UNLOCKING THE BOUNDARIES OF CROSS-TASK VISUAL IN-CONTEXT LEARNING VIA IMPLICIT TEXT-DRIVEN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

In large language models (LLM), in-context learning (ICL) refers to performing new tasks by conditioning on small demonstrations provided in the input context, without any parameter updates. Recent advances in visual domain, i.e. visual in-context learning (VICL), demonstrate promising capabilities for solving downstream tasks by unified vision-language models (VLMs). However, the boundaries of cross-task transfer in VICL remain largely unexplored, particularly for the heterogeneity across low-level vision tasks. This naturally raises the question: *When the visual prompt and the target images originate from different visual tasks, can VLMs still enable VICL?* In the paper, we propose a fully collaborative pipeline, i.e. T2T-VICL, for VLMs to investigate the potential of cross-task VICL. Fundamentally, we design a mechanism to generate and select text prompts that best implicitly describe the differences between two distinct low-level vision tasks, and construct the first cross-task VICL dataset. Building upon this, we present a training strategy from a large VLM to a small vision-language model (sVLM), together with a deployment framework from the sVLM back to the large VLM. Furthermore, we propose a novel inference framework that combines perceptual score-based reasoning with standard evaluation metrics to perform cross-task VICL. Our approach achieves stable results spanning multiple low-level cross-task pairs. During inference, T2T-VICL demonstrates promising performance without requiring any image-based training or model fine-tuning. Our findings highlight the feasibility of enabling cross-task VICL within VLMs, underscoring the utility as a supplementary generalizable paradigm for low-cost vision-language reasoning.

1 INTRODUCTION

In-context learning (ICL) enables models to solve large reasoning tasks by leveraging a few input–output demonstrations without parameter updates (Min et al., 2022; Dong et al., 2022). Recent advances show that large language models (LLMs) exhibit remarkable ICL capabilities across a wide range of natural language processing (NLP) tasks (Highmore, 2024; Wang et al., 2023b; Sia et al., 2024; Li et al., 2025), achieving strong performance in similar linguistic contextual examples. The capabilities of processing visual in-context learning (VICL) have been investigated in multiple vision generalist models, e.g. MAE-VQGAN (Bar et al., 2022), Painter (Wang et al., 2023a), Prompt Diffusion (Wang et al., 2023c), and X-Prompt (Sun et al., 2024b), where the concept of in-context learning extends from language to image. As the generative abilities of vision–language models (VLMs) continue to advance, VICL has been shown to facilitate rapid adaptation to reasoning tasks from small visual demonstrations (Zhou et al., 2024b; Ma et al., 2024), with additional benefits for multimodal integration and transferability.

Intuitively, many visual problems share underlying relationships rather than existing in isolation (Zamir et al., 2018; Pal & Balasubramanian, 2019; Achille et al., 2019). While fully supervised approaches typically learn each task independently, this siloed paradigm is inefficient and demands large amounts of labeled data (LeCun et al., 2015; Chum et al., 2019). Task transfer learning addresses this challenge by exploiting knowledge from source tasks to accelerate and improve performance on new targets (Bao et al., 2019). For example, the Taskonomy (Zamir et al., 2018) framework demonstrates that by computationally modeling transfer dependencies across diverse tasks,

and the possibility to derive a taxonomic map that captures both direct and higher-order relationships has been proven. Such relational phenomena are hypothesized to exist in language domains as well. Viewed through the lens of language, many image-processing–oriented low-level tasks own the same prefix or suffix (Trigka & Dritsas, 2025; Shu et al., 2024), e.g. “deraining”, “denoising”, “dehazing”, “deblurring”, “demoiring” . Textual descriptions of visual differences across such tasks reveal areas of overlap as well as points of divergence. In addition, textual descriptions of generation–oriented low-level tasks such as “colorization”, “light enhancement”, “harmonization”, and “style transfer” (Bar et al., 2022; Ke et al., 2023), primarily hinge on linguistic variations in color, illumination, and contrast. This suggests that certain latent relationships may govern the transformations between them. Meantime, several studies (Brooks et al., 2023; Potlapalli et al., 2023; Conde et al., 2024) have demonstrated the regulatory role of language in guiding/driving visual expressions.

Building on the above observations, we naturally raise the key question: *If the visual prompt and the images to be learned come from two different tasks, can VLMs enable VICL across different tasks?*

In this paper, we develop T2T-VICL, a collaborative pipeline based on multiple VLMs that enable VICL across multiple low-level cross-task pairs. We construct a novel approach that integrates implicit descriptive textural capture, gradient-based token attribution enhancement (Sundararajan et al., 2017; Nguyen et al., 2025), a visual instruction-guided metric (Ku et al., 2023), and image quality assessment (IQA) metrics (Zhang et al., 2012), allowing VLMs to perform cross-task in-context learning without the need for additional training and be evaluated. Our contributions can be summarized in three aspects: (1) We propose the first comprehensive multi-VLM pipeline that automatically generates implicit text prompts to distinguish two low-level vision tasks, which can be used to explore the **boundaries of cross-task VICL**. This work also introduces the **first text–image dataset** with cross-task implicit descriptions, establishing a benchmark expected to exert lasting influence within the field. (2) We design a **VLM→sVLM** and **sVLM→VLM** framework that gets knowledge from a large-scale model into a lightweight model and leverages the compact model to get the text prompt. (3) By coupling score-based reasoning, we establish an automatic inference framework that supports cross-task VICL in VLMs while **eliminating the need for further trining**.

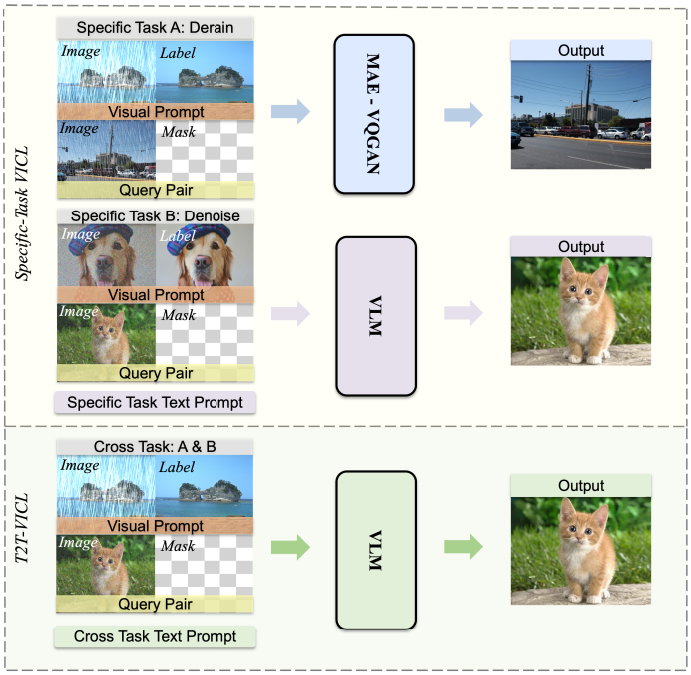

Figure 1: Illustration of Cross-Task Visual In-Context Learning

2 Related Work

2.1 Vision Generalist Model

Vision generalist models seek to replace task-specific vision networks with general-purpose frameworks capable of handling diverse visual tasks under a single paradigm. Early approaches adopted a sequence-to-sequence formulation that casts visual tasks as sequence prediction problems in NLP. Pix2Seq (Chen et al., 2022) pioneered the sequence-to-sequence view, later extended by Unified-IO (Lu et al., 2022) to unify diverse inputs and outputs, while OFA (Wang et al., 2022) advanced it toward multimodal instruction-based learning bridging vision and language. UViM (Zhang et al., 2022) introduced guiding codes as a task-agnostic interface for vision tasks, and Uni-Perceivers (Zhu et al., 2022) extended this idea to multimodal settings. Subsequent works leverage stronger backbones such as transformers and diffusion models, exemplified by Florence-2 (Xiao et al., 2023) for prompt-driven recognition and generation, and InstructCV (Gan et al., 2023), InstructDiffusion (Geng et al., 2023) for conditional denoising under a unified generative framework.

Most notably, VLMs have emerged as unified frameworks that couple visual encoders with large language models under instruction-driven settings. Emu (Sun et al., 2024a; Wang et al., 2024) series integrate diffusion models with LLM-predicted embeddings, functioning as a generalist model capable of handling diverse vision tasks. Chameleon (Chameleon et al., 2024) team follows an early-fusion strategy, integrating text and image tokens from the outset to support interleaved understanding and generation. In contrast, Transfusion (Transfusion et al., 2024) combines next-token prediction for language with diffusion for images, and Show-o (Show-o et al., 2024) extends this hybrid approach with a one-transformer design that merges auto-regression and discrete diffusion. Collectively, these efforts highlight the trend of VLMs shaping the evolution of vision generalist models.

2.2 Visual In-Context Learning

Visual in-context learning (VICL) refers to adapting vision models to downstream tasks through contextual examples rather than explicit fine-tuning, and it can be broadly divided into visual prompting and prompt-driven conditioning. Implicit prompting methods, represented by MAE-VQGAN (Bar et al., 2022) and Painter (Wang et al., 2023a), rely on masked prediction or image completion as objectives, enabling models to generalize under diverse vision tasks. In contrast, explicit prompt-driven approaches leverage explicit conditioning to adapt diverse vision problems. For instance, Prompt Diffusion (Wang et al., 2023c) exploits diffusion-based generation under prompt guidance. While PromptGIP (Liu et al., 2023) adopts QA-style prompt structures. Additional works such as CoOp (Zhou et al., 2022) and VPT (Jia et al., 2022) further extend VICL by introducing learnable prompts for vision transformers. More recently, X-Prompt (Sun et al., 2024b) targets general image generation, compressing contextual signals and unifying diverse vision tasks within an autoregressive framework. Most existing VICL methods focus on performing the visual prompt and query image in the same tasks, our T2T-VICL extends this paradigm to a cross-task setting, where the visual prompt and query image are from different tasks, digging the potential knowledge adaptation across diverse vision problems.

2.3 Task Transfer

Task generalization refers to leveraging knowledge from previously solved tasks to facilitate learning on unseen ones. Early milestones such as Taskonomy (Zamir et al., 2018) provided the first large-scale analysis of transferability across 26 vision tasks, introducing the task affinity matrix as a foundation for cross-task generalization. Following this line, TMT (Pal & Balasubramanian, 2019) extended Taskonomy by applying matrix factorization to refine the estimation of inter-task transferability. Similarly, Task2Vec (Achille et al., 2019) proposed embedding tasks into a vector space using Fisher information, making task similarity measurable and interpretable. Beyond these foundations, (Standley et al., 2020) systematically investigated which tasks should be learned together in multitask settings, highlighting task affinity as a guiding principle . (Dwivedi & Roig, 2019) introduced cross-task consistency to enforce relational constraints. Given the growing visual generation ability of VLMs, the progress in this field motivates our exploration of cross-task generalization in

VLMs and unlocks the boundary of how VLMs can transfer from one vision task to the other vision task.

2.4 Language-Driven Image Restoration and Editing

Language-driven vision restoration has progressed from demonstrating the feasibility of instruction-guided editing, where natural language serves as an efficient component to drive fine-grained manipulation of low-level image features. InstructPix2Pix (Brooks et al., 2023)introduced an instruction-following framework for image editing, enabling localized and semantically consistent modifications based on natural language prompts. InstructIR (Conde et al., 2024) extended this idea to image restoration, demonstrating that human instructions can guide degradation-aware recovery of high-quality images. Similarly, PromptIR (Potlapalli et al., 2023) incorporated prompt learning to adapt restoration networks dynamically under different degradation types.

Beyond, many studies have utilized richer semantic information to guide generation. SPIRE (Qi et al., 2024) leverages semantic and quantitative prompts to jointly control content preservation and restoration strength, while PromptFix (Yu et al., 2024) builds a large-scale instruction-following dataset and proposes high-frequency guidance sampling to preserve local details under natural language control. Perceive-IR (Zhang et al., 2025) focuses on degradation perception, coupling prompt learning with a quality perceiver to generalize across unseen degradations. FPro (Zhou et al., 2024a) proposes frequency-aware prompting to guide restoration in different spectral bands. These works demonstrate that natural language has transitioned from a descriptive signal into a direct driver of low-level vision tasks, highlighting its potential for finer-grained controllability and fidelity.

3 Method

3.1 T2T-VICL: Overview of collaborative VLM pipeline

To explore the boundaries of cross-Task VICL via VLMs, we propose T2T-VICL, which is a collaborative pipeline that leverages the complementary strengths of large and small language VLMs to enhance vision–language reasoning. First, a large pre-trained VLM generates implicit textual descriptions for two distinct tasks, while the linguistic reliability of these descriptions is quantitatively evaluated. To enable efficient deployment, we perform fine-tuning of a sVLM to transfer the reasoning capabilities of the large model while reducing computational overhead. During inference, the sVLM produces text prompts to the final large VLM, thereby guiding the downstream reasoning process. Finally, candidate results are ranked by another VLM with a visual instruction-guided metric and IQA metrics, and the optimal outcome is selected from the top-k candidates.

3.2 Automated Implicit Task Relationship Generation

We begin by automatically generating rich textual descriptions that implicitly capture the differences between two low-level vision tasks. We consider 12 diverse low-level tasks spanning classic degradation/removal problems (e.g. deraining, dehazing, denoising, deblurring, demoiring, shadow removal, reflection removal) as well as generation/enhancement tasks (e.g. colorization, low-light enhancement, harmonization, style transfer). For any arbitrary pair of tasks A and B, our goal is to obtain an implicit text prompt that depicts the difference without explicitly naming either task. To accomplish this, we leverage a state-of-the-art large vision–language model, Qwen2.5-VL-32B-Instruct, which is fed by two image pairs and the ground truth. We design the text prompt P_t to provide structured comparisons of two tasks: (1) the target goal – what the task is trying to achieve (e.g. "remove rain streaks from the scene" or "sharpen blurred details"); (2) the input degradation or attribute – the type of distortion/artifact or initial condition present in the input (rain, haze, blur, noise, poor lighting, etc); and (3) the visual changes from input to output – the perceptible improvements or modifications after applying the task. Crucially, the prompt instructs the VLM to compare these aspects for task A versus task B without ever revealing the task names, ensuring the description is purely implicit. For instance, the model might reveal that denoising and deblurring both remove high-frequency imperfections (random noise vs. motion blur) but differ in the patterns they target. This mechanism forces the VLM to articulate the subtle differences in a narrative manner.

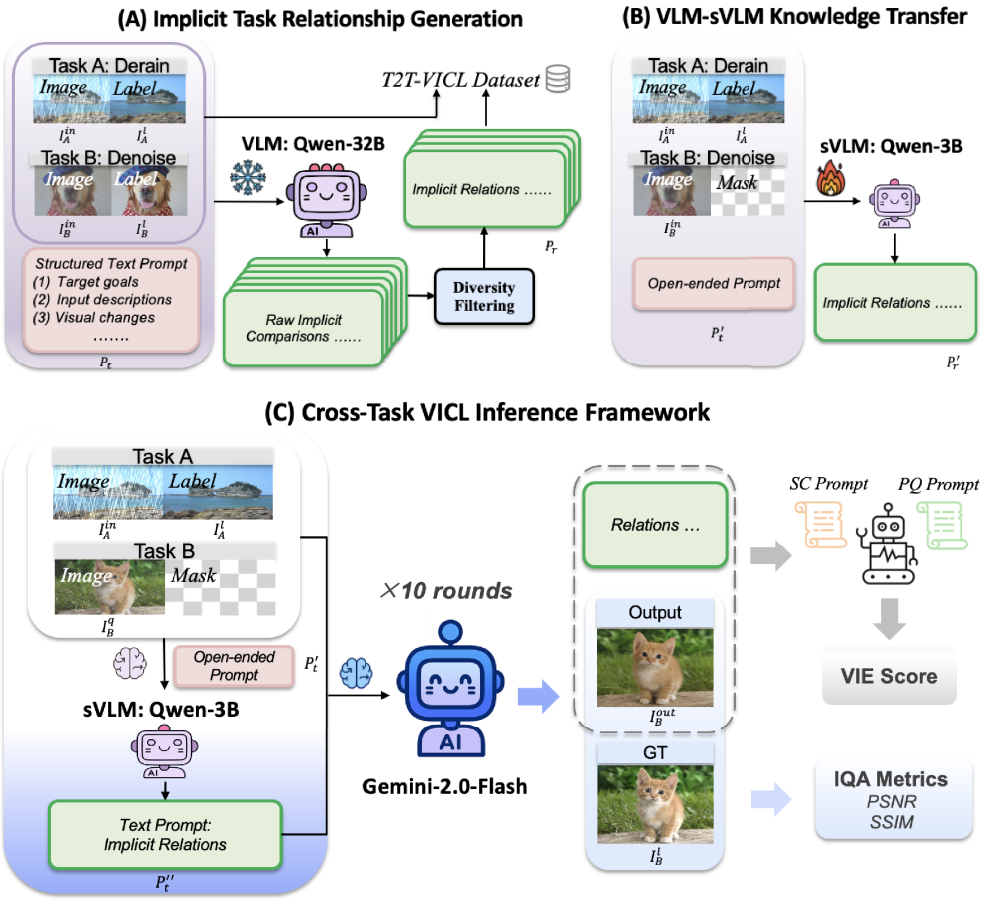

Figure 2: Overview of the proposed pipeline and its workflow

Subsequently, the above procedure is implemented for all the included low-level tasks to build a high-quality benchmark dataset. For each pair of tasks A and B, we sampled a large number of combinations from public datasets and queried the VLM with each sample to get a text output: (i) an input image I_A^{in} and its label I_A^l from task A, and (ii) an input image I_B^{in} with the label I_B^l from different task B. Thus we obtained the implicit textural relations:

$$P_r = M_l(I_A^{in}, I_A^l, I_B^{in}, I_B^l, P_t), \tag{1}$$

The model is trained to generate the corresponding comparison text that the teacher (32B model) had produced for that pairing. However, the VLM can sometimes produce repetitive or overly generic statements, especially if many samples share similar traits. We therefore conducted a diversity filtering method by using semantic sentence embeddings. We encoded each candidate description with a SentenceTransformer (all-MiniLM-L6-v2) to obtain a dense vector representation. This allowed us to quantitatively measure similarities among the descriptions. We then performed clustering in embedding space to detect groups of near-duplicates or redundant phrasing. In this manner, we filtered out repetitive outputs and retained only one representative with the most distinct descriptions. Ultimately, we kept 2,000 diverse descriptions per task pair. To our knowledge, this is the first text-and-image dataset that implicitly captures cross-task relationships, providing the foundation for our next steps.

3.3 VLM–sVLM Knowledge Transfer

3.3.1 Large-to-Small Transfer Framework

Knowledge transfer and teacher-student model construction. Having obtained the implicit task comparisons from the large VLM, we next transfer this knowledge into a Qwen2.5-VL-3B-Instruct

model as a student. By using the generated text prompts in section 3.2 as targets, we fine-tuned the student model with three input images I_A^{in}, I_A^l, and I_B^{in}. Differently, we omit the label of task B I_B^l here. By using an open-ended prompt $P_t^{'}$ (e.g. "Compare the effects observed in these images"), we force the 3B model to learn to infer the essence of task B's effect solely from the characteristics of input and output the implicit textural relations $P_r^{'}$.

$$P_r' = M_s(I_A^{in}, I_A^l, I_B^{in}, P_t'), \tag{2}$$

We format each training instance in the official Qwen-VL conversational style, which includes a system instruction and the images embedded with markdown tags. The student is then trained to produce the teacher's full comparison text as the completion. The training objective is to minimize the cross-entropy loss between the student's generated text and the teacher's reference description. Through many epochs of exposure to different task pairs and images, the sVLM model gradually learns the teacher's "reasoning habit" – specifically, how to discuss goals, degradations, and visual changes for two tasks – and becomes capable of producing these comparisons on its own. In effect, the large model's implicit knowledge of inter-task relationships is transferred into the small model at the language level.

Model choice and training efficiency. We selected Qwen2.5-VL-3B as the student not only for its manageable size, but also because it shares the same architecture and multimodal interfaces as the teacher model. After training, the fine-tuned 3B model is indeed able to take two image pairs from tasks A and B and directly output a coherent comparison of the tasks, much like the 32B model did. In other words, the student has learned to be a "narrative engine" that explains cross-task relationships on demand. This approach is related in spirit to recent prompt-based tuning methods in vision–language models (e.g. CoOp prompt tuning, visual instruction tuning), but with a key difference: instead of learning a static prompt vector or fixed set of words for a model, we train a full generative model to produce dynamic, content-dependent prompts. The student can flexibly generate different comparative descriptions for different image pairs, rather than a single fixed prompt. By using the large model's outputs as textual supervision, we effectively compress its high-level reasoning about tasks into a lightweight model.

3.3.2 Small-to-Large Deployment Framework

Once the student model is trained, we deploy it as a front-end module to assist the large model during inference and provide the text prompt, which is then consumed by the larger VLM for final reasoning and output image generation. This hierarchical approach reduces reliance on heavy computation during the initial stages of inference while preserving the representational capacity of large-scale VLMs. This is essentially a reverse direction of knowledge flow: the Small→Large step uses the 7B model's explanation to guide the final inference model Gemini-2.0-flash. The process employs the trained Qwen-3B model to process the input of the visual prompt (I_A^{in}, I_A^l), query image I_B^q and text prompt P_t', output a prompt representation P_t' as following:

$$P_t'' = M_s(I_i, I_l, I_q, P_t'), \tag{3}$$

where P_s encodes salient task-relevant cues. By offloading this initial abstraction to M_s, we reduce the computational overhead associated with feeding raw multimodal inputs directly into the large model.

The generated prompt P_s is then provided as an input to a larger VLM M_ℓ, which possesses stronger representational power and reasoning ability. Specifically,

$$I_B^{out} = M_\ell(P_s, I_i, I_l, I_q, P_t''), \tag{4}$$

where I_B^{out} denotes the model prediction, and this step enables M_ℓ to focus on higher-level reasoning. The large VLM can focus on executing the described transformation on the query image, rather than figuring out from scratch what the transformation should be. More importantly, this two-stage deployment is highly interpretable: the intermediate text prompt P_t'' clearly explains the intended operation, providing transparency into the system's decision-making. It also offers a point of control – if needed, a human or another module could modify or validate the prompt before the final image generation. This loop leverages the complementary strengths of each model size, the powerful reasoning and generation of large VLM and the efficiency and fast text generation of sVLM to achieve cross-task visual in-context learning that is both effective and practical.

3.3.3 SCORE-BASED REASONING

To enhance decision-making, we introduce perceptual score-based screening based on VIE score (Ku et al., 2023). Since conventional image-synthesis metrics are often task-agnostic and opaque, they provide a single number without revealing why an image is judged good or bad. In contrast, VIE score is a task-aware and explainable evaluator driven by a VLM. Given an instruction I, a synthesized image O, and a set of conditions C^*, the evaluator first produces a natural-language rationale and then a scalar score s:

$$f_{\text{VIE}}(I, O, C^*) = (\text{rationale},\ s), \quad s \in [0, 1]. \tag{5}$$

To reflect the "weakest-link" nature of conditional generation, we decompose the score into *Semantic Consistency* (SC) and *Perceptual Quality* (PQ). Each is formed by minimum aggregation over task-specific sub-scores $\{\alpha_i\}$ and $\{\beta_j\}$ (0–10 scale, later normalized):

$$\text{SC} = \min_i \alpha_i, \qquad \text{PQ} = \min_j \beta_j.$$

The final overall rating uses a geometric mean:

$$O = \big(\text{SC} \cdot \text{PQ}\big)^{1/2}. \tag{6}$$

In practice, we prompt the VLM with explicit rubrics for SC and PQ; notably, PQ is assessed from the synthesized image alone (to avoid instruction confounds), while SC conditions are presented alongside the image. This design yields interpretable rationales, task-aware scores, and robust correlations with human judgments across diverse conditional image tasks.

3.3.4 EVALUATION METRICS

To quantitatively assess our framework, we employ a hybrid metric suite. To complement these, we adopt VIE score for measuring the alignment between generated outputs and task-specific visual improvements, enabling evaluation from a reasoning perspective rather than pixel fidelity alone. For image quality, we report classical fidelity scores including PSNR and SSIM, which remain standard for restoration tasks. PSNR captures pixel-level fidelity via mean-squared error, whereas SSIM reflects perceptual alignment by jointly evaluating luminance, contrast, and structural consistency. For each query image, we performed inference 10 times and selected the result with the highest PSNR as the final output while computing the corresponding VIE score and SSIM. Together, these metrics provide a holistic evaluation that balances low-level fidelity, perceptual similarity, and reasoning coherence.

4 RESULTS

4.1 EXPERIMENTAL SETUP

Datasets. We build our study on a comprehensive collection of twelve representative low-level vision tasks, each paired with a widely adopted benchmark dataset. Specifically, we use GOPRO (Nah, Hyun Kim, and Mu Lee 2017) for image deblurring, D-HAZY (Ancuti, Ancuti, and De Vleeschouwer 2016) for dehazing, UHDM (Yu et al. 2022) for demoiréing, SIDD (Abdelhamed, Lin, and Brown 2018) for denoising, RainCityscapes (Zamir et al. 2021) for deraining, iHarmony4 (Cong et al. 2020) for harmonization, DIV2K (Agustsson and Timofte 2017) with modifications for inpainting and colorization, LoL (Wei et al. 2018) for low-light enhancement, SIR² (Wan et al. 2017) for reflection removal, ISTD (Wang, Li, and Yang 2018) for shadow removal, and Night2Day (Zhu et al. 2017) for style transfer. These datasets collectively span degradations caused by motion, atmospheric conditions, sensor limitations, and lighting deficiencies, as well as creative tasks involving style and appearance transformations.

Implementation details. We conducted our experiments using two NVIDIA A100-SXM4 GPUs with 80 GB of memory each. We follow each dataset's standard train–test split whenever available, or adopt a 70/30 random division when no official split is provided. All images are resized to a unified resolution (448×448 for training inputs, 224×224 for query inputs) to maintain compatibility with the VLM framework.

Main Cross-Task VICL Results Table 1 presents the quantitative comparisons across eight representative tasks. Figure 3 illustrates representative examples of cross-task in-context learning, demonstrating how the VLM model adapts flexibly to diverse tasks driven by conditioning implicit text prompts. Several key observations can be drawn:

Robust overall performance. We present results on eight cross-task pairs to systematically evaluate the generalization ability of our VICL framework. Our framework demonstrates robust performance across a wide range of visual task pairs, with a particular strength in handling low-level cross-task pairs where in-context transfer is traditionally difficult. Unlike prior approaches that are often confined to semantically similar domains, our method adapts seamlessly across both prefix-aligned tasks (e.g., deraining vs. denoising, deblurring vs. demoiring) and prefix-divergent tasks (e.g., shadow removal vs. deraining). This ability to generalize across heterogeneous task pairs indicates that visual in-context learning is not constrained by task similarity. These findings highlight the scalability and flexibility of our approach, showing that VICL can naturally bridge both perceptually related and unrelated tasks under a single model.

Cross-task diversity.

The proposed VICL exhibits consistent gains on image processing tasks (e.g. denoising, deraining, deblurring, low-light enhancement) while preserving a single, decoder-free interface. Beyond processing, the same prompting mechanism transfers to generation-oritented tasks (composition, object insertion/removal, attribute/style manipulation), maintaining spatial layout and identity fidelity. Cross-task transfer holds both within prefix-aligned pairs and across prefix-divergent pairs, without task-specific retraining. Qualitative assessments show reduced artifacts, sharper structures, and stable color consistency, establishing that VICL extends reliably from semantic understanding to both restoration and generative settings.

Generalization to semantically distant tasks. Beyond absolute numbers, an important property is cross-task consistency. This consistency underlines the advantage of framing all tasks under a unified in-context prompting mechanism. More importantly, when applied across a few tasks with large semantic gaps (e.g. denoising vs. light enhancement), VICL maintains stable predictions, whereas task-specific models deteriorate significantly. This highlights robustness as a key byproduct of the unified in-context paradigm.

Table 1: Comparisons.

Task	Metric	Method	
		M1	M2
Task 1 → Task 2	PSNR↑	0	0
	SSIM↑	0	0
	VIE Score↑	0	0
Task 1 → Task 2	PSNR↑	0	0
	SSIM↑	0	0
	VIE Score↑	0	0
Task 1 → Task 2	PSNR↑	0	0
	SSIM↑	0	0
	VIE Score↑	0	0
Task 1 → Task 2	PSNR↑	0	0
	SSIM↑	0	0
	VIE Score↑	0	0

5 CONCLUSION

In conclusion, we propose the first cross-task VICL pipeline, i.e T2T-VICL, that enables collaboration among multiple VLMs, together with an automatic reasoning mechanism. Our framework systematically explores the largely uncharted boundaries of cross-task transfer in VICL. This collaborative paradigm fills a gap in understanding how heterogeneous VLMs can jointly reason and adapt across tasks, and we anticipate that our findings will encourage further investigation into the mecha-

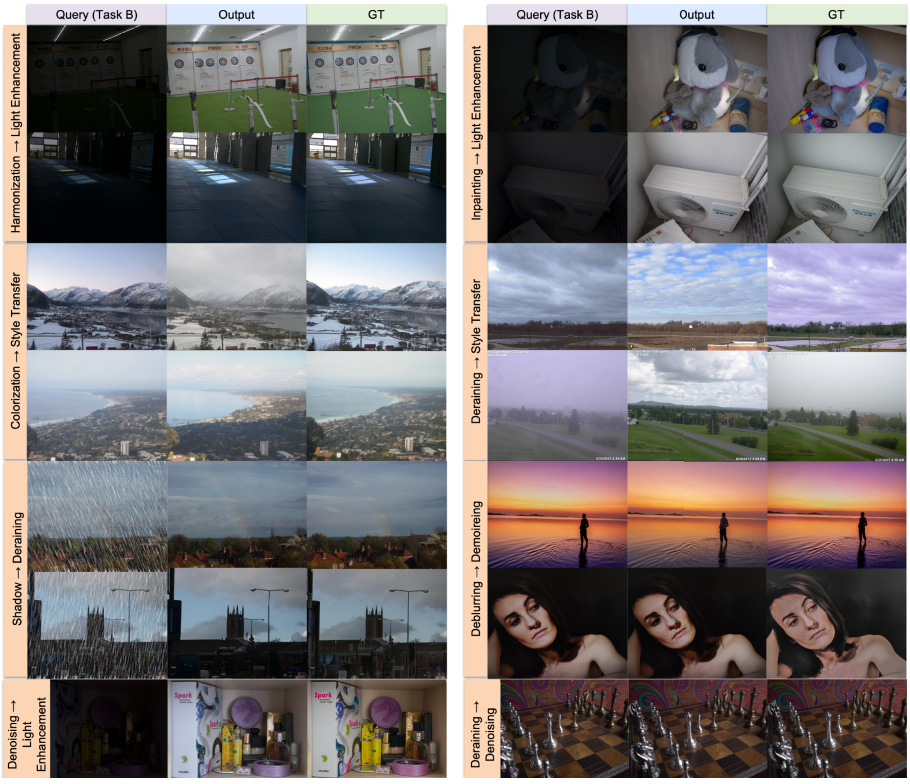

Figure 3: Representative Examples of Cross-Task In-Context Learning

nisms that underpin robust cross-task transfer and the broader generalization capacity of multimodal models.

6 Statement

Ethics Statement. All authors have read and adhere to the ICLR Code of Ethics. If applicable, we discuss below potential ethical concerns, mitigations, or unintended risks associated with our work. While we do not explicitly use sensitive attributes (e.g. race, gender) in training or evaluation, downstream misuse—such as biased editing or manipulations—may pose fairness or privacy risks.

Reproducibility Statement. To facilitate reproducibility, we provide an anonymous repository containing our full implementation, data preprocessing scripts, trained checkpoints, and evaluation code. All datasets are publicly available (or accompanied by instructions to obtain them). Proofs, additional ablation studies, and extended experimental results are included in the supplementary material. We also include instructions to reproduce the main figures and tables in the repository.

LLM Usage Details. LLMs were used to assist in gathering and organizing related work, with all selections and final writings verified by the authors.

References

Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nick Watters, Alexander Lerchner, and Irina Higgins. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6430–6439, 2019.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, pp. 2309–2313. IEEE, 2019.

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Team Chameleon et al. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022.

Lovish Chum, Anbumani Subramanian, Vineeth N Balasubramanian, and CV Jawahar. Beyond supervised learning: A computer vision perspective. *Journal of the Indian Institute of Science*, 99 (2):177–199, 2019.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *ECCV*, 2024.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy and transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12387–12396, 2019.

Y. Gan, S. Park, A. Schubert, A. Philippakis, and A. M. Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint arXiv:2310.00390*, 2023.

Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Hu, D. Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*, 2023.

Clyde Highmore. In-context learning in large language models: A comprehensive survey. 2024.

Menglin Jia, Lijie Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pp. 709–727. Springer, 2022.

Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson WH Lau. Neural preset for color style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14173–14182, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jia Li, Chongyang Tao, Jia Li, Ge Li, Zhi Jin, Huangzhao Zhang, Zheng Fang, and Fang Liu. Large language model-aware in-context learning for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(7):1–33, 2025.

Xiaodong Liu, Xiangyu Zhang, Kaiming He, and Jian Sun. Unifying image processing as visual prompting question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yujie Lu, Gaurav Mishra, Shijie Wang, Zhe Zhang, Zihang Liu, Han Zhang, Christoph Feichtenhofer, Chunyuan Li, Luke Zettlemoyer, and Kai-Wei Chang. Unified-io: A unified model for vision, language, and multi-modal tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yecheng Jason Ma, Joey Hejna, Chuyuan Fu, Dhruv Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, et al. Vision language models are in-context value learners. In *The Thirteenth International Conference on Learning Representations*, 2024.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Grains: Gradient-based attribution for inference-time steering of llms and vlms. *arXiv preprint arXiv:2507.18043*, 2025.

Arvind Pal and Vineeth N. Balasubramanian. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2189–2198, 2019.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023.

Chenyang Qi, Zhengzhong Tu, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Qifeng Chen, and Hossein Talebi. Spire: Semantic prompt-driven image restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Team Show-o et al. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Yan Shu, Weichao Zeng, Zhenhang Li, Fangmin Zhao, and Yu Zhou. Visual text meets low-level vision: A comprehensive survey on visual text processing. *arXiv preprint arXiv:2402.03082*, 2024.

Suzanna Sia, David Mueller, and Kevin Duh. Where does in-context learning happen in large language models? *Advances in Neural Information Processing Systems*, 37:32761–32786, 2024.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 9120–9132, 2020.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality, 2024a. URL https://arxiv.org/abs/2307.05222.

Zeyi Sun, Ziyang Chu, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. X-prompt: Towards universal in-context image generation in auto-regressive vision language foundation models. *arXiv preprint arXiv:2412.01824*, 2024b.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Team Transfusion et al. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Maria Trigka and Elias Dritsas. A comprehensive survey of deep learning approaches in image processing. *Sensors*, 25(2):531, 2025.

Peng Wang, An Yang, Rui Men, Junyang Lin, Yuxiao Lin, Ming Zhou, Chang Zhou, Jingren Lin, and Xu Sun. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning (ICML)*, pp. 23318–23340. PMLR, 2022.

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*, 2023a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for incontext learning. *Advances in Neural Information Processing Systems*, 36:15614–15638, 2023b.

Z. Wang, Y. Jiang, Y. Lu, P. He, W. Chen, Z. Wang, M. Zhou, et al. In-context learning unlocked for diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8542–8562, 2023c.

Zehui Xiao, Li Yuan, Dongdong Zhang, Lu Zhang, Chunyuan Liu, et al. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. A comprehensive evaluation of full reference image quality assessment algorithms. In *2012 19th IEEE International Conference on Image Processing*, pp. 1477–1480. IEEE, 2012.

Xu Zhang, Jiaqi Ma, Guoli Wang, Qian Zhang, Huan Zhang, and Lefei Zhang. Perceive-ir: Learning to perceive degradation better for all-in-one image restoration. *IEEE Transactions on Image Processing*, 2025.

Yutong Zhang, Qibin Hou, Alexander Kolesnikov, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Uvim: A unified modeling approach for vision with learned guiding codes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8751–8760, 2022.

Shihao Zhou, Jinshan Pan, Jinglei Shi, Duosheng Chen, Lishen Qu, and Jufeng Yang. Seeing the unseen: A frequency prompt guided transformer for image restoration. *IEEE Transactions on Image Processing*, 2024a.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024b.

Xizhou Zhu, Jinguo Li, Wenhai Dai, Zehuan Yuan, Qiao Yu, and Jifeng Liu. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16804–16815, 2022.

A Appendix

You may include other additional sections here.