GENIE: A Visual-Only Diffusion Framework for Task-Agnostic Image Transformation

Anonymous authors
Paper under double-blind review

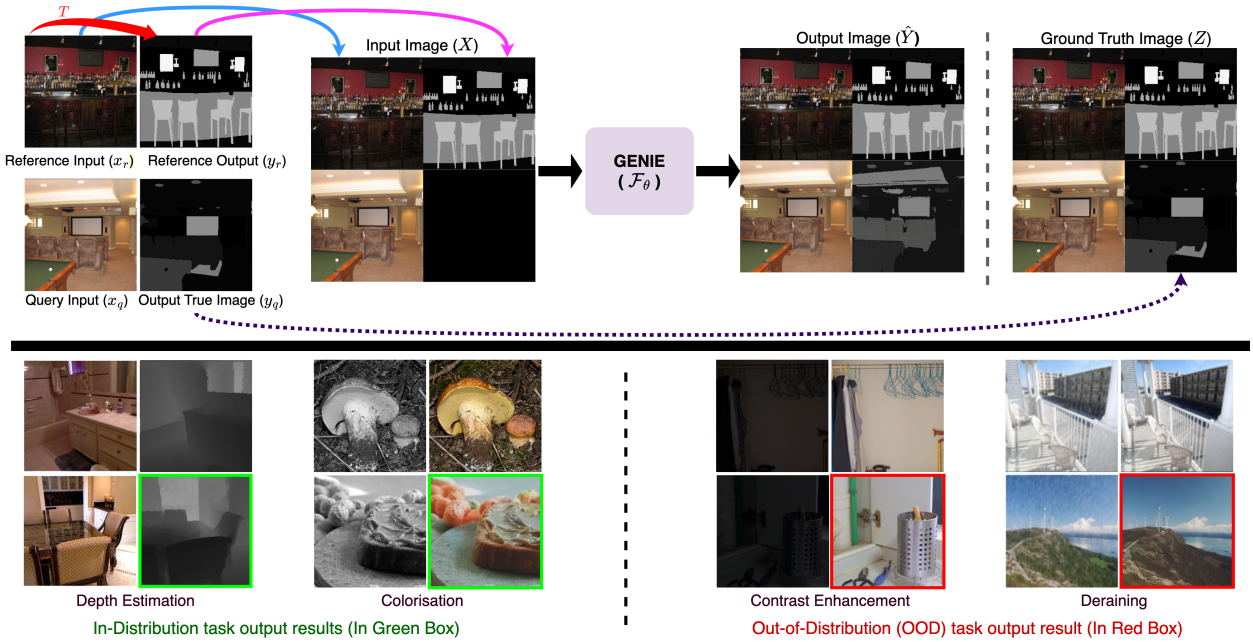

Figure 1: **Illustration of the proposed model, GENIE** ($\mathcal{F}_\theta$), performing diverse visual tasks solely through visual guidance. The **top section** illustrates the model's workflow, where a 2×2 input grid X contains reference input-output pairs (x_r, y_r), a query input image (x_q), and a masked region. GENIE generates a predicted output $\hat{Y}$, shown alongside the ground truth Z; the bottom-right cell represents the target output (y_q). The **bottom section** demonstrates GENIE's predictions across various tasks, highlighting its effectiveness on both in-distribution tasks (green boxes) and out-of-distribution (OOD) tasks (red boxes).

Abstract

Designing a unified vision model capable of handling diverse visual transformation tasks without task-specific modifications remains a significant challenge, particularly in scaling and generalizing beyond narrowly defined objectives. We propose **GENIE**[1], a novel *ControlNet-Diffusion* framework that performs task-based image generation solely through **visual exemplars**, eliminating dependence on textual prompts or auxiliary metadata. Unlike conventional prompt-driven diffusion models, GENIE employs a **dual visual conditioning** mechanism—combining implicit guidance via ControlNet and explicit task encoding through CLIP-based visual arithmetic—to infer task intent directly from reference input-output pairs. To improve semantic alignment between visual exemplars and generated outputs, we introduce a lightweight **task consistency loss**, which encourages representational coherence in the embedding space across transformed pairs. While not a multitask learner in the classical sense, GENIE enables task switching across multiple tasks without

[1]Code will be published upon acceptance.

any task-specific modifications in architecture or task-specific loss functions. Evaluations across **seven vision tasks**—inpainting, colorization, edge detection, deblurring, denoising, semantic segmentation, and depth estimation—and **two out-of-distribution (OOD) tasks**—deraining and contrast enhancement—demonstrate that GENIE achieves an average performance gain of **10%** over visual-conditioned baselines, showcasing its effectiveness for scalable and text-free visual generation.

1 Introduction

The field of computer vision has seen tremendous advancements through specialized models targeting isolated tasks like classification He et al. (2016); Dosovitskiy et al. (2021); Huang et al. (2017); Tan & Le (2019) and segmentation Kirillov et al. (2023); Chen et al. (2018); Ronneberger et al. (2015). Despite their high performance, these models are typically architecturally rigid and limited in their ability to generalize across tasks or domains without retraining. As a result, scalability and adaptability remain critical bottlenecks when attempting to build general-purpose visual systems.

In contrast, *in-context learning* (ICL) Brown et al. (2020) has transformed natural language processing (NLP) by enabling models to perform new tasks at inference time through conditioning on a few input-output examples, without retraining. This strategy, successful in zero-shot and few-shot NLP Chowdhery et al. (2023); Du et al. (2022); Wei et al. (2022), has inspired growing interest in extending ICL principles to computer vision.

However, translating ICL to vision is not straightforward, and recent attempts fall into two distinct categories. *Multi-modal approaches* integrate vision and language, relying on textual prompts to specify the desired task Alayrac et al. (2022); Lu et al. (2023); Wang et al. (2023a); Geng et al. (2024). While powerful, they are inherently limited in settings where text is unavailable, ambiguous, or insufficient to capture complex visual intent. *Vision-only models* circumvent this by using grids of visual examples to condition the model purely through input-output image pairs Bar et al. (2022); Wang et al. (2023b); Bai et al. (2024) (Fig.1). These models often lack mechanisms for explicit task-specific alignment, leading to ambiguity, reduced output fidelity, and poor generalization when faced with complex or out-of-distribution (OOD) tasks (Fig.2).

A representative failure case of text-reliant models is shown in Fig.3, where a vision and text-conditioned model Wang et al. (2023c); Meng et al. (2024) might fail to infer the intended transformation when textual prompt is removed and generate inconsistent result with only visual references. This highlights the need for robust and vision-centric alternatives that can infer intent directly from visual cues.

To address these limitations, we introduce **GENIE** (***Ge**nerative **N**etwork using **I**mage **E**xemplars*), a diffusion-based, text-free framework for exemplar-guided image transformation. GENIE is trained jointly on multiple tasks as a single system with one backbone, one output head, and a lightweight loss function. At

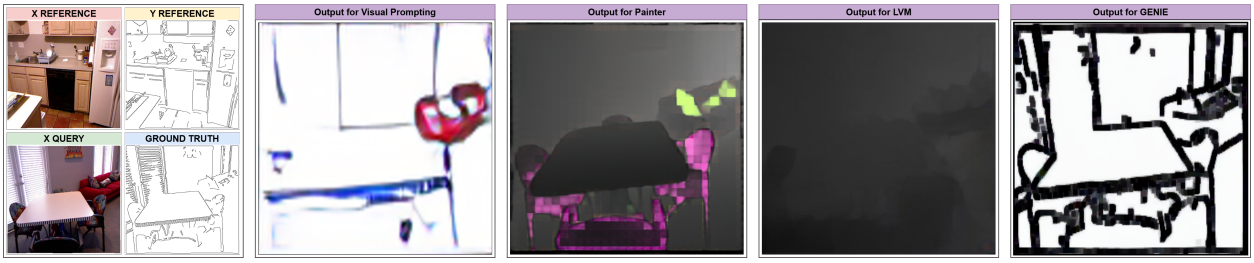

Figure 2: **Comparison of outputs on the OOD scribble conversion task:** (a) Input example with both query and reference images; (b) Output from Visual Prompt Bar et al. (2022), showing limited task adaptation; (c) Output from Painter Wang et al. (2023b), struggling to interpret the task; (d) Output from LVM Bai et al. (2024), also failing to interpret the task effectively; (e) Output from GENIE (ours), demonstrating robust and effective handling of the OOD task. *All models were retrained under the same conditions for fair comparison.*

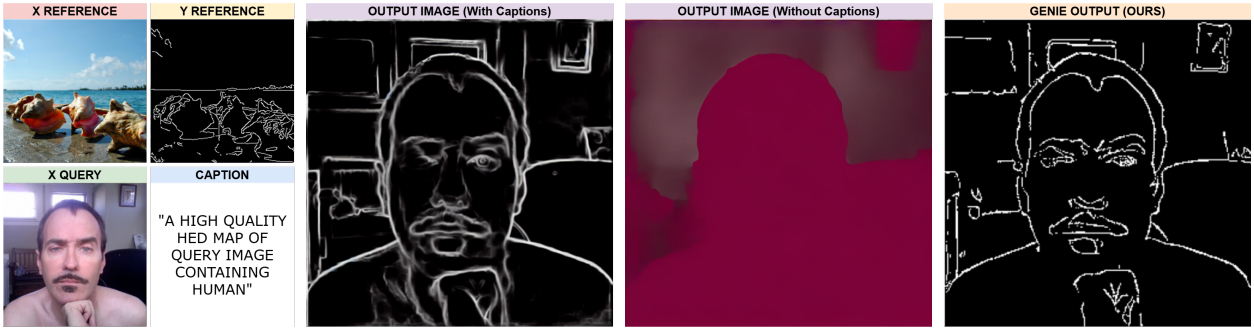

Figure 3: **GENIE vs. Prompt-Diffusion under prompt ablation:** (a) Input exemplar grid with reference and query samples; (b) Output from Prompt-Diffusion using a caption prompt; (c) Output from Prompt-Diffusion without a caption prompt, showing a significant drop in performance due to prompt dependence; (d) Output from GENIE (ours), which performs robustly without requiring textual input, demonstrating strong generalization using purely visual guidance.

inference time, it seamlessly switches between transformations by ingesting different input-output exemplar pairs, demonstrating promising zero-shot generalization to new tasks and domains without any architectural or loss-function modifications.

At the core of GENIE are three key technical components designed to tackle the challenges in exemplar-guided image transformation: **(a)** A *dual visual conditioning mechanism* that combines implicit spatial features from ControlNet (implicit conditioning) with explicit task representations from CLIP-based Radford et al. (2021) visual arithmetic (explicit conditioning). This directly addresses the task alignment problem seen in prior vision-only models and also removes textual dependency from diffusion models. **(b)** A *masked image modeling objective*, which allows generalization and allows the single architecture to support a wide variety of transformation tasks without needing output-space redesigns for each task. **(c)** A lightweight yet effective *task consistency loss* that operates in CLIP's embedding space to enforce semantic coherence between the reference and predicted transformations, ensuring the model preserves transformation intent, especially under domain shifts.

While both InstructGIE Meng et al. (2024) and Prompt Diffusion Wang et al. (2023c) use a reference–query grid (akin to our implicit conditioning), our explicit conditioning goes beyond this by replacing text with CLIP-based visual arithmetic to inject the transformation from reference to query and thereby replacing text conditioning, making entire architecture purely visually conditioned. This is made effective by our task consistency loss, which ensures that the that the transformation vectors of the reference and query, captured by visual arithmetic remain aligned and transfers task intent effectively.

Collectively, these components enable GENIE to generalize effectively across task boundaries, outperforming chosen vision-only baselines—including Painter Wang et al. (2023b), Visual Prompt Bar et al. (2022), and LVM Bai et al. (2024). While numerous related methods exist, these mostly fall into two categories—masked image modeling (MIM) and autoregressive next-token prediction (NTP). We select three representative proxies: Painter for MIM, LVM for autoregressive NTP, and Visual Prompt for MIM with self-supervision on open-set tasks. These baselines are strictly vision-only and not compared against architectures with textual inputs.

Experiments conducted across seven in-distribution(ID) and two OOD tasks reveal that GENIE achieves a consistent **10%** average performance improvement over these baselines, all without text, retraining, or task-specific modules. Our model's distinct advantages are further detailed in Table 1. In summary, our contributions are:

- We propose **GENIE**, a unified, text-free image transformation framework based on diffusion model that integrates both *implicit* (ControlNet-based spatial guidance) and *explicit* (CLIP-based visual feature arithmetic) conditioning to infer task intent directly from images.

Model	Textual Conditioning	Visual Conditioning	External Visual Conditioning	Reconstruction Loss	Task Constraint Loss
Visual Prompt Bar et al. (2022)	×	✓	×	✓	×
Unified-IO Lu et al. (2023)	✓	✓	×	✓	×
Painter Wang et al. (2023b)	×	✓	×	✓	×
Prompt Diffusion Wang et al. (2023c)	✓	✓	×	✓	×
LVM Bai et al. (2024)	×	✓	×	✓	×
PromptGIP Liu et al. (2024b)	×	✓	×	✓	×
InstructGIE Meng et al. (2024)	✓	✓	×	✓	✓
GENIE (ours)	×	✓	✓	✓	✓

Table 1: **Comparison of visual conditioning models** across key architectural attributes. GENIE uniquely combines external visual guidance and task-specific loss without textual prompts.

- We design a **task consistency loss** that operates in CLIP embedding space to enforce semantic alignment between reference and query transformations, enabling task adherence even under domain and task distribution shifts.

- We demonstrate the effectiveness of GENIE across seven diverse in-distribution tasks and two out-of-distribution tasks, showing up to **10% performance gains** over state-of-the-art visual prompting methods.

2 Related Works

Masked Image Modeling. Masked Image Modeling is a core strategy in self-supervised learning. Methods like Context Encoders Pathak et al. (2016) and MAE He et al. (2022) learn powerful, general-purpose visual representations by reconstructing masked image patches. While effective for pretraining, these models are not designed to infer and perform tasks from visual exemplars directly.

In-Context Learning (ICL) with Diffusion Models. ICL, widely used in NLP Brown et al. (2020); Radford et al. (2019), multimodal extensions such as Flamingo Alayrac et al. (2022), Unified-IO Lu et al. (2023), etc. In the visual generative domain, diffusion models Ho et al. (2020); Rombach et al. (2022) are state-of-the-art. ControlNet Zhang et al. (2023) introduced strong visual conditioning but requires separate modules for each task. Subsequent works like Prompt Diffusion Wang et al. (2023c), InstructDiffusion Geng et al. (2024) and InstructGIE Meng et al. (2024) achieve task generalization by fusing visual and textual prompts. However, their fundamental dependence on text limits their use in scenarios that are purely visual or where text is ambiguous.

Visual Prompting. To eliminate text dependency, text-free visual prompting methods like PromptGIP Liu et al. (2024b), Visual Prompt Bar et al. (2022) and Painter Wang et al. (2023b) condition models on masked input-output image grids, following BERT Devlin et al. (2019) style training. These approaches rely on token-level self-attention to infer the task. While effective for some in-distribution tasks, this lack of explicit task alignment can lead to poor generalization on fine-grained or OOD transformations.

Large Vision Models. A recent class of Large Vision Models Bai et al. (2024); Hao et al. (2024) follows a vision-only, autoregressive paradigm, similar to language models predicting the next token. These models show impressive generality but do not incorporate external conditioning, which may limit their effectiveness on tasks that benefit from structured visual guidance.

In summary, current approaches either rely on textual prompts or, if purely visual, lack explicit task grounding, leading to ambiguity in complex or OOD settings. This gap motivates our proposed framework: a text-free diffusion model with dual visual conditioning and a task consistency loss for robust, exemplar-guided transformation.

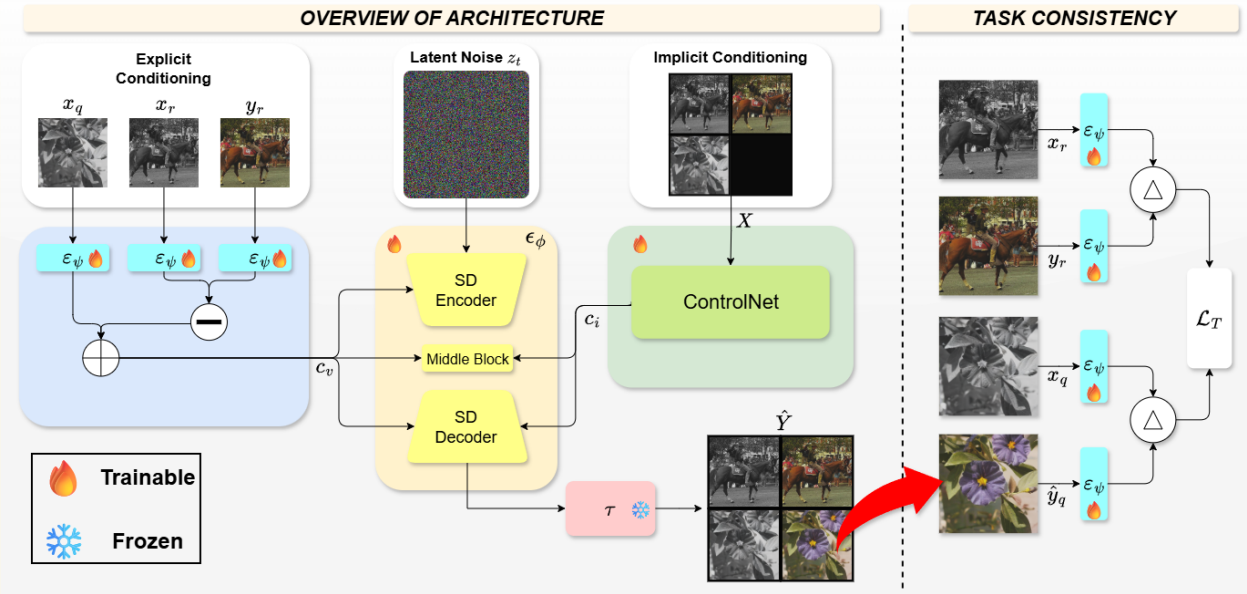

Figure 4: **Overview of the GENIE architecture.** The **left section** illustrates GENIE's dual visual conditioning framework. A 2×2 image grid (X), comprising a query input (x_q), reference input (x_r), and reference output (y_r), is processed through: (i) *explicit conditioning*, where a CLIP-based encoder (ε_ψ) computes a visual task representation (c_v) via visual arithmetic, and (ii) *implicit conditioning*, where ControlNet extracts spatial cues (c_i) directly from X. These embeddings jointly condition the latent diffusion model (ε_ϕ) to generate the query output ($\hat{y}_q$). The **right section** depicts the *task consistency* module, wherein the CLIP encoder (ε_ψ) encodes x_r, y_r, x_q, and $\hat{y}_q$ to enforce semantic consistency between the reference and generated transformations through the task consistency loss ($\mathcal{L}_T$).

3 Proposed Methodology

We propose **GENIE**, a visual-only diffusion framework for prompt-guided conditional generation that performs multiple tasks without any task-specific modifications, using only visual examples without textual prompts or task labels. Its key novelty is a *dual visual conditioning* mechanism that combines implicit token-level guidance through self-attention and explicit semantic guidance via vector arithmetic, conditioned through cross-attention, which replaces the need for textual prompts. Additionally, our novel *task consistency loss* enforces that the semantic transformation from reference input to output is preserved in the query input to its generated output by aligning their embedding. Detailed discussions follow.

3.1 Problem Formulation

Let (x_r, y_r) denote a reference input–output pair, where $y_r = T(x_r)$ and T is an unknown transformation (e.g., colorization, edge detection, or depth estimation). Given a query input x_q, the goal is to generate an output $\hat{y}_q$ such that $\hat{y}_q \approx T(x_q)$, purely by inferring T from the exemplar (x_r, y_r). The true output is denoted as $y_q = T(x_q)$. All images are of shape $\mathbb{R}^{H \times W \times 3}$, with $H = W = 256$.

We arrange the inputs into a 2×2 spatial grid forming a composite image $X \in \mathbb{R}^{2H \times 2W \times 3}$, posing the task as image completion where the model fills in the masked region for y_q, following the setup in Wang et al. (2023b); Bar et al. (2022) as is illustrated in Fig. 4.

$$X = \begin{bmatrix} x_r & y_r \\ x_q & [\text{MASK}] \end{bmatrix}$$

During training, a corresponding completed version Z is constructed by populating the masked region with y_q. The model $\mathcal{F}_\theta$ is trained to reconstruct Z from X and a noisy latent z_t, sampled for timestep t of the

denoising diffusion process, with a focus on generating the missing query output:

$$\hat{y}_q = \mathcal{F}_\theta(z_t, X, x_r, y_r, x_q), \quad \text{and} \quad y_q = T(x_q)$$

Unlike models such as Painter Wang et al. (2023b), our framework normalizes all images to $[0, 1]$ across tasks, avoiding any data-specific modifications and enabling seamless integration of heterogeneous transformations.

3.2 Architecture Overview

As shown in Fig. 4, GENIE is built upon the latent diffusion model Rombach et al. (2022), augmented with a ControlNet module Zhang et al. (2023) for implicit visual guidance (c_i) and a dedicated visual encoder (ε_ψ) for explicit visual guidance (c_v). The model consist of the following components:

(a) VQ-GAN (τ)**:** A pre-trained VQ-GAN Esser et al. (2021) encoder encodes the $512 \times 512 \times 3$ input grid Z into a $64 \times 64 \times 4$ latent z_0, enabling efficient training in a compact space. The decoder reconstructs $\hat{y}_q$ from the denoised latent.

(b) ControlNet Module (Implicit Guidance): The masked grid X is processed by the ControlNet conditioning module to extract spatially aligned implicit features c_i via self-attention. This enables the model to learn the task transformation from the reference pair and apply it to the query content during end-to-end training. The extracted features are then passed to the denoising network to fill the masked region.

(c) CLIP Encoder (Explicit Guidance): The encoder ε_ψ maps x_r, y_r, and x_q to 1024-D CLIP embeddings. Using task vector arithmetic, it constructs c_v to represent the intended transformation direction, serving as a semantic prior.

(d) Denoising Network (ϵ_ϕ)**:** A Transformer-U-Net predicts noise at each timestep t using conditioning from both c_i and c_v. Each processing block includes convolution, ResNet layers He et al. (2016), and ViT modules Dosovitskiy et al. (2021) to capture both local detail and global context.

All components of the model are trained end-to-end ensuring both structural fidelity and semantic alignment. For further architectural details, including interaction among components within the unified model $\mathcal{F}_\theta$, refer to Sup Mat.

3.3 Dual Visual Conditioning

(a) Implicit Conditioning via ControlNet. The composite image X, containing the reference pair (x_r, y_r) and the query input x_q, is fed into a ControlNet module. Through self-attention, the model extracts implicit features c_i that capture task-relevant visual transformations such as color changes, structural modifications, or denoising patterns, and adapts them to the query content. These features are injected into multiple layers of the denoising network, providing fine-grained spatial cues without any explicit task labels. The entire conditioning process is trained end-to-end to complete the masked region in X.

Since X implicitly defines the transformation through exemplar placement, this structure allows GENIE to generalize across task types. While the current approach samples (x_r, y_r) randomly, future work can incorporate selection strategies based on visual or semantic similarity.

(b) Explicit Conditioning via Visual Arithmetic. To capture high-level task semantics, we utilize a CLIP ViT-B/32 encoder $\varepsilon_\psi(\cdot)$ to extract 1024-D semantic embeddings for x_r, y_r and x_q. Compared to alternatives like VQ-VAE Rombach et al. (2022), ViTs Dosovitskiy et al. (2021), or LLaVA-style caption generators Liu et al. (2024a), CLIP embeddings were found to offer the best trade-off between task-discriminability and image content (see Sup Mat).

$$e_{x_r} = \varepsilon_\psi(x_r), \quad e_{y_r} = \varepsilon_\psi(y_r), \quad e_{x_q} = \varepsilon_\psi(x_q)$$

We compute the reference task vector via:

$$v_r = e_{y_r} - e_{x_r} \tag{1}$$

This difference vector represents the semantic shift induced by the transformation T. To apply T to x_q, we generate a target embedding via:

$$c_v = e_{x_q} + v_r \tag{2}$$

This analogy-inspired construction draws from the insight that CLIP's embedding space $\mathcal{E}$ encodes meaningful transformations as linear paths Radford et al. (2021); Reed et al. (2015). If T maps $x_r \to y_r$, then c_v approximates the desired location of $\hat{y}_q$ in $\mathcal{E}$.

During denoising, c_v is injected into all cross-attention layers of the denoising network, which helps to disambiguate tasks that have visually similar structures but distinct semantics, and removes the need for text conditioning.

To prevent the model from overfitting to the explicit signal c_v and ignoring c_i, we randomly zero out c_v in 50% of training iterations. This forces the model to learn from both structural and semantic cues in a balanced manner, improving robustness under domain shift.

3.4 Learning Objective

GENIE is trained using a composite loss that combines generative fidelity with semantic alignment.

(a) Diffusion Loss. We adopt the standard latent denoising objective from diffusion models Rombach et al. (2022). Let z_0 denote the clean latent encoding of the composite grid Z, and z_t its noisy version at timestep t. The denoising U-Net ϵ_ϕ predicts the added noise ϵ, conditioned on both implicit (c_i) and explicit (c_v) visual cues:

$$\mathcal{L}_D = \mathbb{E}_{z_0,t,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\phi(z_t, t, c_v, c_i)\|_2^2\right] \tag{3}$$

(b) Task Consistency Loss. To ensure that the model performs the correct semantic transformation on the query image based solely on the reference pair, we propose a task consistency loss $\mathcal{L}_T$ that promotes structural equivalence of transformations in the embedding space. Rather than relying on explicit task labels, we assume that transformations manifest as vector displacements in a shared semantic manifold $\mathcal{E} \subset \mathbb{R}^d$, parameterized by a CLIP-based visual encoder ε_ψ.

Inspired by analogy-based learning Reed et al. (2015); Hertzmann et al. (2023), we model a transformation T as the difference between the reference pair embeddings:

$$v_r = \varepsilon_\psi(y_r) - \varepsilon_\psi(x_r). \tag{4}$$

After conditioning the diffusion model on the reference pair and query input x_q, the output $\hat{y}_q$ is reconstructed by decoding the denoised latent:

$$\hat{y}_q = \tau\left(z_t - \epsilon_\phi(z_t, t, c_v, c_i)\right), \tag{5}$$

and the query transformation is similarly computed as:

$$v_q = \varepsilon_\psi(\hat{y}_q) - \varepsilon_\psi(x_q). \tag{6}$$

To enforce semantic consistency, we define the task consistency loss to minimize the difference in squared displacement magnitudes between the two transformation vectors:

$$\begin{aligned}\mathcal{L}_T &= \mathbb{E}\left[\left(\|v_r\|^2 - \|v_q\|^2\right)^2\right] \\ &= \mathbb{E}\Big[\Big(\|\varepsilon_\psi(y_r) - \varepsilon_\psi(x_r)\|^2 - \|\varepsilon_\psi(\hat{y}_q) - \varepsilon_\psi(x_q)\|^2\Big)^2\Big]\end{aligned}$$

Further Insights to $\mathcal{L}_T$: From an analytical point of view, the loss operates in a feature space structured by contrastive pretraining of ε_ψ where semantically meaningful concepts are organized as direction. t-SNE clustering in the Sup Mat supports this assumption, showing that input–output difference vectors, indicating the direction from input to output, naturally group by task.

Category	Dataset	Associated Tasks
In-Distribution (ID) Data & Task	ImageNet Deng et al. (2009)	Deblurring, Super-resolution, Inpainting, Colorization, Edge Detection, Denoising
	ADE20K Zhou et al. (2017)	Semantic Segmentation
	NYU Depth V2 Silberman et al. (2012)	Depth Estimation
OOD Data & OOD Task	Deraining Zamir et al. (2022)	Deraining
	LOL Wei et al. (2018)	Low-light Enhancement
OOD Data & ID Task	DomainNet Peng et al. (2019)	Denoising, Inpainting, Edge Detection, Colorization

Table 2: **Overview of dataset–task configurations** used for evaluation. OOD settings—either due to novel tasks or unseen domains—are highlighted in gray.

The task consistency loss, based on above explanation, is designed to ensure that the semantic transformation captured by the reference pair is similar to the semantic transformation between the query input and its generated output. It does not assume uniformity across all tasks (e.g., deraining under light vs. heavy rain) but instead ensures that the relative semantic shift observed in the reference pair is mirrored in the query. This is done by minimizing the difference between their task representation (Eqn. 4 and Eqn. 6) in embedding space, providing a lightweight inductive bias that stabilizes the strength of transformations inferred from exemplars. Thus encouraging both pairs to represent the same task strength without requiring identical pixel-level changes.

Geometrically, it enforces that v_r and v_q lie on the same hypersphere in $\mathcal{E}$, preserving the geodesic radius of the transformation. Operating in a contrastively structured feature space, this avoids both under- and over-editing, thereby supporting generalization in both ID and OOD settings where task style may vary but transformation strength should remain coherent.

Final Objective. GENIE is trained to using:

$$\mathcal{L} = \mathcal{L}_D + \lambda_T \cdot \mathcal{L}_T \tag{7}$$

We empirically set $\lambda_T = 0.05$ to balance generation quality and transformation alignment. This weighting allows GENIE to consistently perform high-fidelity synthesis while preserving the intended transformation semantics across a wide range of tasks. Increasing λ_T further improves task alignment but leads to visual artifacts, revealing a trade-off between semantic fidelity and perceptual quality. This choice is also motivated by findings in prior works Isola et al. (2017), where a high weighting ($\lambda = 100$) is placed on reconstruction objectives to emphasize alignment.

4 Experimental Evaluations

4.1 Training Setup and Dataset Details

For each task, we curated a diverse set of 20,000 randomly sampled images from multiple datasets (detailed in Table 2), to support a range of vision tasks. ImageNet Deng et al. (2009), ADE20K Zhou et al. (2017), and NYU Depth V2 Silberman et al. (2012) were used as primary sources to ensure visual diversity and broad task coverage. To assess GENIE's generalization, we structure the evaluation into three categories (Table 2): (a) in-distribution (ID) tasks on in-distribution data (e.g., denoising), (b) out-of-distribution (OOD) tasks from unseen datasets (e.g., deraining), and (c) ID tasks evaluated on visually dissimilar OOD domains using DomainNet Peng et al. (2019). Additional dataset details are in the Sup Mat.

All components of the GENIE are jointly trained end-to-end on all tasks, without any task specific modification to architecture or loss function, using AdamW optimizer with a weight decay of 0.01 Loshchilov & Hutter (2019) and a learning rate of 1×10^{-5} on a single NVIDIA A100 GPU with 80GB of memory, using a batch size of 16 for 20 days, ensuring fairness in comparison and evaluation.

4.2 Evaluation Strategy

We evaluate deblurring, super-resolution, denoising, colorization, and inpainting on the ImageNet validation set Deng et al. (2009) using task-specific metrics: PSNR for deblurring and super-resolution; PSNR and

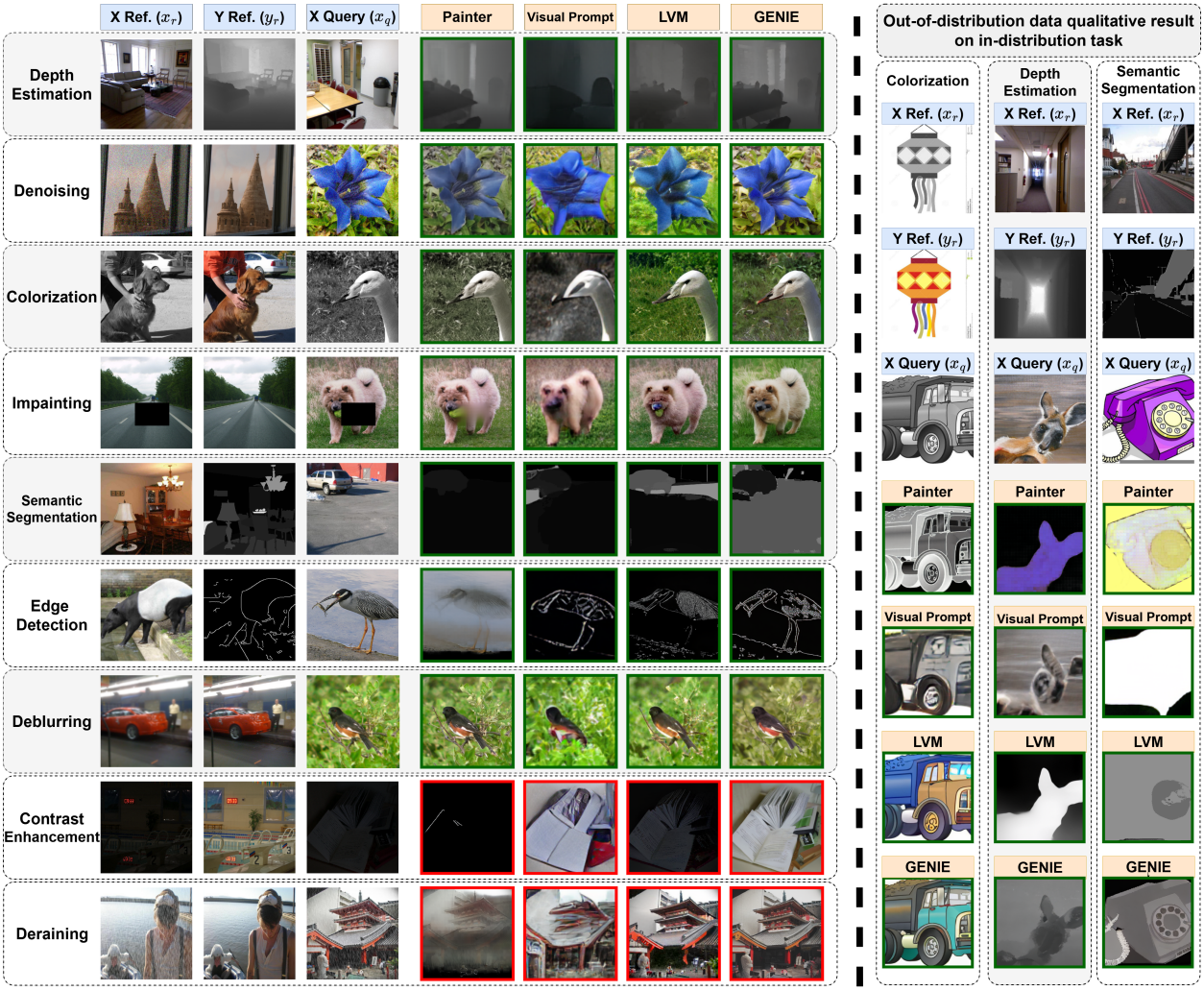

Figure 5: **Comparison of model results across various tasks.** The left portion shows results on in-distribution (green outline) and OOD tasks (red outline), while the right portion displays performance on a different domain data Peng et al. (2019). While *Painter* Wang et al. (2023b) struggles with OOD tasks and *Visual Prompt* Bar et al. (2022) often distorts structures, both *LVM* Bai et al. (2024) and *GENIE* show strong performance across many tasks. However, *GENIE* demonstrates superior generalization, producing better results than *LVM*. Also, *LVM*, *Visual Prompt*, and *GENIE* successfully segment objects but often misassign labels due to patch reconstruction in latent space.

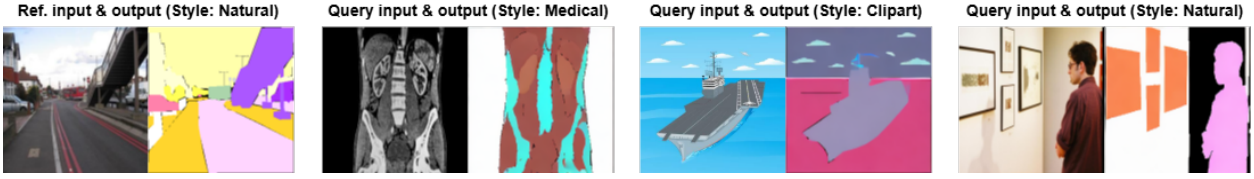

Figure 6: The model maintains consistent segmentation quality across diverse input domains, including natural images, medical scans, and clipart-style scenes.

In Distribution Task														
	Depth Estimation	Denoising			Super Resolution		Deblurring		Colorization		Inpainting			Semantic Segmentation
Models	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	MSE ↓	LPIPS ↓	SSIM ↑	MSE ↓	LPIPS ↓	mIoU ↑
Visual Prompt Bar et al. (2022)	0.112	17.513	0.532	0.411	17.786	0.397	18.020	0.384	0.028	0.337	0.374	0.044	0.557	-
Painter Wang et al. (2023b)	0.063	19.060	**0.842**	0.366	19.230	0.386	**24.331**	**0.161**	0.022	0.264	0.429	0.030	0.442	40.620
LVM Bai et al. (2024)	0.071	19.413	0.793	0.373	18.687	0.372	22.246	0.377	0.019	0.261	0.558	0.018	0.411	-
GENIE	**0.061**	**23.712**	0.827	**0.261**	**22.323**	**0.271**	23.210	0.218	**0.016**	**0.255**	**0.632**	**0.013**	**0.381**	-

Table 3: **Quantitative results of models on ID Tasks**. Metrics marked with "↓" indicate lower is better, and metrics marked with "↑" indicate higher is better. The hyphen symbol "-" indicates that the model outputs unrelated or ambiguous results, making quantitative evaluation unsuitable. Best results are in "bold" and second-best are underlined.

SSIM for denoising; MSE for colorization; and SSIM and MSE for inpainting. Depth estimation and semantic segmentation are evaluated using RMSE on the NYU Depth V2 validation set Silberman et al. (2012) and mIoU on the ADE20K validation set Zhou et al. (2017), respectively. We also report LPIPS Zhang et al. (2018) for all image reconstruction tasks to assess perceptual quality.

To assess GENIE's generalization to OOD tasks, we evaluate its performance on deraining using the dataset Zamir et al. (2022) and on contrast enhancement using the LOL dataset Wei et al. (2018), using PSNR, SSIM and LPIPS. We further conduct qualitative analysis on the Painting, Sketch, and Clipart subsets of DomainNet Peng et al. (2019) to examine robustness under domain shift.

GENIE is compared against three vision-only baselines: Visual Prompting Bar et al. (2022), Painter Wang et al. (2023b), and LVM Bai et al. (2024). All models are retrained and evaluated across a shared set of tasks as described above to ensure fair and consistent comparison.

4.3 Qualitative Results

We qualitatively compare GENIE with Painter, Visual Prompting, and LVM on both in-distribution (green boxes) and OOD tasks (red boxes) in Fig. 5.

(a.) In-Distribution Generalization: For ID tasks, Painter produces high-quality outputs. Visual Prompting correctly infers the task but often introduces structural artifacts. In contrast, LVM and GENIE consistently generate high-fidelity, semantically accurate outputs that preserve fine image details across all tasks.

(b.) Out-of-Distribution Generalization: On OOD tasks like contrast enhancement and deraining, Painter's performance degrades significantly; it fails to adapt and often reverts to identity mappings or familiar in-distribution behaviors. Visual Prompting again shows moderate task inference but produces distorted and blurred results.

GENIE and LVM generalize far better to these unseen tasks. However, their performance varies: LVM excels at deraining but struggles with contrast enhancement and scribble generation (Fig. 2). This is likely because deraining shares low-level priors (e.g., texture removal) with LVM's pretraining, whereas tasks like contrast enhancement require domain-specific color transformations that GENIE is better equipped to model. Notably, GENIE also visually outperforms LVM on in-distribution tasks when applied to OOD-style inputs (e.g., Painting and Clipart datasets). While both models show some degradation on OOD tasks, the outputs remain visually coherent, highlighting their potential as generalized models. More qualitative results for GENIE are available in the Sup Mat.

4.4 Quantitative Results

We evaluated GENIE on a wide range of in-distribution and OOD tasks, with results summarized in Table 3 and Table 4. Baselines that failed to generate valid outputs on certain tasks were excluded from comparison.

(a.) In-Distribution Tasks: GENIE excels across ID tasks, achieving top performance in depth estimation, super-resolution, colorization, and inpainting. For *denoising*, both GENIE and LVM achieve higher PSNR than Painter but have slightly lower SSIM and LPIPS. This is expected, as their latent-space reconstruction

prioritizes structural fidelity, which can affect perceptual metrics like SSIM. Similar observation was made for deblurring. For *segmentation*, GENIE successfully segments multiple objects in complex scenes (Fig. 5 & 8) and across domains (Fig. 6). However, due to a lack of explicit pixel-level supervision in generation, it occasionally assigns incorrect class labels, a limitation also present in models predicting in latent space, such as LVM and Visual Prompting. This likely stems from them optimizing for semantic understanding rather than exact pixel prediction making metrics like mIoU an incomplete indicator of quality. Overall, GENIE improves performance by an average of **12.2%** over the second best across all ID tasks.

(b.) Out-of-Distribution Tasks: GENIE also performs robustly on unseen OOD tasks like contrast enhancement and deraining, consistently outperforming the next-best models, Visual Prompting and LVM. In *contrast enhancement*, GENIE outperforms LVM by 12.6% in PSNR, 25.9% in SSIM and 5.8% in LPIPS. In *deraining*, it achieves 13.3% higher SSIM while maintaining a competitive PSNR and LPIPS. Across OOD tasks, GENIE shows a mean improvement of **6.30%** over LVM, demonstrating its strong generalization capabilities.

In total, GENIE achieves an average performance gain of approximately **10%** across all evaluated tasks, validating its effectiveness as a unified framework for diverse image-to-image translation.

4.5 Inference Time

GENIE runs for ∼8s with 50 ddim steps, ∼3s with 20 ddim steps, and ∼1s with 5 ddim steps on a single A100 GPU. As seen in Fig. 7, 20 and 50 steps are qualitatively the same, so inference time is ∼3s per image. This performance is on par with any typical diffusion model.

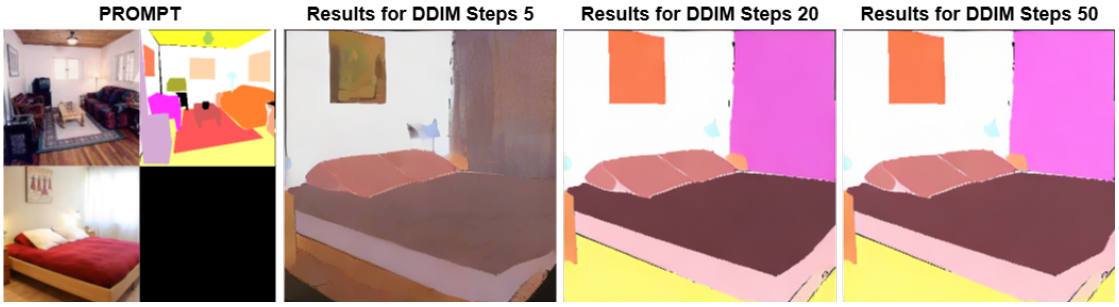

Figure 7: Segmentation result (color coded for each class) for different DDIM steps.

4.6 Ablation Study

To isolate the contribution of GENIE's key components, we conducted an ablation study, summarized in Table 5. We evaluated four model variants on five representative tasks (three ID, two OOD), starting from a minimal baseline to the full architecture.

Out of Distribution Task						
Models	**Contrast Enhancement**			**Deraining**		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Visual Prompt Bar et al. (2022)	12.090	0.308	0.929	14.983	0.391	0.473
Painter Wang et al. (2023b)	-	-	-	-	-	-
LVM Bai et al. (2024)	<u>16.440</u>	<u>0.533</u>	<u>0.619</u>	**18.112**	<u>0.454</u>	**0.310**
GENIE	**18.513**	**0.671**	**0.585**	<u>17.414</u>	**0.513**	<u>0.357</u>

Table 4: **Quantitative results of models on OOD Tasks**. Metrics marked with "↓" indicate lower is better, and metrics marked with "↑" indicate higher is better. The hyphen symbol "-" indicates that the model outputs unrelated or ambiguous results, making quantitative evaluation unsuitable. Best results are in "bold" and second-best are <u>underlined</u>.

Our study starts with **Model A**, which uses only implicit conditioning (ControlNet). We then introduce our proposed components individually: **Model B** adds the **task consistency loss**, which significantly boosts performance on generalization-heavy tasks like depth estimation and OOD transformations. **Model C** adds **explicit visual conditioning**, which excels on content-sensitive tasks such as inpainting and colorization. The full architecture, **Model D**, integrates both explicit conditioning and the task consistency loss. This complete model achieves superior results across all tasks, yielding an average performance improvement of **6.8%** over the next-best variant. This outcome confirms that explicit visual guidance and semantic consistency are complementary, validating GENIE's integrated design.

			In-Distribution Tasks			OOD Task	
Model	**Explicit Conditioning**	**Task Consistency**	**Depth Estimation** (RMSE ↓)	**Colorization** (MSE ↓)	**Inpainting** (SSIM ↑)	**Contrast Enhancement** (PSNR ↑)	**Deraining** (SSIM ↑)
A	×	×	0.081	0.023	0.519	14.477	0.452
B	×	✓	0.069	0.021	0.597	16.789	0.504
C	✓	×	0.073	0.017	0.613	16.343	0.481
D	✓	✓	0.061	0.016	0.632	18.513	0.513

Table 5: **Ablation study results across selected tasks**. Models: A) Implicit conditioning only, B) Implicit conditioning + task consistency, C) Implicit + explicit conditioning, D) GENIE.

Figure 8: **Qualitative results showing GENIE's limitations. Top:** An OOD task, transitioning from map-style to aerial satellite imagery from Isola et al. (2017). GENIE attempts to fill the map and results in ambiguous and unrealistic outputs. **Bottom:** A segmentation task where GENIE struggles with class label assignments, producing inconsistent segmentation masks.

5 Conclusions & Future Directions

We introduced GENIE, a unified diffusion framework for text-free multi-task visual prompt learning. Driven by its dual visual conditioning and a novel task consistency loss, GENIE achieves superior generalization and fidelity across diverse ID and OOD tasks, outperforming state-of-the-art visual-only baselines.

Despite its strong performance, GENIE shows limitations, particularly in segmentation where latent-space modeling can cause label misalignment, and on highly divergent OOD tasks (Fig. 8). Future work could address this by incorporating pixel-level supervision for generated output and expanding the training data to further enhance GENIE's adaptability for real-world vision challenges.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems*, volume 35, pp. 23716–23736, 2022.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22861–22872, 2024.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25005–25017, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

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. In *9th International Conference on Learning Representations ICLR 2021*, 2021.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 5547–5569, 2022.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 12709–12720, 2024.

Zhiwei Hao, Jianyuan Guo, Chengcheng Wang, Yehui Tang, Han Wu, Han Hu, Kai Han, and Chang Xu. Data-efficient large vision models through sequential autoregression. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 557–570. Association for Computing Machinery, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pp. 6840–6851, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, 2024a.

Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Zichong Meng, Changdi Yang, Jun Liu, Hao Tang, Pu Zhao, and Yanzhi Wang. Instructgie: Towards generalizable image editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 18–34, 2024.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Boyu Peng, Chi Wang, Yang Shi, Yixin Zhang, Xiaochun Wei, Qixing Huang, and Chen Change Loy. Moment alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1738–1747, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Scott Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, pp. 1252–1260, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 61501–61513, 2023a.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023b.

Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8542–8562, 2023c.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *10th International Conference on Learning Representations, ICLR 2022*, 2022.

Zhe Wei, Yiyi Zhang, Zhen Wu, Nanning Zheng, and Jing Zhang. Low-light image enhancement using a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1557–1565, 2018.

Syed Waqas Zamir, Aditya Arora, Salman Hameed Khan, Hayat Munawar, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Appendix

A Introduction

This supplementary material provides additional analysis of GENIE's explicit conditioning mechanism, its performance on out-of-distribution (OOD) datasets and tasks, and an examination of its limitations. In Section B, we detail the design choices behind explicit conditioning, including our rationale for selecting CLIP-ViT Radford et al. (2021) as the visual encoder over alternatives such as VLMs Liu et al. (2024a), VQ-VAE Esser et al. (2021), and ViT Dosovitskiy et al. (2021). Section D outlines the datasets used for training and evaluation, highlighting their characteristics and relevance to task diversity. Section E presents qualitative results on OOD datasets from varied visual domains, including Clipart, Painting, and Sketch subsets of DomainNet, showcasing GENIE's generalization ability. This section also includes failure cases, focusing on GENIE's challenges in extreme OOD tasks and semantic segmentation scenarios.

Figure 9: The figure illustrates the result of LLaVA Liu et al. (2024a) (VLM) in generating captions for reference pair transformations. Blue-bordered images (top) represent correctly generated captions. Red-bordered images (bottom) indicate incorrect captions, where the VLM misinterprets the transformation.

B Explicit Conditioning Choice

B.1 Vision-Language Models (VLMs) for Task Transformation Understanding

We explored VLMs, specifically LLaVA Liu et al. (2024a), to We explored the feasibility of interpreting transformations between reference image pairs $(x_{\mathrm{r}}, y_{\mathrm{r}})$ in a zero-shot setting using a vision-language model (VLM). The goal was to replace predefined, task-specific text prompts with captions generated by the VLM that describe the transformation from input to output in the reference pair. These captions would then be passed through the CLIP text encoder to provide conditioning information within the stable diffusion pipeline.

To facilitate this, the VLM was prompted with the following instruction:

Caption: *The image is a concatenation of two images side by side. Tell me the relationship between the images. You can instruct like "The right image is* `<task>` *of the left image." Choose the closest* `<task>` *from: "Segmentation," "Denoised," "Colorization," "Hed map," "Boundary image," "Grayscale," "Black-Masked," "Depth-image" based on your knowledge.*

The VLM performed reasonably well on simpler transformations. For instance, it correctly identified colorization tasks with captions like "The right image is a colorized version of the left image." However, it frequently failed to recognize more complex transformations such as segmentation, edge detection, and masked-image reconstruction (Fig. 9). In these cases, the model often produced ambiguous or incorrect descriptions—e.g.,

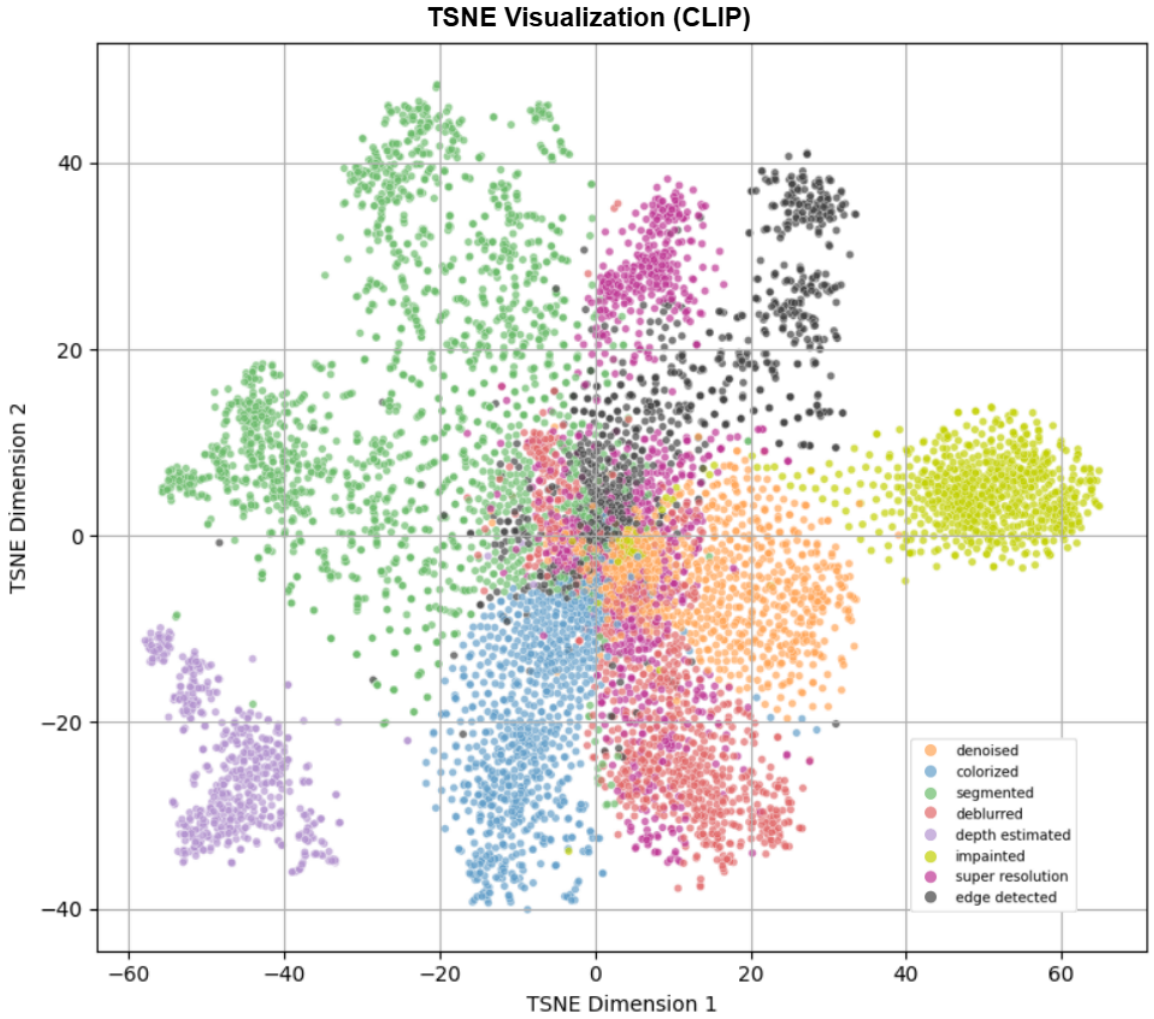

Figure 10: **t-SNE Visualization of CLIP-ViT task-specific embeddings.** The embeddings represent differences between the transformed image and the original image for each task. CLIP-ViT provides well-separated task clusters.

interpreting edge detection as "black and white image" or generating generic captions lacking task-specific clarity.

These limitations stem from the VLM's training objectives, which prioritize general content understanding over fine-grained task-specific transformations. While effective for high-level semantic alignment, this makes VLMs unreliable for providing precise task context. As a result, we conclude that such captioning mechanisms are inadequate for replacing textual prompts in the CLIP text encoder within our framework. Consequently, we shift our focus to visual encoders for explicit conditioning, aiming to fully eliminate textual dependencies in GENIE.

B.2 Encoder Evaluation

To identify the most suitable visual encoder for capturing task-specific transformations in our framework, we experimented with three encoders: CLIP-ViT Radford et al. (2021), ViT Dosovitskiy et al. (2021), and VQ-VAE. Each encoder's embeddings were analyzed using t-SNE van der Maaten & Hinton (2008) to visualize how well the embeddings differentiate various transformation tasks such as deblurring, masking,

noise removal, and edge detection. The t-SNE projections represent the **difference embeddings**, which capture the difference between the transformed image and the original image for each task.

B.2.1 Task-Specific Embeddings

- **CLIP-ViT and ViT**: We evaluated both CLIP-ViT and ViT in zero-shot settings, and their respective t-SNE projections (Figures 10 and 11) reveal well-separated task clusters, indicating their effectiveness in capturing task-specific visual semantics. While their performance was comparable, we chose CLIP-ViT for our framework. This decision is motivated by the fact that CLIP's visual and textual encoders are jointly trained using contrastive learning, which aligns the visual embedding space with text semantics. As a result, using CLIP-ViT to replace the textual encoder in the diffusion model's conditioning pathway ensures greater compatibility and semantic consistency within the latent diffusion architecture.

- **VAE**: In contrast, the t-SNE projection for VAE-based embeddings (Fig. 12) shows substantial overlap between clusters, particularly for visually similar tasks such as deblurring and super-resolution. This poor task separability suggests that VAE embeddings lack the discriminative capacity to model fine-grained task transformations, making them unsuitable for use in GENIE's conditioning pipeline.

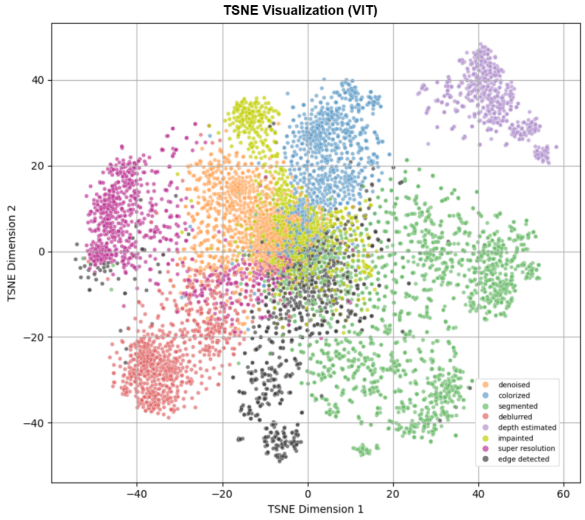

Figure 11: **t-SNE Visualization of ViT task-specific embeddings.** Task clusters are well-separated, demonstrating strong feature extraction capabilities.

Figure 12: **t-SNE Visualization of VAE task-specific embeddings.** The embeddings show significant overlap between tasks, particularly deblurring and super-resolution.

Although the learned task-specific embeddings capture meaningful task semantics, they may also contain irrelevant information such as content-related features or noise in a few of the tasks. We rely on training to help the model suppress this noise and emphasize task-relevant signals. As part of future work, we plan to improve the robustness of these embeddings by explicitly disentangling task-specific information from content and noise.

B.2.2 Content-Specific Embeddings

We further evaluated whether Diffusion-VAE and CLIP-ViT can effectively cluster images based on object information, with each cluster corresponding to a distinct object class. If an encoder can successfully cluster objects, it indicates that the encoder is capable of deciphering the semantic information about the object. This ability is crucial for conditioning the diffusion model in downstream tasks, as the extracted object semantics and transformations can be used to guide the generation process.

As seen in Fig.13, Diffusion-VAE embeddings show significant overlap between object classes, indicating a struggle in capturing object-specific features. This limitation suggests that Diffusion-VAE may not be well-suited for conditioning the diffusion model based on object semantics. In contrast, Fig.14 demonstrates that CLIP-ViT embeddings form distinct clusters for each object class. This clustering suggests that CLIP-ViT is capable of extracting the necessary features and can be further used to condition a diffusion model.

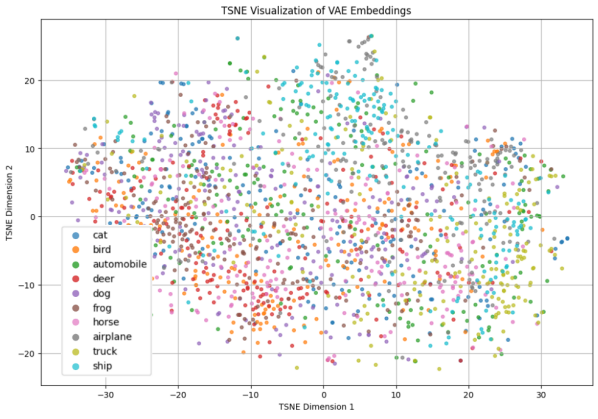

Figure 13: **t-SNE Visualization of VAE content-specific Embeddings.** The embeddings show significant overlap between object classes, indicating limitations in capturing object-specific features.

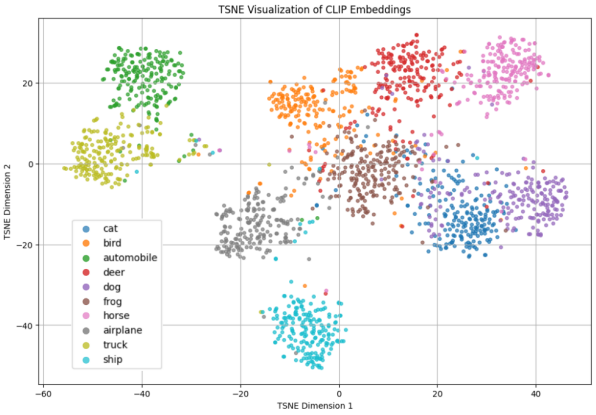

Figure 14: **t-SNE Visualization of CLIP-ViT content-specific Embeddings.** The embeddings form distinct clusters for each object class, indicating effective extraction of object semantics.

C Architecture Details

GENIE's full model $\mathcal{F}_\theta$ integrates the ControlNet-based denoising architecture, a CLIP-based visual encoder, and a VQ-GAN encoder-decoder for latent compression. We summarize the core equations and flow used in GENIE for clarity and reference.

C.1 Latent Compression

The unmasked 2×2 grid $Z \in \mathbb{R}^{512\times512\times3}$ is encoded using a pre-trained VQ-GAN Esser et al. (2021) encoder (τ_{enc}) to obtain the latent representation z_0 as shown below. VQ-GAN also includes a decoder (τ_{dec}); together, they are denoted as τ.

$$z_0 = \tau_{enc}(Z) \tag{8}$$

C.2 Visual Conditioning

The implicit condition c_i is extracted from the masked composite 2×2 grid $X \in \mathbb{R}^{512\times512\times3}$ using ControlNet Zhang et al. (2023), while the explicit condition c_v is computed via CLIP (ε_ψ) embeddings using transformation arithmetic:

$$e_{x_r} = \varepsilon_\psi(x_r), \quad e_{y_r} = \varepsilon_\psi(y_r), \quad e_{x_q} = \varepsilon_\psi(x_q) \tag{9}$$

$$c_v = e_{y_r} - e_{x_r} + e_{x_q} \quad \text{(explicit visual condition)} \tag{10}$$

$$c_i = \text{ControlNet}(X) \quad \text{(implicit condition)} \tag{11}$$

C.3 Latent Diffusion

The latent vector z_0 undergoes progressive noise addition over t timesteps to produce the noisy latent $z_t \in \mathbb{R}^{64\times64\times4}$. The denoising network ϵ_ϕ then predicts the noise $\hat{\epsilon}$ conditioned on timestep t, explicit

Dataset	Number of Training Samples	Number of Testing Samples	Tasks
ImageNet	20000 per task	20000	Impainting, Deblurring, Colorization, Edge Detection, Super-Resolution, Denoising
ADE 20K	20000	3000	Segmentation
NYU V2	20000	654	Depth Estimation
LOL	-	15	Contrast Enhancement
Deraining	-	500	Deraining

Table 6: Summary of datasets used for training and evaluating GENIE, along with their associated tasks. The " - " symbol indicates datasets that were exclusively used for evaluation and not included in GENIE's training pipeline.

condition c_v, and implicit condition c_i:

$$\hat{\epsilon} = \epsilon_\phi(z_t, t, c_v, c_i) \tag{12}$$

C.4 Fetching Output

To reconstruct the query output $\hat{y}_q$, the predicted noise $\hat{\epsilon}$ is subtracted from z_t to obtain the denoised latent $\tilde{z}_0$, which is then passed through the VQ-GAN decoder:

$$\tilde{z}_0 = z_t - \hat{\epsilon} \tag{13}$$

$$\hat{y}_q = \tau_{dec}(\tilde{z}_0) \tag{14}$$

D Datasets

D.1 In-Distribution Training Dataset

ImageNet-22K: ImageNet-22K Deng et al. (2009) is a large-scale, hierarchical visual dataset containing over 14 million images categorized into 22,000 classes.

ADE20K: ADE20K Zhou et al. (2017) is a scene-centric dataset comprising over 25,000 images annotated for 150 semantic categories, including both indoor and outdoor environments. It is widely used for segmentation tasks, with 20,000 images allocated for training, 2,000 for validation, and 3,000 for testing .

NYU Depth V2: NYU Silberman et al. (2012) Depth V2 contains RGB-D images captured in 464 indoor scenes using a Microsoft Kinect camera. It provides dense depth annotations alongside RGB data, with 24,000 images in the training set and 654 images in the test set.

D.2 Out-of-distribution (OOD) Testing Dataset

LOL: The Low-Light Dataset (LOL) Wei et al. (2018) consists of paired low-light and enhanced images, allowing models to learn to enhance visibility and detail in poorly lit scenes. This dataset is commonly used for low-light image enhancement tasks. It contsist a total of 500 image pairs, split into 485 images for training and 15 pairs for testing.

DomainNet: DomainNet Peng et al. (2019) is a large-scale dataset spanning six diverse domains, including real, sketch, clipart, and painting. Since our model is trained on ImageNet Deng et al. (2009), NYUv2 Silberman et al. (2012), and ADE20k Zhou et al. (2017), we perform qualitative evaluations on the painting,

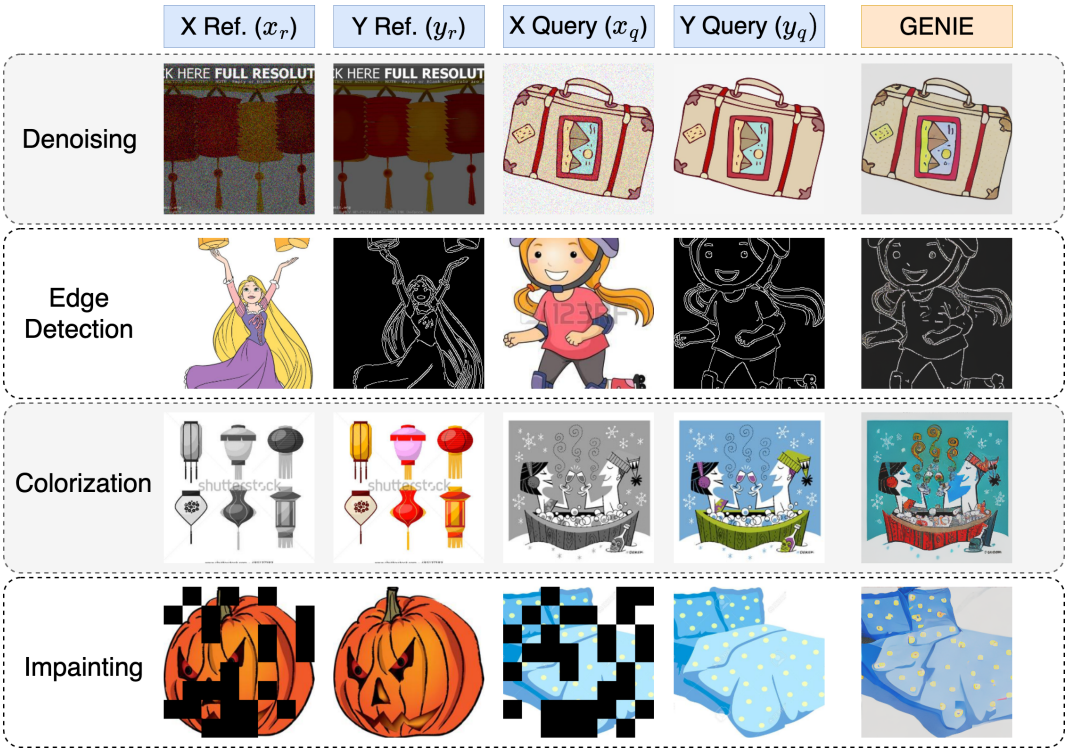

Figure 15: **Qualitative result of GENIE on various tasks on CLIPART domain from domain net Peng et al. (2019) dataset.**

sketch, and clipart domains. These domains differ significantly from the natural image datasets used during training, offering an opportunity to assess our model's adaptability.

Deraining Dataset: The Deraining dataset Zamir et al. (2022) contains images with rain streaks and their corresponding clear versions. It is used to test how well a model can remove the rain and restore clear images. This dataset is a common benchmark for evaluating performance in OOD scenarios. The dataset consists of 500 images for training and 500 images for testing. In our experiments, we utilized all 500 test images to test our model.

E Qualitative Results

This section provides qualitative examples of GENIE's performance across a range out-of-distribution (OOD) data. figures 15, 17, and 16 display results on three different domains from DomainNet Peng et al. (2019). These examples highlight the model's capacity to generalize task transformations under OOD data scenarios.

While GENIE demonstrates robust performance in some OOD tasks (as shown in Fig.18), Fig.19 highlights some of the model's limitations. These includes challenges in accurately segmenting objects into the correct classes as intended in the ADE20k dataset, challenges in deciphering scribble inputs to generate realistic outputs, and the inability to capture intricate details of pose estimation tasks from the reference image. These examples showcase areas where further refinement is needed, particularly in handling complex input-output transformations and finer task intricacies.

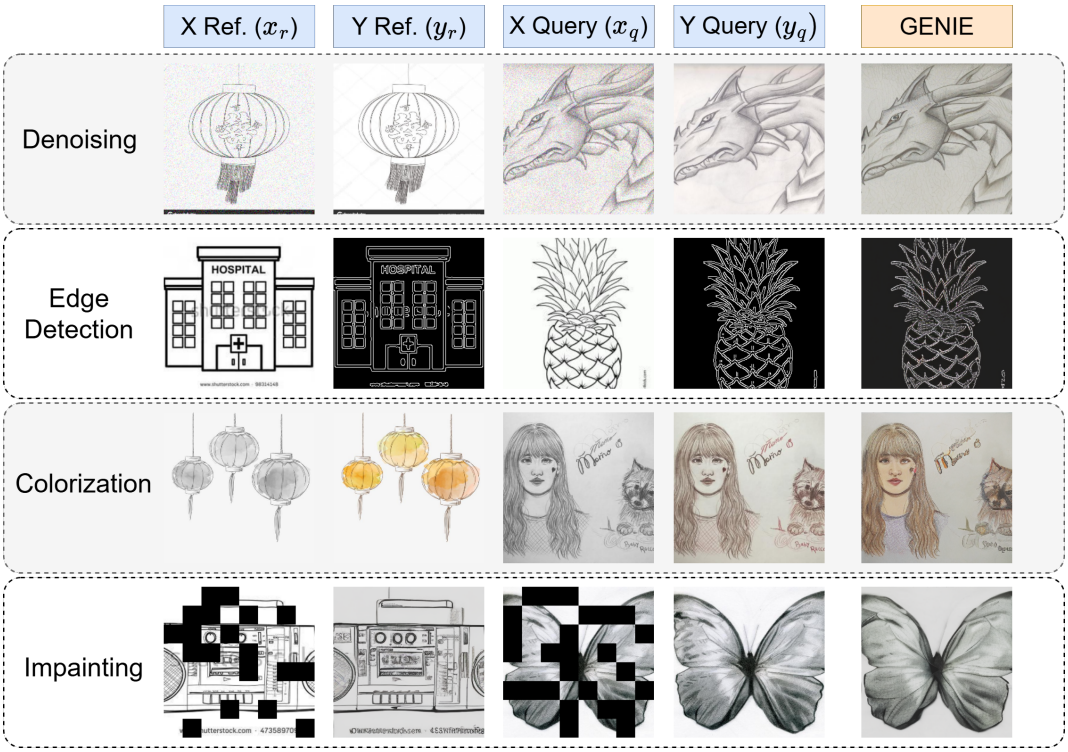

Figure 16: Qualitative result of on various tasks on SKETCH domain from domain net Peng et al. (2019) dataset.

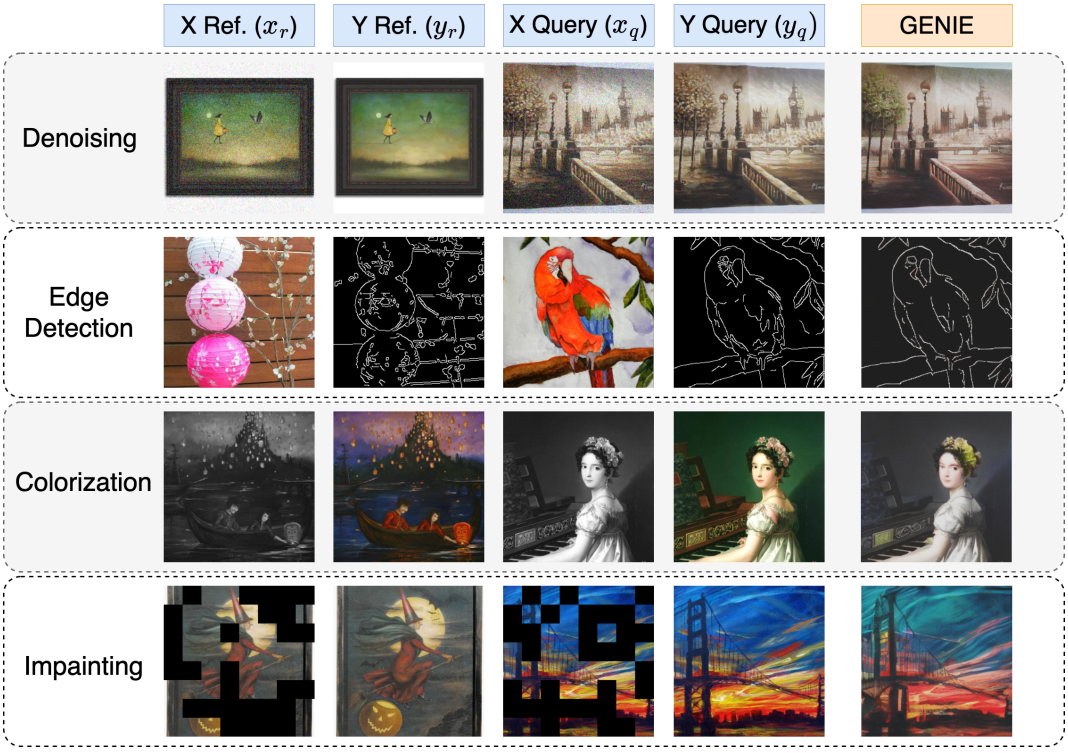

Figure 17: Qualitative Result result of GENIE on various tasks on PAINTING domain from domain net Peng et al. (2019) dataset.

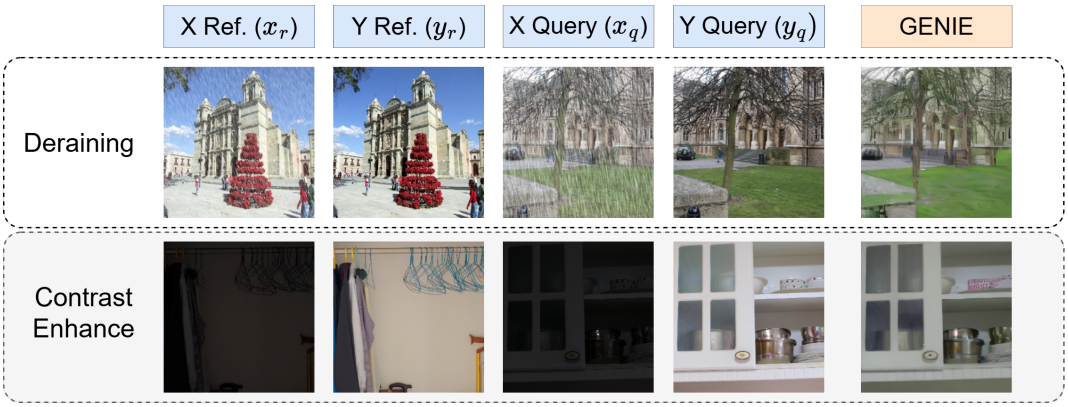

Figure 18: Qualitative result on Out-of-distribution tasks.

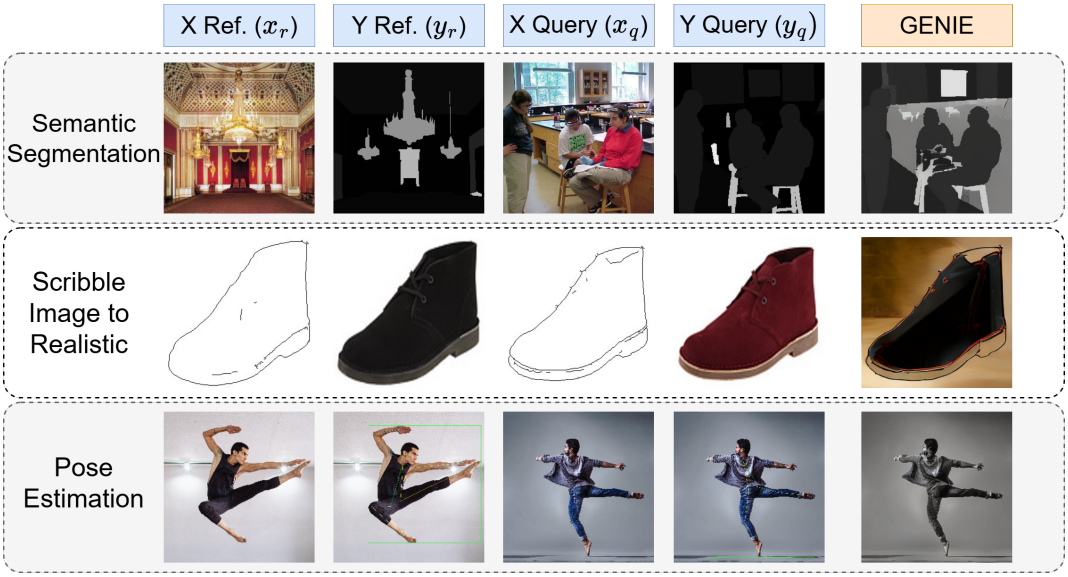

Figure 19: **Limitations of GENIE:** The figure illustrates limitations in semantic segmentation where though the model is able to segment the objects in scene but it struggles to assign the class as intended in the dataset, challenges in generating realistic shoes from scribble input by filling with random colors, and difficulties in capturing fine details for pose estimation from the reference image.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in neural information processing systems*, volume 35, pp. 23716–23736, 2022.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22861–22872, 2024.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25005–25017, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

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. In *9th International Conference on Learning Representations ICLR 2021*, 2021.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 5547–5569, 2022.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 12709–12720, 2024.

Zhiwei Hao, Jianyuan Guo, Chengcheng Wang, Yehui Tang, Han Wu, Han Hu, Kai Han, and Chang Xu. Data-efficient large vision models through sequential autoregression. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 557–570. Association for Computing Machinery, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pp. 6840–6851, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, 2024a.

Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Zichong Meng, Changdi Yang, Jun Liu, Hao Tang, Pu Zhao, and Yanzhi Wang. Instructgie: Towards generalizable image editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 18–34, 2024.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Boyu Peng, Chi Wang, Yang Shi, Yixin Zhang, Xiaochun Wei, Qixing Huang, and Chen Change Loy. Moment alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1738–1747, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Scott Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, pp. 1252–1260, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 61501–61513, 2023a.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023b.

Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8542–8562, 2023c.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *10th International Conference on Learning Representations, ICLR 2022*, 2022.

Zhe Wei, Yiyi Zhang, Zhen Wu, Nanning Zheng, and Jing Zhang. Low-light image enhancement using a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1557–1565, 2018.

Syed Waqas Zamir, Aditya Arora, Salman Hameed Khan, Hayat Munawar, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.