Personalized Vision via Visual In-Context Learning

Anonymous Author(s)
Affiliation
Address
email

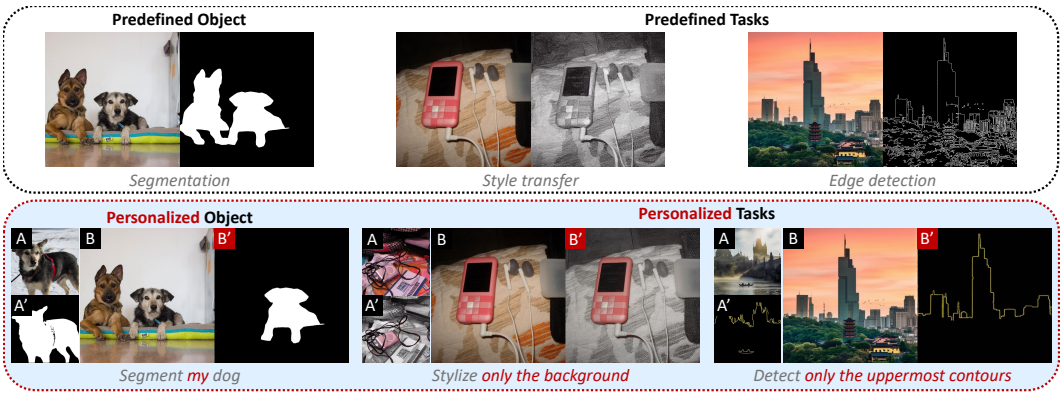

Figure 1: **Predefined vs. Personalized Vision.** Illustration of traditional vision tasks (top) and the personalized tasks enabled by our proposed **PICO** (bottom). Given a contextual example pair ($A \rightarrow A'$) defining the desired visual transformation, and a query image B, our model infers the task and generates the corresponding B' at test time.

Abstract

Modern vision models, trained on large-scale annotated datasets, excel at predefined tasks such as segmentation but struggle to adapt flexibly to personalized vision tasks—tasks defined at test-time by users with customized objects or novel objectives. Existing personalization approaches typically rely on synthesizing additional training data or fine-tuning the entire model, limiting flexibility and incurring significant computational cost. Inspired by recent advances in natural language processing, we explore a new direction: leveraging visual generative models for personalized vision via in-context learning. We introduce a structured four-panel input format, where a single annotated example specifies the personalized visual task, allowing the model to interpret and generalize the task to new inputs without further fine-tuning. To enable this one-shot capability, we construct a Visual-Relation tuning dataset tailored to personalized vision in-context learning. Extensive experiments demonstrate that our approach (i) surpasses fine-tuning and synthetic-data baselines on personalized segmentation, (ii) enables test-time definition of novel personalized tasks, and (iii) generalizes across both visual recognition and generation settings. Our work establishes a new paradigm for personalized vision, combining the adaptability of in-context learning with the visual reasoning capabilities of generative models.

1 Introduction

Modern vision models [1, 2, 3, 4, 5], trained on large-scale annotated datasets, have achieved impressive performance in both visual recognition and generation. However, these models typically

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

succeed on predefined object categories (*e.g.*, cars, people) or standard task formats (*e.g.*, object detection, semantic segmentation) where abundant labeled data exists. They often struggle to adapt flexibly to personalized vision—tasks defined by users at test-time, involving customized objects or novel task definitions. With growing demand for personalized vision systems that quickly adapt to individual needs, a critical question emerges: *How can we achieve flexible and high-performing personalized vision?*

A traditional approach to personalized vision uses generative models to synthesize additional training data tailored to specific personalized objects. For example, Personalized Representation (PRPG) [6] employs DreamBooth [7] to generate synthetic data for target concepts, then adapting general-purpose feature representations into personalized ones. While these methods [6, 8] make strides toward personalized vision by adapting to personalized objects, they remain constrained to predefined task (*e.g.*, segmentation or classification) and fail to generalize flexibly to arbitrary user-defined tasks. Besides, adapting to a new subject often requires computationally expensive fine-tuning of the entire model.

Inspired by recent breakthroughs in natural language processing (NLP), where the paradigm has shifted from task-specific fine-tuning toward in-context learning [9, 10], models can now perform novel tasks defined only at test time. Motivated by this shift, we explore a new direction: *leveraging visual generative models for personalized vision via in-context learning.* Unlike NLP tasks, which are typically well-defined and easily described with text, vision tasks often involve ambiguous perceptual inputs that are hard to specify through language alone. Furthermore, current visual generative models [4, 5], primarily pretrained on image generation, are incapable of directly reasoning about novel visual tasks at test-time.

To bridge this gap, we extend the idea of vision in-context learning (ICL) by introducing a four-panel input format. In this setting, a single annotated example (an input-output pair) is provided as a visual context, implicitly specifying the personalized task. The model, named Personalized In-context Operator (**PICO**), interprets this visual context to understand the personalized task, subsequently adapting it to new inputs to generate corresponding outputs. We construct the Visual-Relation Dataset (**VisRel**), a tuning dataset composed of diverse and structurally organized visual tasks, based on the proposed four-panel ICL setup to excite the model's ability to understand and reason about personalized vision tasks.

We conduct extensive experiments to validate the effectiveness of our proposed paradigm for personalized vision. First, we demonstrate that our method achieves superior performance compared to fine-tuning-based methods for personal subjects within conventional vision tasks. Second, we show, for the first time, that our method provides unprecedented flexibility in dynamically accommodating novel, user-defined tasks at test-time. Finally, our method achieves strong performance across diverse personalized vision scenarios, spanning both visual recognition and generation.

In summary, our key contributions are:

- We explore visual in-context learning, introducing a novel paradigm that directly leverages generative models for personalized vision, instead of relying on synthetic data generation.
- We propose an in-context fine-tuning strategy and construct a corresponding dataset, enabling pretrained image diffusion models to become effective visual in-context reasoners.
- We demonstrate promising results across a wide range of personalized vision tasks, spanning both recognition and generation, and covering varied subjects and task definitions.

2 Related Work

Personalized Vision. Existing personalized vision methods [6, 8, 11, 12, 13, 14, 15] typically adapt vision or vision-language models (VLMs) to handle user-specific concepts within predefined tasks like retrieval and segmentation. For example, PerSAM [14] segments user-indicated regions using cosine similarity on pretrained segmentation features [3], while PDM [15] leverages intermediate features from text-to-image (T2I) models [4] to localize personalized instances. PRPG [6] generates synthetic training data to enhance personalized representations for downstream tasks. However, these methods are inherently restricted to fixed task formats, lacking flexibility to accommodate arbitrary user-defined tasks at test-time. Real-world personalization often demands versatile, dynamically defined tasks. For instance, users may want to insert specific objects into images or annotate them

using custom formats. Such scenarios motivate our approach to enable personalized vision systems to rapidly adapt beyond fixed frameworks.

Visual In-Context Learning. Visual ICL, inspired by prompt-based task adaptation in NLP [9], aims to adapt vision models to downstream tasks through contextual examples. Bar *et al.* [16] first propose visual prompting by framing vision tasks as quad-grid masked image inpainting. Painter [17], a ViT-based model [18] trained through masked image modeling, shows strong ICL capabilities across various dense prediction tasks, and SegGPT [19] further enhances this ability specifically for segmentation. However, existing training-based visual ICL methods rely heavily on extensive, task-specific pretraining, limiting generalization to unseen tasks. In contrast, inference-based methods [20, 21, 22, 23, 24] attempt to interpret visual demonstrations by translating them into textual instructions. These methods do not fully use the visual instructions, resulting in inaccuracies due to the ambiguity of the textual descriptions. Additionally, they remain largely confined to semantically-driven editing tasks. Our work advances visual ICL by explicitly formulating personalized vision as visual relations within a unified space, enabling robust, flexible one-shot personalization tailored to individual needs.

Diffusion Priors. Diffusion models have emerged as the defacto paradigm for image synthesis [4, 5], demonstrating powerful generative priors beneficial for diverse vision tasks, including dense prediction [25, 26, 27], image restoration [28, 29, 30, 31], style transfer [32, 33], etc. Within data-scarce personalized vision settings, diffusion models are commonly employed to synthesize additional training, augmenting limited examples for downstream finetuning [7, 34]. However, this two-stage process [6] is computationally intensive, limiting practicality for frequent adaptation to personalized concepts. Recent work such as In-Context LoRA [35] have highlighted the intrinsic ICL capability of diffusion transformers [36]. Building upon these insights, we directly utilize diffusion priors as in-context learners, enabling flexible, immediate adaptation to arbitrary user-defined personalized visual tasks without relying on synthetic data augmentation.

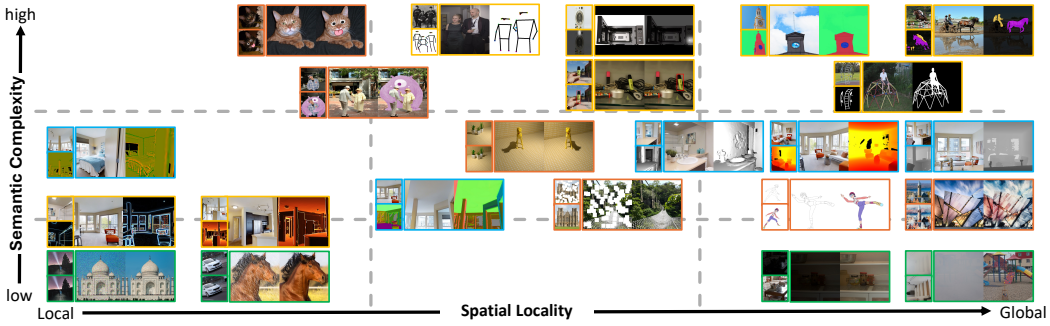

Figure 2: **Structured Visual Relation Space.** Tasks are organized by semantic complexity (low to high) and spatial locality (local to global), covering diverse task types, color-coded as: ■: Restoration/Enhancement; ■: Physical/Geometric Estimation; ■: Semantic Perception; ■: Generative Manipulation.

3 Method

Our objective is to achieve flexible visual personalization through a task-agnostic framework capable of adapting to user-defined tasks at inference without extra finetuning. We reformulate personalized vision as a visual ICL problem, where a single input-output exemplar defines the task objectives. The model infers user intent from contextual visual demonstration and applies it to new queries. Central to our approach is training on a broad visual relation space, repurposing pretrained diffusion transformers into in-context visual reasoners.

3.1 Data: A Visual Relation Space

ICL succeeds in NLP because every task (*e.g.*, translation, summarization, question answering, etc.) shares a unified language generation interface. In vision, however, different tasks have heterogeneous output format (*e.g.*, pixel arrays, masks, coordinates), limiting the potential for unified in-context generalization. We address this by unifying visual tasks as image-to-image transformations repre-

sented as RGB inputs and outputs [16, 17]. Our key insight is that a robust visual ICL model should similarly embed tasks within a unified visual relation space, enabling interpolation and composition of transformations at test time. To learn this space, we curate VisRel, a diverse collection of more than 25 visual tasks, aiming to span the space of common 2D transformations (see Figure 2). The dataset construction considers three design principles.

Task Taxonomy. We structure the visual relation space along two intuitive axes: (1) *Semantic Complexity* measures the level of semantic understanding required, spanning low-level (pixel/color adjustments), mid-level (structure/shape manipulation), and high-level (object/class reasoning) transformations. (2) *Spatial Locality* defines the spatial context dependency, ranging from local (neighboring pixels), intermediate (objects patches), to global (full-image context) operations.

Intra-task Diversity. To prevent overfitting to narrow task variants, we maximize diversity within each task. For example, inpainting includes masks of varying colors, shapes, and transparency; segmentation supports different colors, transparency mask; and restoration tasks (denoising, deblurring) incorporate different noise levels or blur kernels. By exposing the model to a rich space of transformations, we encourage learning fundamental transformation principles rather than memorizing task-specific patterns. This design is important for zero-shot generalization to novel personalized tasks defined through contextual visual demonstrations.

Minimal Text Label. The model primarily trained to infer transformation intent from visual exemplars (nputs-output pairs), without relying on explicit task identifiers. However, to resolve ambiguities between potential conflicts of interest tasks (*e.g.*, local vs. global edits; black and white depth estimation vs.colorful style transfer), we introduce minimal text prompts (*e.g.*, "edit.. vs. estimate..") as soft boundaries.

3.2 Training: PICO

Given an input-output demonstration pair $\{A, A'\}$ illustrating a visual relation $r : A \to A'$ and a query image B, our training objective is to generate an image B' that adheres to the underlying visual transformation provided by the examples. We represent tasks via a quad-grid input format: $I = \text{Grid}(\{A, A', B, X\})$, where X is a noisy placeholder. This format allows task specification without explicit labels, enabling personalized adaptation through visual exemplars. The overall training pipeline is illustrated in Figure 3. We build upon a pretrained T2I diffusion transformer (DiT) [37], finetuned using LoRA [38]. Conditions are visual exemplars $c_{vp} = \mathcal{E}(\{A, A', B\})$ encoded by a VAE encoder $\mathcal{E}(\cdot)$, and minimal textual prompts c_T. Unlike In-Context LoRA [35], which injects noise into all latents, we maintain clean latent representations for c_{vp} while only applying noise to $x_0 = \mathcal{E}(X)$ to obtain x_t. This clean conditioning ensure stable preservation of visual relation logic during the noisy denoising process, and prevent potential corruption of given conditions. The model learns to refine the noisy latent x_t by predicting a velocity field v conditioned on c_T and c_{vp}. The conditional velocity predictor $v = v_\Theta(x_t, t|c_T, c_{vp})$ is trained with conditional flow matching (CFM) loss:

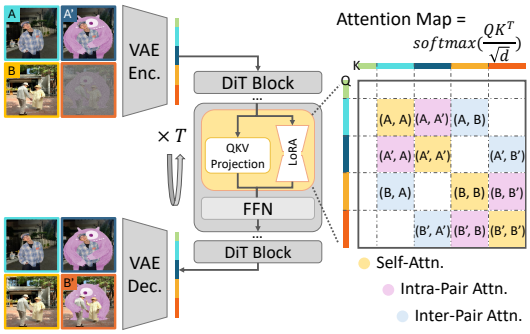

Figure 3: **Overall pipeline of PICO.**

$$\mathcal{L}_{\text{CFM}}(\Theta) = \mathbb{E}_{t,x_t}\left[||v_\Theta(x_t, t|c_T, c_{vp}) - \hat{v}(x_t, t)||^2\right], \tag{1}$$

where Θ denotes the model parameters and $\hat{v}(x_t, t)$ is the ground-truth velocity for the noisy component x_t at time t. This objective trains the model to iteratively refine the noisy placeholder X into a valid output B' that faithfully inherits the transformation demonstrated by (A, A'). This quad-grid arrangement allows the DiT's attention mechanism [5] to naturally capture intra-relationships between the examples (A, A') as well as their inter-correlations with the query image (A, B), guiding the iterative denoising of X into the desired output B'.

3.3 Inference: One-Shot Personalization

At inference time, personalized adaptation is achieved through one-shot visual prompting, mirroring the training procedure. We replace the placeholder cell with pure noise $x_T \sim \mathcal{N}(0, 1)$, and iteratively denoise it over T steps, conditioned on the context latents $c_{vp} = \mathcal{E}(\{A, A', B\})$ and optional text cue c_T. Formally, the clean latent x_0 is obtained by integrating the learned conditional velocity field v_Θ:

$$x_0 = x_T + \int_T^0 v_\Theta(x_t, t \mid c_T, c_{vp})\, dt, B' = \mathcal{D}(x_0), \tag{2}$$

where x_t follows the flow $\frac{dx_t}{dt} = v_\Theta(x_t, t \mid c_T, c_{vp})$, and $\mathcal{D}(\cdot)$ is the VAE decoder. The model seamlessly transfers the visual transformation demonstrated by (A, A') to the query B, supporting flexible, test-time personalization without fine-tuning.

4 Experiments

We validate our method through extensive experiments addressing three key questions: (1) Does visual ICL surpass traditional personalized fine-tuning on standard tasks like personalized segmentation? (2) Can the framework handle novel, user-defined tasks at inference? (3) Does it extend across recognition and generation tasks?

4.1 Implementation Details

We build PICO upon FLUX.1-dev [37], a latent rectified flow transformer model, finetuning with LoRA [38] (rank 256) on the VisRel dataset for $30,000$ steps using a single H100 GPU. Training is conducted at a resolution of 1024×1024, where each image in the quad-grid is structured at 512. We use the Prodigy optimizer [39] with safeguard warmup, bias correction enabled, and a weight decay of 0.01. The training dataset consists of 315 samples across 27 diverse tasks. Examples of task types are shown in Figure 2. The dataset is constructed from existing sources [40, 41, 42, 43, 44, 45, 46, 47, 48, 49]. Due to space constraints, full details of data construction are provided in the supplementary. Code, model and dataset will be released.

4.2 Personalized Image Segmentation

Datasets. We evaluate across four personalized segmentation benchmarks: PerSeg [14], DOGS [6], PODS [6], and PerMIS [15]. While PerSeg and DOGS mainly contain either single instances or distinct instances easily segmented using semantic cues, PODS is more challenging due to variations in viewpoints, scales, and distractors. PerMIS, sourced from video frames, further increases the difficulty by emphasizing instance-level segmentation.

Baselines. We compare PICO with three groups of state-of-the-art methods: (i) Large-scale pretrained segmentors: PerSAM [14] and SegGPT [19], both trained on extensive collections of annotation segmentation masks. (ii) Personalized representation learners: PDM [15] (diffusion features) and

Table 1: **Quantitative Comparison on personalized segmentation.** We evaluate PICO against large-scale pretrained, personalized, and generalist ICL methods. ★: best, ☆: second-best, and ◆: third-best.

Method	PerSeg [14]			DOGS [6]			PODS [6]			PerMIS [15]		
	mIOU↑	bIOU↑	F1↑	mIOU↑	bIOU↑	F1↑	mIOU↑	bIOU↑	F1↑	mIOU↑	bIOU↑	F1↑
large-scale												
PerSAM [14]	90.50◆	72.79◆	94.07☆	86.87☆	71.06★	53.18	67.45☆	56.63☆	45.60★	51.77☆	37.95☆	21.71☆
SegGPT [19]	95.77★	81.58★	99.16★	91.16★	65.93☆	85.14★	65.22◆	50.75◆	42.45	77.90★	47.10★	38.61★
personalized												
PDM [15]	29.99	10.97	2.79	21.03	8.95	0.11	26.39	10.98	1.12	23.62	9.10	1.27
PDM+PerSAM	50.09	60.08	33.37	64.36	53.82	41.85	35.56	45.34	22.33	28.93	25.25	11.72
PRPG [6]	-	-	-	81.52◆	37.34	68.74☆	60.68	34.56	40.41◆	-	-	-
generalist												
VP [16]	24.83	18.11	0.03	38.50	14.34	4.86	17.48	12.10	0.14	8.87	4.16	0.10
Painter [17]	56.56	51.58	29.76	72.07	49.75	56.88◆	26.93	25.44	6.87	19.53	15.59	4.20
PICO (ours)	90.97☆	76.13☆	62.82◆	71.02	54.71◆	49.84	68.72★	60.26★	44.88☆	49.52◆	33.63◆	14.90◆

PRPG (personalized features via synthetic-data finetuning), followed by using attention maps for instance localization. (iii) Generalist ICL models: Visual Prompting (VP) [16] and Painter [17].

Evaluation Metrics. Following [15, 6], we report mIOU, bIOU and F1@0.50 scores over all benchmarks. All the baseline methods we use its official code base and default settings.

Results. Table 1 shows that PICO outperforms generalist ICL models (VP, Painter) and personalized representation methods (PDM, PRPG), particularly on the more challenging PODS and PerMIS datasets. While PRPG achieves competitive results on DOGS, its reliance on per-instance synthetic data generation makes it computationally costly and difficult to scale (see Table 2). Thus, we omit its results on PerSeg and PerMIS, where over 500 unique instances are each accompanied by a single reference image. In contrast, PICO's generative in-context learning paradigm enables instant adaptation to new instances at inference without retraining, offering strong practical advantages. Notably, compared to large-scale pretrained segmentors, PICO achieves comparable performance while using up to four orders of magnitude fewer labeled data (see Table 3), highlighting its superior data efficiency enabled by generative priors. Interestingly, whereas traditional segmentation methods rely heavily on deterministic visual features, our results reveal that generative priors can act as strong inductive biases, warranting further exploration for structured vision tasks. Qualitative results are shown in Figure 4(a).

Free-Form Inputs and Task Flexibility. Beyond dense masks, PICO supports sparse annotations (*e.g.*, bounding boxes, circles), enabling intuitive, coarse-grained personalization tasks such as detection, shown in Figure 4(b). The method also extends seamlessly from single-instance segmentation to part-level parsing, respecting arbitrary color coding and transparency levels specified by users at test time. As shown in Figure 4(c), our model successfully follows contextual appearance cues and consistently segments out specific semantically identical components. semantic components consistently. Although never trained on facial data, it performs well on out-of-domain tasks (*e.g.*, face parsing), demonstrating its robustness and flexibility.

Table 2: **Comparison of personalized segmentation.**

Method	Use of Generative Prior	Features	Seg. Method	Test-time New Instance?
PDM [15]	Feature extractor	SDXL-turbo [50]	Attention map	✓
PRPG [6]	Synthetic data generator	Personalized DINOv2 [2]	Attention map	✗(retraining required)
PICO (ours)	In-context learner	–	Direct output	✓

Table 3: **Comparison of large-scale pretrained methods.** PICO uses minimal supervision and adopts a generative diffusion backbone.

Method	Seg. Data / Total Data	Training	Loss
PerSAM [14]	11M / 11M	Finetuned from MAE-pretrained ViT-H [51] (encoder)	Cross-entropy
SegGPT [19]	254K / 254K	Finetuned from Painter [17]	Smooth L1
Painter [17]	138K / 192K	Finetuned from MAE-pretrained ViT-Large [51]	Smooth L1
PICO (ours)	**40 / 315**	Finetuned from FLUX (DiT-based) [37]	Flow-matching

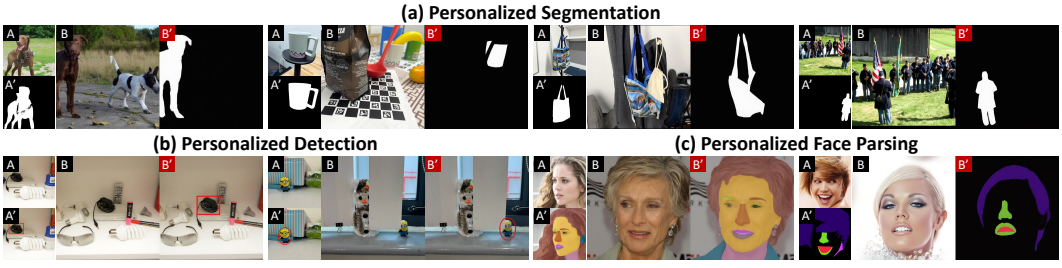

Figure 4: **Results of free-form personalized segmentation adaptation.** PICO supports a range of personalized settings: (a) Personalized object segmentation; (b) Personalized detection using sparse annotations; (c) Arbitrary part-level face parsing with in-context color and transparency cues.

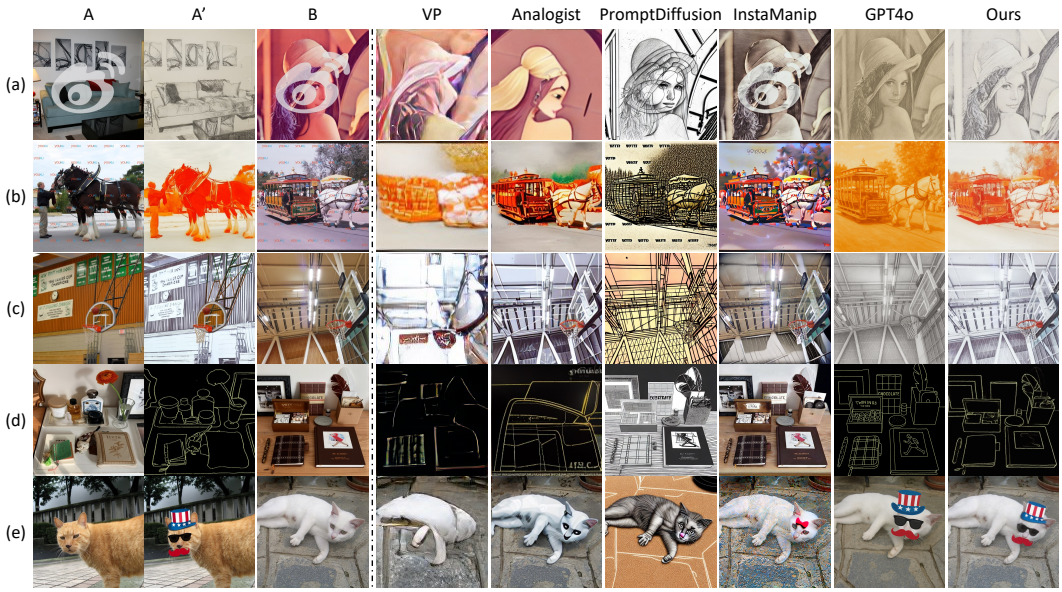

Figure 5: **Qualitative comparisons on test-time personalized tasks.** We compare our method with five representative baselines. Each task is defined by a visual example pair $(A \rightarrow A')$, including (a)(b) watermark removal + style transfer; (c) background-only stylization; (d) contour-only edge detection; and (e) add the same stickers.

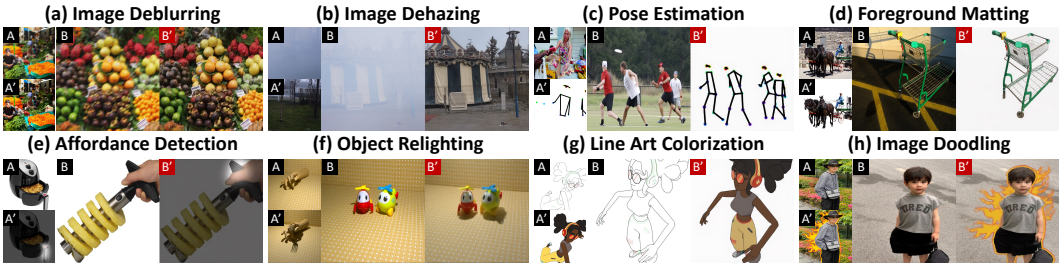

Figure 6: **Results of supported tasks.** PICO supports a diverse range of tasks, including restoration (a–b), perception (c–e), and generation (f–h).

4.3 Personalized Test-time Task Generalization

Task Definition. We evaluate test-time personalization on user-defined visual tasks that differ from conventional CV setups. Specifically, we focus on: (i) **Composite tasks** requiring multi-step operations (*e.g.*, watermark removal followed by stylization). (ii) **Spatially constrained tasks**, traditionally performed globally but here applied locally or selectively (*e.g.*, contour-only edge detection, background-only stylization). (iii) **Semantic-conditional tasks** demanding context-aware edits (*e.g.*, adding stickers to semantically relevant image regions).

Baselines. Given these novel tasks, we compare PICO with representative state-of-the-art methods supporting visual instructions, including: (i) Inference-based method: VP [16], Analogist [23]; (ii) Training-based method: PromptDiffusion [52], InstaManip [24]; (iii) *Commercial multimodal models*: GPT-4o [53]. Textual instructions for these methods follow Analogist's GPT-4o-based reasoning procedure.

Results. Qualitative comparisons in Figure 5 show that PICO effectively handles diverse test-time defined novel tasks, clearly surpassing baseline methods. Training-based methods (PromptDiffusion, InstaManip) primarily target semantic-driven editing and thus struggle to match demonstrated appearances, especially in non-RGB outputs (*e.g.*, edge maps as shown in Figure 5(d)). They also fail at composite tasks, notably failing to remove watermarks before stylization (Figure 5(a,b)). Inference-based methods (VP, Analogist) can roughly mimic target transformations, but their outputs suffer

Method	Pers. Seg↑	Normal↓	Z-depth↓
VTM (10-Shot) [54]	-	11.4391	**0.0316**
Ours w/o Text	66.88	12.7105	0.0432
(± std)	-	(± 3.0854)	(± 0.0228)
Ours w Text	**68.72**	**10.5306**	0.0377
(± std)	-	(± 2.2856)	(± 0.0199)

Method	2DEdge↓	2DKeypoint↓	Reshading↓
VTM (10-Shot) [54]	0.0791	0.0639	**0.1089**
Ours w/o Text	0.0538	0.0609	0.1518
(± std)	(± 0.0170)	(± 0.0128)	(± 0.0553)
Ours w Text	**0.0515**	**0.0497**	0.1364
(± std)	(± 0.0172)	(± 0.0137)	(± 0.0522)

Table 4: Quantitative ablation studies. For reference, we include 10-shot results from VTM [54].

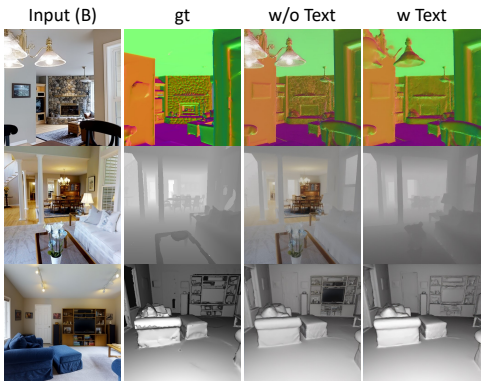

Figure 7: Qualitative comparisons on with and without text prompts.

from poor fidelity and noticeable visual artifacts or misalignment. GPT-4o [53] shows promising in-context understanding ability, capturing the high-level intent conveyed by examples. However, two major limitations are observed. (1) Spatial misalignment: While the semantic content is preserved, the pixel-wise layout is distorted. This poses challenges for tasks requiring spatial precision, such as contour detection (Figure 5(e)), and fails in scenarios involving local edits like adding the same hat (Figure 5(c)). (2) Over-reliance on abstract concepts: Rather than faithfully imitating the visual exemplars, GPT-4o appears to rely on high-level semantic embeddings. In stylization tasks (Figure 5(a)(b)(c)), the output fails to match the reference style, but instead defaults to generic "sketch" or "orange-tone" effects. In contrast, PICO produces outputs consistently aligned in spatial and semantic detail with provided examples, highlighting its robust visual reasoning capability.

Supported Tasks. In addition to personalized user-defined tasks, PICO also supports various standard visual tasks spanning restoration, perception, and generation, as illustrated in Figure 6. Although trained on these standard tasks, the model generalizes remarkably well to novel instances from as few as 10 examples per task. Notably, for the task of object relighting, *i.e.*, transforming an object from one lighting condition to another, PICO is able to predict physically plausible shadows aligned with previously unseen query objects (Figure 6(f)). This indicates an implicit understanding of lighting and object interactions, highlighting its strong generalization capability to novel physical transformation tasks.

4.4 Ablation Studies

Effects of Text Prompts. We first quantify the importance of minimal textual prompts in resolving ambiguities among multiple visual tasks. Specifically, we evaluate our model on personalized segmentation (PODS) as well as five dense prediction tasks from Taskonomy [40] (surface normal, Z-buffer depth, texture edge, 2D keypoints, and reshading). We prepare $1,000$ quad-grid formatted test examples per task from the "Muleshoe" building [40]. Evaluation metrics follow [54]: mean error (mErr) for surface normal, and RMSE for other tasks. RGB predictions are converted to respective raw outputs for metric computation.

The quantitative results in Table 4 show obvious performance improvements with text prompts, alongside lower variance, indicating that minimal text cues effectively reduce task ambiguity compared to visual prompts alone. Figure 7 illustrates typical failures without text cues, where the model confuses distinct output spaces (*e.g.*, outputting RGB-like results instead of proper surface normal maps). With text prompts, the model clearly separates these tasks, highlighting the necessity of textual guidance as soft task boundaries. For reference, we include VTM [54], a state-of-the-art 10-shot fine-tuning method for dense prediction. Remarkably, our generative in-context learner surpasses this specialized approach on tasks such as surface normal estimation and texture edge detection despite substantially lower supervision, highlighting strong generalization and data efficiency enabled by generative priors. Additional ablation studies are provided in the supplementary material.

Task vs. Data Scaling. We systematically investigate how task diversity and data quantity affect model generalization. Keeping LoRA rank (r=128) and training steps (10k) fixed, we evaluate

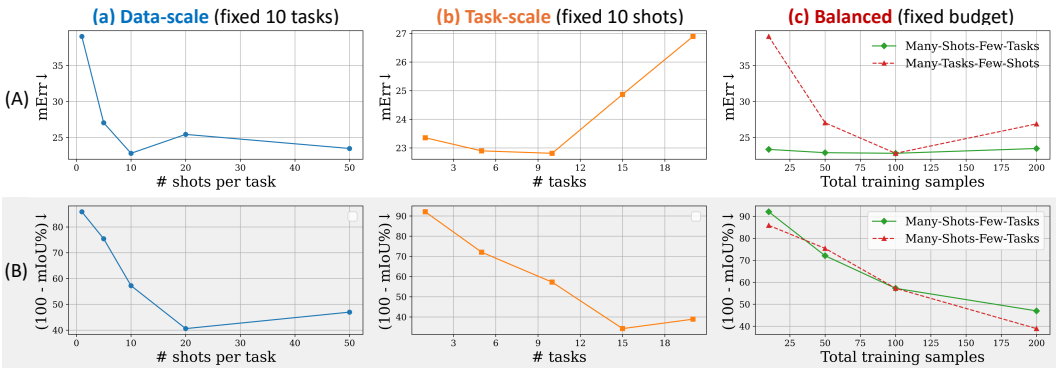

Figure 8: Quantitative comparisons across three scaling strategies on **seen** (A: surface normal estimation) and **unseen** (B: personalized segmentation) tasks. We report mean error (mErr) for surface normal and use $(100 - \text{mIoU}\%)$ for segmentation to maintain consistent interpretation (lower is better↓). Notably, in the fixed-budget setting (B-c), scaling task diversity improves generalization in unseen tasks, supporting our visual relation space hypothesis.

three scenarios: (i) **Data-scale sweep:** Fixing 10 dense prediction tasks, vary shots per task: ($K \in 1, 5, 10, 20, 50$). (ii) **Task-scale sweep:** Fixing 10 shots per task, vary number of tasks ($N \in 1, 5, 10, 15, 20$). (iii) **Balanced sweep:** Fixing total training images constant ($10, 50, 100, 200$), compare many-tasks–few-shots ($N > K$) against few-tasks–many-shots ($N < K$) regimes. We evaluate on both in-domain tasks seen during training (*e.g.*, surface normal estimation) and out-of-domain tasks not seen during training (*e.g.*, personalized segmentation).

Quantitative results are shown in Figure 8. For in-domain tasks, more data volume consistently improves performance (Figure.8A-a), while adding tasks hurt (Figure.8A-b), indicating limited capacity for memorizing multiple tasks. Under fixed budgets, concentrating data on fewer tasks is best. (Figure.8A-c). For out-of-domain generalization, performance improves with more data per task only up to 20 shots, after which it declines due to over-specialization (Figure.8B-a). Greater task diversity consistently boosts generalization (Figure.8B-b). Under fixed budgets, the many-tasks–few-shots strategy increasingly outperforms fewer-tasks–many-shots as task count grows (Figure.8B-c). These findings support our *visual-relation–space* hypothesis: increased data enhances memorization of seen tasks, while greater task diversity is crucial for robust generalization to unseen, user-defined visual tasks.

5 Conclusion

In this paper, we introduced a novel approach for personalized vision by reformulating it as a visual in-context learning (ICL) problem. Unlike existing methods that rely heavily on task-specific fine-tuning or synthetic data augmentation, we proposed learning a unified visual relation space, enabling pretrained diffusion transformers to reason about user-defined visual tasks given a single visual demonstration. Our method, termed **PICO**, demonstrates superior flexibility and effectiveness across diverse personalized vision scenarios, including complex compositional tasks. Extensive experiments validate its strong capacity to adapt robustly and efficiently to novel, test-time personalized tasks, highlighting its practical value for real-world applications and unlocking new potential for generative image models as versatile visual in-context reasoners.

Limitation and Future Work. Although PICO shows strong generalization within the visual-relation space seen during training, it is less reliable on entirely novel task types outside that space. This aligns with human learning, *i.e.*, people also extrapolate best within familiar domains, but broadening the method to truly novel tasks remains an open challenge. Additionally, the quad-grid input format, while effective, inherently limits the number of contextual examples and their complexity. Future research could explore richer context formats, or long-context vision sequential model [55] capable of supporting an arbitrary number of demonstration examples or task images, such as video sequences, enabling more comprehensive task specifications and sophisticated visual reasoning.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[6] Shobhita Sundaram, Julia Chae, Yonglong Tian, Sara Beery, and Phillip Isola. Personalized representation from personalized generation. In *ICLR*, 2025.

[7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

[8] Yunhua Zhang, Hazel Doughty, and Cees GM Snoek. Low-resource vision challenges for foundation models. In *CVPR*, 2024.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[11] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. MyVLM: Personalizing vlms for user-specific queries. In *ECCV*, 2024.

[12] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *ECCV*, 2022.

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

[14] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In *ICLR*, 2024.

[15] Dvir Samuel, Rami Ben-Ari, Matan Levy, Nir Darshan, and Gal Chechik. Where's Waldo: diffusion features for personalized segmentation and retrieval. *NeurIPS*, 2024.

[16] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *NeurIPS*, 2022.

[17] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: transformers for image recognition at scale. *ICLR*, 2021.

[19] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: segmenting everything in context. In *ICCV*, 2023.

[20] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. *NeurIPS*, 2023.

[21] Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. ImageBrush: learning visual in-context instructions for exemplar-based image manipulation. *NeurIPS*, 2023.

[22] Ruoyu Zhao, Qingnan Fan, Fei Kou, Shuai Qin, Hong Gu, Wei Wu, Pengcheng Xu, Mingrui Zhu, Nannan Wang, and Xinbo Gao. InstructBrush: learning attention-based instruction optimization for image editing. *arXiv preprint arXiv:2403.18660*, 2024.

[23] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *TOG*, 2024.

[24] Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for few-shot image manipulation. In *CVPR*, 2025.

[25] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *ICLR*, 2025.

[26] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. GeoWizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024.

[27] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

[28] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. DiffIR: Efficient diffusion model for image restoration. In *ICCV*, 2023.

[29] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. SinSR: diffusion-based image super-resolution in a single step. In *CVPR*, 2024.

[30] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE TPAMI*, 2025.

[31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

[32] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, 2024.

[33] Yuxin Jiang, Liming Jiang, Shuai Yang, Jia-Wei Liu, Ivor Tsang, and Mike Zheng Shou. Balanced image stylization with style matching score. *arXiv preprint arXiv:2503.07601*, 2025.

[34] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

[35] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

[36] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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

[38] 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. *ICLR*, 2022.

[39] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. In *ICML*, 2024.

[40] Amir R Zamir, Alexander Sax, , William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018.

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[42] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017.

[43] Codruta O. Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. In *ICIP*, 2019.

[44] Wenhan Yang Jiaying Liu Chen Wei, Wenjing Wang. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018.

[45] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, 2020.

[46] Shijie Huang, Yiren Song, Yuxuan Zhang, Hailong Guo, Xueyin Wang, Mike Zheng Shou, and Jiaming Liu. PhotoDoodle: Learning artistic image editing from few-shot pairwise data. *arXiv preprint arXiv:2502.14397*, 2025.

[47] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018.

[48] Yuekun Dai, Shangchen Zhou, Qinyue Li, Chongyi Li, and Chen Change Loy. Learning inclusion matching for animation paint bucket colorization. *CVPR*, 2024.

[49] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. ReLight My NeRF: A dataset for novel view synthesis and relighting of real world objects. In *CVPR*, 2023.

[50] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024.

[51] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[52] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *NeurIPS*, 2023.

[53] OpenAI. Hello GPT-4o. `https://cdn.openai.com/gpt-4o-system-card.pdf`, 2024.

[54] Donggyun Kim, Jinwoo Kim, Seongwoong Cho, Chong Luo, and Seunghoon Hong. Universal few-shot learning of dense prediction tasks with visual token matching. In *ICLR*, 2023.

[55] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: These claims are supported by the method and experiments (see Section 3 and Section 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper is empirical and does not include theoretical results or formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4 describes the necessary implementation details to ensure the reproducibility of our experiments. Code will be released upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Not currently. We use public datasets, so data used is available. We are working on a formal approval to publicly release the code, upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The implementation details are described in Section 4. Additional configurations are in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report mean and standard deviation in Table 4. Discussion of initial random seed can be found in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 specifies the implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work involves training vision models on public datasets and does not raise foreseeable ethical concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader of impact is discussed in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credited previous works and codes in Section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our training dataset constructed from existing publicly available data, so the data used is available. We plan to open-source them along with the official code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not include research with crowdsourcing and human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only use LLM for proof-reading.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.