Computer Vision. pp. 172–188. Springer (2024) 5
21. Gandikota, R., Wu, Z., Zhang, R., Bau, D., Shechtman, E., Kolkin, N.: Sliderspace: Decomposing the visual capabilities of diffusion models. arXiv preprint arXiv:2502.01639 (2025) 5
22. Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: Ganalyze: Toward visual definitions of cognitive image properties. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 5744–5753 (2019) 4
23. Gong, Y., Song, Y., Li, Y., Li, C., Zhang, Y.: Relationadapter: Learning and transferring visual relation with diffusion transformers. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026), https://openreview.net/forum?id=DOb47fj0cl 13
24. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Sy2fzU9gl 4
25. Huang, R., Wang, C., Yang, J., Lu, G., Yuan, Y., Han, J., Hou, L., Zhang, W., Hong, L., Zhao, H., et al.: Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. arXiv preprint arXiv:2504.01934 (2025) 3, 8, 10
26. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017) 5
27. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 4
28. Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: International conference on artificial intelligence and statistics. pp. 2207–2217. PMLR (2020) 4
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 2

30. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t 7
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 1
32. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z 7
33. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019) 4
34. Miller, G.A.: WordNet: A lexical database for English. In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994), https://aclanthology.org/H94-1111/ 8, 9
35. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? arXiv preprint arXiv:2501.09038 (2025) 5
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 6
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 6
38. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd birds-200-2011. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 5
39. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019) 10, 12
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 1, 13
41. Xu, J., Gandelsman, Y., Bar, A., Yang, J., Gao, J., Darrell, T., Wang, X.: Improv: Inpainting-based multimodal prompting for computer vision tasks. arXiv preprint arXiv:2312.01771 (2023) 3
42. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 8
43. Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: Causalvae: Structured causal disentanglement in variational autoencoder. arXiv preprint arXiv:2004.08697 (2020) 4