Cycle-Consistent Tuning for Layered Image Decomposition

Zheng Gu[1,5] Min Lu[2] Zhida Sun[1] Dani Lischinski[3] Daniel Cohen-Or[1,4] Hui Huang[1*]
[1]CSSE, Shenzhen University [2]Shenzhen University [3]Hebrew University of Jerusalem [4]Tel Aviv University
[5]State Key Lab. for Novel Software Technology, Nanjing University
{guzheng.szu, lumin.vis, zhdsun, danix3d, cohenor, hhzhiyan}@gmail.com

Figure 1. Our method learns to disentangle overlaid logos from their supporting surfaces and recompose them seamlessly onto other objects. Each example shows two input photographs of objects with distinct logos. We first decompose each image into its logo and object layers, and then cross-compose the separated logos onto the other objects. These results demonstrate accurate separation and faithful re-integration across challenging non-linear cases involving complex geometry, lighting, and viewpoint changes.

Abstract

Disentangling visual layers in real-world images is a persistent challenge in vision and graphics, as such layers often involve non-linear and globally coupled interactions, including shading, reflection, and perspective distortion. In this work, we present an in-context image decomposition framework that leverages large diffusion foundation models for layered separation. We focus on the challenging case of logo-object decomposition, where the goal is to disentangle a logo from the surface on which it appears while faithfully preserving both layers. Our method fine-tunes a pretrained diffusion model via lightweight LoRA adaptation and introduces a cycle-consistent tuning strategy that jointly trains decomposition and composition models, enforcing reconstruction consistency between decomposed and recomposed images. This bidirectional supervision substantially enhances robustness in cases where the layers exhibit complex interactions. Furthermore, we introduce a progressive self-improving process, which iteratively augments the training set with high-quality model-generated examples to refine performance. Extensive experiments demonstrate that our approach achieves accurate and coherent decompositions and also generalizes effectively across other decomposition types, suggesting its potential as a unified framework for layered image decomposition. Project page: link.

1. Introduction

Image decomposition has long been a challenging problem in computer vision and computer graphics, aiming to factorize images into physically meaningful components. Classic approaches such as intrinsic decomposition [5, 6, 44] separate reflectance from shading, relying on explicit priors and rigid task formulations. More recent methods leverage alpha-channels [45] for layered decomposition [36, 40, 43], yielding structured outputs via additive layer compositing.

*Corresponding author

These techniques, however, are largely confined to settings where components interact linearly (e.g., via alpha blending [36, 45]). In contrast, isolating a logo from a product photographed under non-frontal viewpoints involves highly non-linear and globally coupled interactions driven by shading, perspective distortion, surface reflectance, and material-dependent appearance [25]. Such cases cannot be resolved by local or patch-based analysis alone; they require non-local reasoning and semantic understanding of what constitutes an object versus an overlaid element. Addressing this challenge calls for data-driven priors that capture scene- and object-level context, as encoded by modern foundation models [22, 33].

In this work, we leverage the representational power of large foundation models to tackle logo-object image decomposition. Rather than relying on handcrafted priors or local statistics [6], we adopt a pretrained image inpainting diffusion model to separate an overlaid logo from its supporting object, producing (i) a rectified logo layer (fronto-parallel and largely illumination-invariant) and (ii) a "clean" object image. We realize this adaptation via a lightweight LoRA fine-tuning [17] tailored to this task. Our training follows the In-Context Learning (ICL) paradigm [18], where the supervision is presented in a single three-panel grid image. This encourages the model to internalize the operation of removing or isolating overlaid elements while preserving the underlying structure. Importantly, our method is not a training-free, one-shot scheme [3, 15] conditioned on example pairs during inference; instead, we adopt the data-driven route and train on a carefully curated dataset to ensure the outputs remain contextual, consistent, and faithful to the input. This design imparts task-specific decomposition capabilities while retaining the broad generalization of the pretrained model, with performance that scales with the availability and quality of training data.

To make this training effective and robust, we introduce a coupled decomposition–composition framework with cycle consistency. The decomposition module predicts two outputs from a composite input: the rectified logo and the logo-free object. A complementary composition module then reconstructs the original image from these predicted components. A cycle-consistency loss enforces agreement between the reconstruction and the input, allowing the two modules to supervise each other and reducing the need for densely annotated ground truth. Practically, we organize supervision as triplets (composite, logo, clean object) when available, and fall back on the cycle signal when ground-truth layers are incomplete or imperfect. This joint training substantially stabilizes learning and improves fidelity in the presence of real-world non-linearities.

We further propose a self-improving data loop that progressively enlarges and refines the training set. Starting from a small seed of annotated triplets, we train an initial LoRA, use it to generate additional candidate labeled triplets, and select high-quality results to refine the LoRA and yield more reliable labeled data. After that, we train the cycle-consistent model and use it to propose additional decompositions on unlabeled images, selecting high-quality results via automated filters and simple human-in-the-loop checks. The accepted results are then added back into the training pool for subsequent rounds. Combined with our cycle-consistent objective, this bootstrapping strategy steadily improves robustness and semantic accuracy across diverse viewing conditions, materials, and lighting.

We validate the effectiveness of the approach through extensive experiments, as illustrated in Figure 1. While our primary focus is logo–object decomposition, we also demonstrate the generality of the framework by applying it to two distinct problems: foreground–background separation and ambient-level decomposition of albedo and lighting. These results suggest a general paradigm for image decomposition that can handle complex, non-linear, and semantically coupled interactions well beyond the specific application studied in this work.

2. Related Work

2.1. Diffusion Models

Diffusion models [16] have emerged as the leading framework for high-fidelity visual generation in recent years. They model complex data distributions through a progressive denoising process [34]. Since the Stable Diffusion [33] and the Diffusion Transformer [30], these approaches have advanced text-to-image synthesis [4], image translation [24], and controllable image editing [29, 46]. Recent variants such as the FLUX family [22, 23] further improves contextual awareness, enabling localized and spatially consistent manipulations. Building on this progress, we extend diffusion models to image decomposition, where a single input is split into distinct components.

2.2. Visual In-Context Learning

In-Context Learning (ICL) [12], originated from large language models (LLMs), indicates the ability of pretrained models to adjust their behavior by leveraging contextual signals [21]. Early works of Visual ICL focus on discriminative tasks such as image segmentation [37, 38, 47]. Some later works explore generative visual ICL, such as conditional image generation [3, 39, 42] and example-based image editing [7, 15, 26]. More recently, researchers have found that existing Diffusion Models possess inherent in-context abilities during their generation process [18], which enables applications such as image editing through text instructions [23, 48]. Although these techniques demonstrate strong capability, they are mostly confined to *single-input-single-output* mappings. In this work, we unlock the *single-input-multi-output* potential of visual ICL. We achieve this via a simple but effective cycle-consistent tuning strategy, which decomposes and recomposes with contextual information to enforce mutual correspondence between layers.

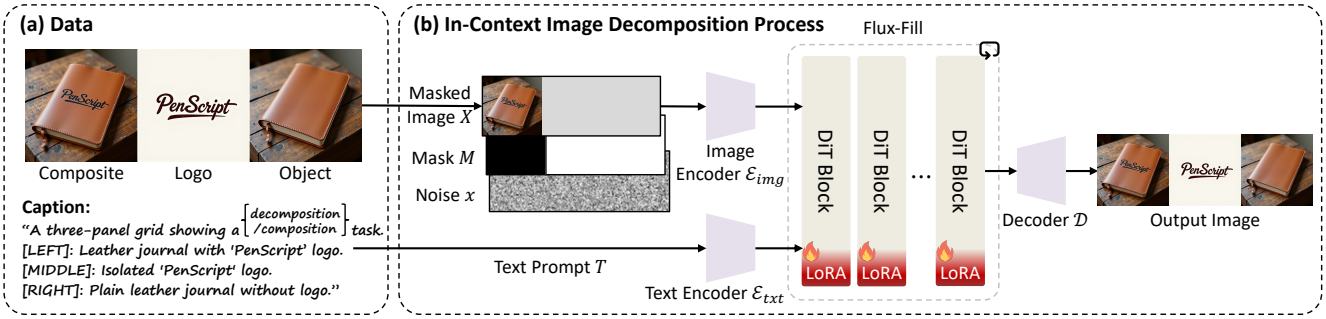

Figure 2. Overview of the image decompostion framework. Given a composite image, the model receives a masked input, a binary mask indicating the logo region, and a noise latent, and predicts both the isolated logo and the clean object. The process is implemented by tuning a LoRA on top of Flux-Fill. The composition scheme is similar but uses a complementary mask to produce the composite image.

2.3. Asset Extraction via Generative Models

Recent work has explored the use of generative models for extracting reusable assets directly from images. Affara et al. [1] explored asset extraction for urban street assets through rectified object proposals. RRM [14] introduces radiance-guided material extraction to obtain relightable 3D assets. AssetDropper [25] formulates this problem through conditional image-to-image translation, learning to isolate standardized assets from user-specified regions. It employs a reward-driven optimization process that aligns the extracted assets with visual priors. In contrast, we view this problem from a more general layered decomposition angle, where the decomposition is learned jointly with another composition process. Rather than relying on explicit masks or fixed extraction views, our approach infers the layered structure directly from context, enabling decomposition that preserves layers faithfully beyond asset extraction.

3. Method

In this work, we present a framework for image decomposition using large generative models. Leveraging the contextual understanding capabilities of Diffusion Transformers, we fine-tune a pretrained image inpainting model with LoRA to specialize in separating components from single composite images (Section 3.1). To ensure consistency and plausibility between input and decomposed layers, we jointly train a complementary composition model under a cycle-consistency constraint (Section 3.2). Furthermore, we adopt a progressive data collection with a self-improvement strategy that iteratively expands the training using generated pseudo-decompositions filtering (Section 3.3).

3.1. Preliminaries

Image Inpainting with DiTs. Our method is built upon FLUX.1-Fill-dev [22], a Diffusion Transformer model designed for image inpainting. Given a masked image X, a binary mask M, and noise latent x, the model iteratively refines the masked region through a conditional Transformer guided by a text prompt T. As shown in Figure 2, at each timestep t, the forward process can be formulated as:

$$x_{t-1} = \phi\left(v_\theta\left(\left[x_t, \mathcal{E}_{img}\left(X\right), M\right], t, \mathcal{E}_{txt}\left(T\right)\right)\right), \quad (1)$$

where v_θ denotes the model, $\mathcal{E}_{img}$ and $\mathcal{E}_{txt}$ are image and text encoders, and ϕ is the flow-matching [13] scheduler. After denoising, the latent x_0 is fed into the VAE decoder $\mathcal{D}$ to reconstruct the complete image.

Low-Rank Adaptation (LoRA). We adopt Low-Rank Adaptation (LoRA) [17] to efficiently fine-tune the Diffusion Transformer. Given a parameter matrix $W \in \mathbb{R}^{d\times k}$, LoRA learns the following matrices:

$$W' = W + UV, \quad (2)$$

where $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{r\times k}$ are LoRA parameters ($r \ll d$). This enables lightweight adaptation while preserving the pretrained model's capacity.

Specifically, to finetune the inpainting model for decomposition, flow matching loss is utilized to encourage the Transformer to predict the velocity field of the forward process as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{x,t}\left[\left\| v_\theta\left(x_t, M, t, \tau\right) - \frac{\partial x_t}{\partial x}\right\|_2^2\right], \quad (3)$$

where $\theta = (U, V)$. M is set to zeros in the regions to be preserved and ones in the regions to be inpainted.

3.2. Cycle Consistent Decomposition

Typically, image decomposition is an ill-posed problem since the number of unknowns exceeds the number of inputs. Although we can collect ground truth decomposition images to some extent, the supervision is still insufficient to constrain the model. To address this challenge, we introduce a cycle consistency constraint to provide additional regularization to the solution space. Unlike decomposition, composition is a well-defined and deterministic process. Therefore, by jointly learning decomposition and composition in a cyclic manner, we enable the model to leverage

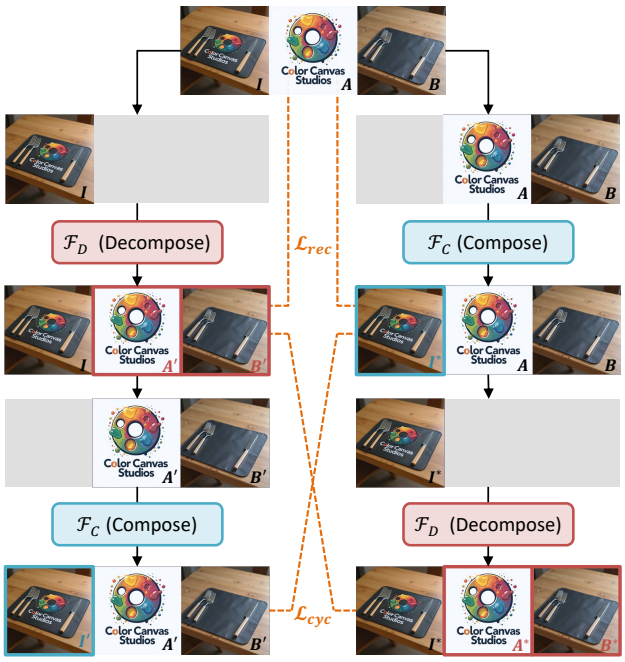

Figure 3. Illustration of our cycle-consistent training. The model jointly learns decomposition and composition, ensuring the decomposed layers recompose into the original image and vice versa.

complementary knowledge from both tasks, thereby stabilizing the training and reducing ambiguity.

As shown in Figure 3, the model is given an example triplet of images $\langle I, A, B\rangle$, where I denotes the combination of A and B. Note that this combination can be either linear alpha blending or more complex nonlinear interactions, such as logo-object composition. Our goal is to learn the following two functions at the same time:

$$\begin{cases} \mathcal{F}_D(I) = \langle A, B\rangle, \\ \mathcal{F}_C\left(\langle A, B\rangle\right) = I, \end{cases} \tag{4}$$

where $\mathcal{F}_D$ is the decomposition function and $\mathcal{F}_C$ is the composition function. To learn both functions, during each training step, we operate decomposition and composition symmetrically to supervise each other in two parallel tracks:

1. Starting from I, first predict $\langle A', B'\rangle = \mathcal{F}_D(I)$, then recompose them into $I' = \mathcal{F}_C(\langle A', B'\rangle)$;
2. Starting from A and B, first predict $I^* = \mathcal{F}_C(\langle A, B\rangle)$, then decompose it into $\langle A^*, B^*\rangle = \mathcal{F}_D(I^*)$.

We employ a cycle consistency loss $\mathcal{L}_{cyc}$ to drive the above learning process:

$$\begin{aligned} \mathcal{L}_{cyc} = \mathbb{E}_{x,t_1}\Big[\|v_\theta(x^I_{t_1}, M_D, t_1, \tau_D) - v_\theta(x^{I^*}_{t_1}, M_D, t_1, \tau_D)\|_2^2\Big] \\ + \mathbb{E}_{x,t_2}\Big[\|v_\theta(x^{\langle A,B\rangle}_{t_2}, M_C, t_2, \tau_C) - v_\theta(x^{\langle A',B'\rangle}_{t_2}, M_C, t_2, \tau_C)\|_2^2\Big], \end{aligned} \tag{5}$$

where t_1 and t_2 are two different timesteps. $\{M_D, \tau_D\}$ and $\{M_C, \tau_C\}$ are the binary mask and text tokens for decomposition and composition, respectively. In practice, the two functions share the same LoRA parameter space θ, which enhances parameter efficiency and stabilizes the training.

3.3. Progressive Data Collection

Training high-quality decomposition models usually requires large paired datasets. However, collecting such data for our logo–object decomposition is costly and hard to scale. To overcome this data scarcity, we propose a progressive self-improvement strategy that iteratively expands the training corpus using model-generated data.

Seed Data Collection. We initiate the process with a small seed dataset comprising 100 manually curated $\langle I, A, B\rangle$ triplets, with the assistance of GPT-4o [19] (Figure 4(a)). This seed set is used to train an initial IC-LoRA [18], which is then employed to generate a large candidate pool of new triplets from text prompts.

Iterative Data Generation with IC-LoRA. Given the sparse initial data, the first-generation IC-LoRA exhibits instability and often fails to produce plausible decompositions. A naive expansion of the dataset with these low-quality samples would lead to error propagation. To address this, we generate data iteratively (Figure 4(b)). We use Qwen-VL [2] to filter generated triplets based on visual plausibility and decomposition consistency. This filtered dataset is then used to train IC-LoRA for the next round. This iterative process of generation, filtering, and retraining progressively enhances the data generation stability.

Self-improving the Cycle-Consistent Model. We apply a similar iterative self-improving methodology to train the cycle-consistent decomposition and composition model (Figure 4(c)). In each iteration, we first generate a large batch of composite images. The cycle model from the previous iteration is then used to perform a full decomposition-recomposition cycle on this new data. Examples that are evaluated as high fidelity are considered pseudo-samples. These samples are retained and incorporated into the training set for each iteration. This self-curation process ensures that the model continuously benefits from increasingly reliable supervision, resulting in demonstrable improvements in decomposition stability, coherence, and generalizability.

4. Experiments

4.1. Comparison Methods

To benchmark our approach, we compare our method against baselines from two categories: asset-focused and instruction-based editing. For asset-focused editing, we compare against AssetDropper [25], in which GroundingDINO [28] is required to locate the logo first. For instruction-based editing, we evaluate IC-Edit [48], Flux-Kontext [23], and Gemini [9].

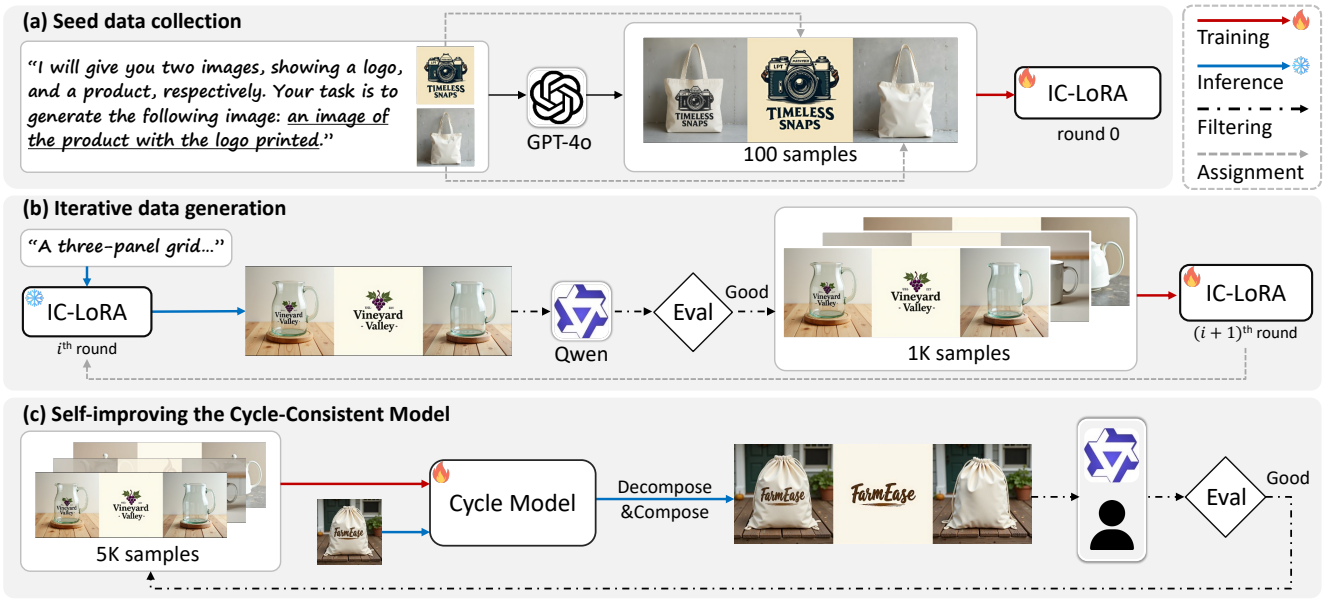

Figure 4. Illustration of progressive data collection. (a) We first collect a seed dataset to obtain an IC-LoRA as the initial data generator. (b) In each round, we select high-quality samples generated by the current LoRA and reintroduce them to the training set. (c) During the training of the cycle model, we use it to produce consistent data by decomposing an image, and then re-composing it. High-quality recomposition samples are added back into the training set.

Table 1. Quantitative comparison using VQAScore and VLMScore, higher is better. To prevent potential evaluation bias, we utilize three VLMs as evaluators (*i.e.*, Qwen, GPT-4o, and Gemini). The **best** and second best results are highlighted.

Method	**VQAScore** ↑		**VLMScore** ↑ (Qwen / GPT4o / Gemini)				
	Logo	Object	Logo Isolation	Logo Consistency	Object Isolation	Object Consistency	Average
AssetDropper [25]	0.42	—	4.49 / **2.87** / **4.84**	3.84 / 4.00 / 3.90	—	—	—
ICEdit [48]	0.31	0.31	1.09 / 2.70 / 1.02	1.56 / 2.58 / 1.00	2.35 / 3.93 / 3.88	3.80 / 3.42 / 3.32	2.55
Kontext [23]	0.40	0.32	3.27 / 2.73 / 3.28	4.17 / 4.04 / 2.80	2.80 / 3.75 / 4.64	4.75 / 4.26 / **4.94**	3.79
Gemini [9]	0.42	**0.32**	4.39 / 2.63 / 4.72	4.05 / 4.36 / 3.72	3.09 / **4.39** / **4.72**	**4.81** / 4.63 / 4.93	4.20
Ours	**0.43**	0.31	**4.50** / 2.77 / 4.76	**4.22** / **4.43** / **3.94**	**3.35** / 4.33 / 4.37	4.70 / 4.62 / 4.67	**4.22**

4.2. Quantitative Results

Quantitative evaluation is conducted on 1.5K synthetic test samples using two metrics: (a) VQAScore [27], which measures text-image alignment. VQAScore is computed independently on the decomposed logo and object; (b) VLMScore, in which we use different VLMs to assess decomposition results on a 1-5 scale across four aspects: Logo Isolation, Logo Consistency, Object Isolation, and Object Consistency. Please check out the supplementary material for more detail on the evaluation.

Results in Table 1 demonstrate that we achieve the highest logo VQAScore and VLMScore, confirming our superiority in separating logos. In contrast, while instruction-based models (Gemini [9], Flux-Kontext [23]) excel at logo removal (high object scores), they struggle to accurately isolate the logo. AssetDropper [25], while specialized for asset extraction, degrades in complex non-linear scenarios and, crucially, fails to recover the underlying object.

4.3. Qualitative Results

We present a qualitative comparison with baseline methods in Figure 5. The figure showcases five challenging scenarios: illumination variation, perspective distortion, logos on non-planar (3D) surfaces, textual elements, and transparent materials. These examples demonstrate the methods' varying abilities to handle non-linear layer interactions. Across all cases, our approach consistently produces cleaner isolated logos and more coherent objects. In contrast, competing methods exhibit noticeable artifacts, incomplete separations, or fail to preserve layer consistency. Figure 6 shows our method's generalization on in-the-wild photographs.

4.4. Ablation Studies

We conduct ablation studies to validate our component design choices, with results in Figure 7. We first establish a baseline by training the decomposition model only on data from the initial Round 0 IC-LoRA generator, which

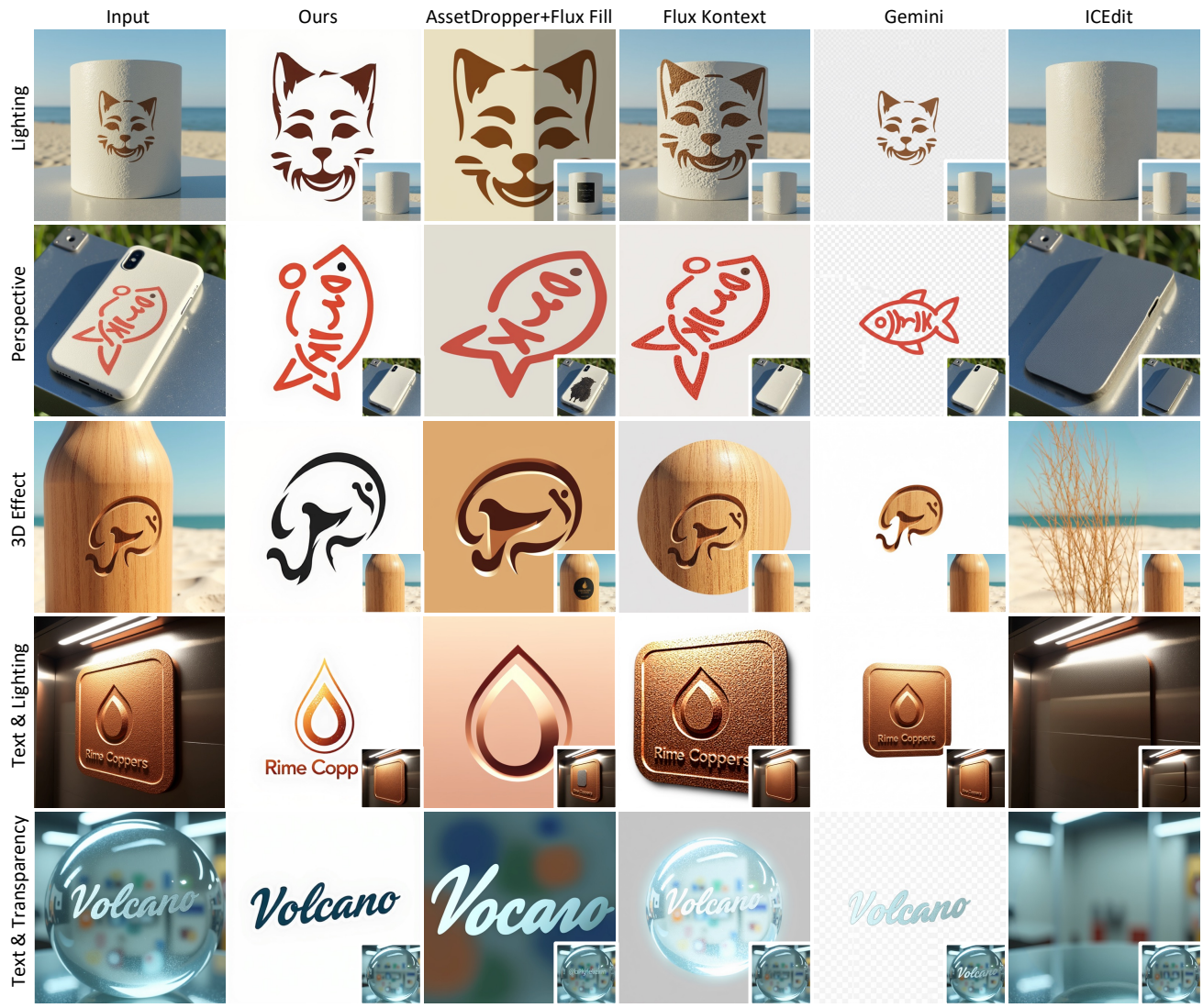

Figure 5. Qualitative comparison on challenging scenarios on synthetic data. The first column shows the inputs, while the following columns present results from our approach and four baselines: AssetDropper [25], Flux-Kontext [23], Gemini [9], and ICEdit [48]. The decomposed object layers appear at the bottom-right of each sample. For AssetDropper [25], we use FLUX-Fill [22] to inpaint the logo region as the object layer. Note that all the synthetic images are generated from a prompt, not composited from a logo and a clean object.

yields poor separations. Introducing Iterative Data Generation measurably improves decomposition, demonstrating the benefit of higher-quality data. Adding Cycle-Consistency provides a significant enhancement in logo fidelity and isolation. Finally, our full model further refines object consistency and realism. The increasing proportion confirms the data generation quality iteratively improves, providing superior supervision for subsequent training.

4.5. User Study

To complement VLM-based evaluation, we conducted a user study consisting of 20 questions with 30 participants. In each question, participants are required to rank four anonymized decomposition results from best to worst according to two criteria: (a) consistency, measuring how accurately the logo is extracted, and (b) perceptual reasonableness, assessing whether the result appears natural without non-linear artifacts.

As illustrated in Figure 10, our method is ranked top-1 in over 50% of the cases. AssetDropper [25], as a task specific logo extraction method, outperforms Flux-Kontext [23] and Gemini [9], which is consistent with its design focus. While Gemini [9] achieves strong VLM scores, it does not show the same advantage in user study, likely due to tendencies such as preserving original size or producing pseudo-transparent outputs, which are less favored by human.

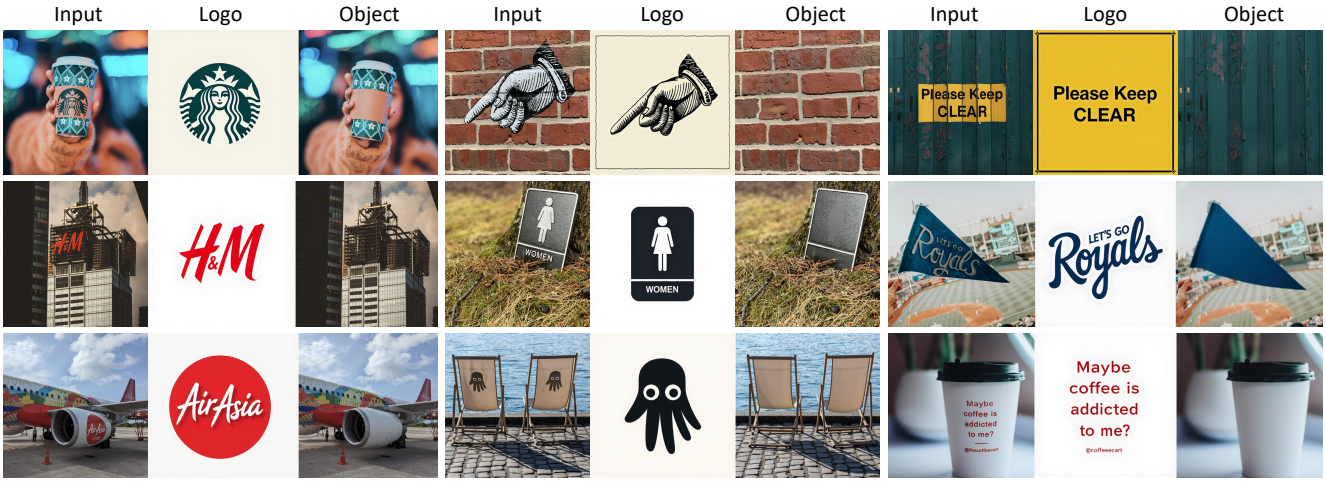

Figure 6. Decomposition results on real photographs. Our approach successfully decomposes both well-known brand logos and uncommon text signs under diverse lighting and perspective conditions, demonstrating strong generalization in real-world environments.

Figure 7. Ablation study. Each row visualizes the decomposition results with progressively adding the proposed components. The decomposed logos and objects become clearer and more consistent from the base model to the full model.

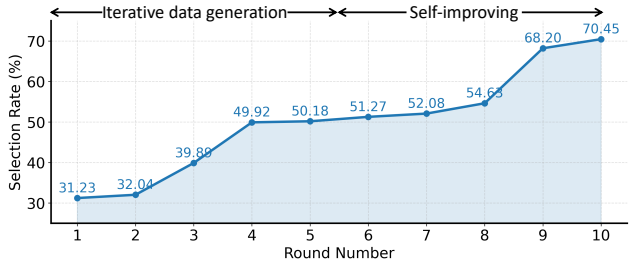

Figure 8. Selection rate of high-quality samples generated by our approach across different data collection rounds. The selection rate increases steadily as the round number grow from 1 to 10.

Table 2. Quantitative comparison on intrinsic decomposition. For the compared methods, we report their results presented by Careaga et al. [6]. Lower values indicate better performance.

Method	Intensity (×100)↓	Chromaticity ↓
Careaga et al. [5]	0.57	6.56
Kocsis et al. [20]	1.13	5.35
Chen et al. [8]	0.98	4.12
Careaga et al. [6]	**0.54**	**3.37**
Ours (w/o Cycle)	0.59	3.65
Ours	0.57	3.54

4.6. Generalization to Other Decomposition Tasks

Intrinsic Decomposition. We evaluate our method on intrinsic decomposition, which factors an image into reflectance (albedo) and shading. Without task-specific priors, we train our cycle model on the Hypersim dataset [32] and evaluate on the MAW dataset [41]. As shown in Table 2, our approach achieves comparable accuracy to SOTA methods specifically tailored for this task. Crucially, an ablation against a variant lacking cycle consistency shows consistent improvement, demonstrating its benefit in stabilizing this decomposition. Qualitative results in Figure 9(a) further verify our framework's effectiveness.

Foreground-background Decomposition. We also test our method on foreground-background decomposition, which aims to disentangle salient objects from their contextual environments. We make a training set of ∼5K triplets by generating composite images (FLUX.1-dev [22]), creating the background layer (Flux-Kontext [23]), and extracting the foreground (GroundingDINO [28] and SAMv2 [31]). After training our cycle model on this dataset, it produces clear decompositions, as shown in Figure 9(b). These results further demonstrate the generalization of our framework to other separation tasks.

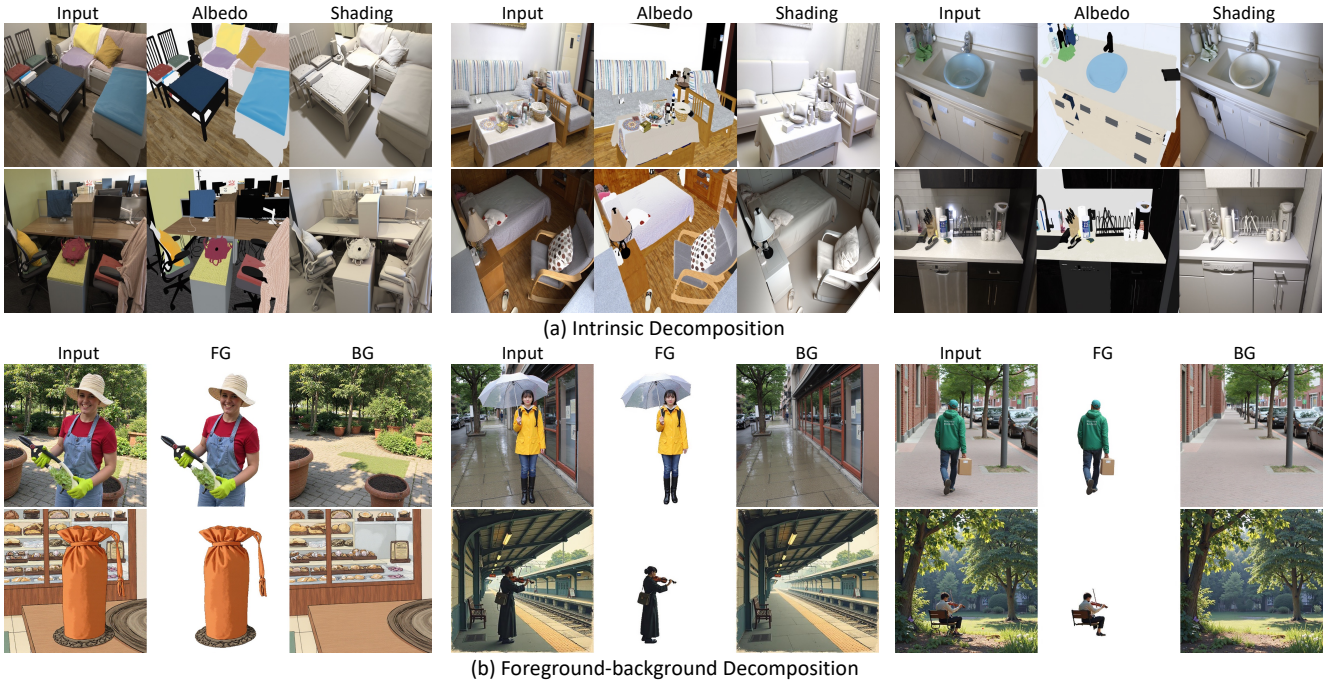

Figure 9. Our decomposition results on (a) intrinsic decomposition, which decomposes an input image into albedo and shading layers, and (b) foreground-background decomposition, which separates a foreground object from its background scene. These qualitative results are supported by quantitative evaluation (e.g., Table 2), confirming the consistency and reliability of our decomposition framework.

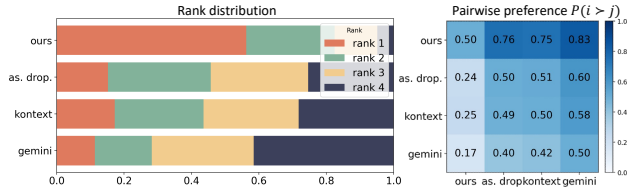

Figure 10. Visualization of user study results. Left: Rank distribution; Right: Pairwise preference heat map.

5. Conclusions, Limitations, and Future Work

In this paper, we revisited the problem of layered image decomposition through the lens of in-context learning. While diffusion-based in-context models have mostly been explored for composition and editing, we showed that the same paradigm can also be extended to the inverse process to decompose or unmix an image into its constituent layers. Central to our framework is a cross-cycle training scheme, where decomposition and composition models are trained jointly to validate each other, enabling the handling of non-linear and globally coupled interactions between layers through one model. This is different from previous work DecompDiffusion [35], which addresses image decomposition by training separate models.

While our method shows strong and consistent performance across a wide range of settings, it is not entirely bulletproof. The model still struggles with certain out-of-domain cases, particularly when the overlaid element dominates the scene, such as very large brand logos on walls or billboards. These cases often fall outside the distribution of training data and remain challenging for current models trained under limited diversity. Another limitation is the handling of multi-layer decomposition. Our current formulation is designed for separating up to two layers and cannot easily generalize to images containing multiple overlaid elements, such as posters with several distinct logos or text layers. This constraint mainly arises from the limited-grid paradigm used for inpainting-based visual in-context learning, a limitation shared with other VICL approaches [3, 48].

Beyond these limitations, the formulation reflects a broader idea that generative models can learn not only to compose but also to disassemble. Treating composition and decomposition as dual, cross-linked processes allows the model to internalize how image layers interact, rather than relying on explicit supervision or handcrafted priors.

Looking ahead, coupling decomposition and composition through mutual supervision opens several future work, potentially extending to motion, illumination, or multimodal information such as audio and 3D structure. Another promising direction is improving the robustness of the framework, where recent advances in robust model adaptation [10, 11] could inspire more resilient decomposition. More broadly, it encourages models that discover structure from weak or implicit supervision, moving toward a unified understanding of visual composition.

Acknowledgements

This work was supported in part by ICFCRT (W2441020), GD S&T Program (2024B0101050004), Shenzhen Science and Technology Program (KJZD20240903100022028, KQTD20210811090044003, RCJC20200714114435012), NSFC (62472288), GD Natural Science Foundation (2026A1515010423), Israel Science Foundation (3441/21, 1473/24, 2203/24), and Scientific Development Funds from Shenzhen University.

References

[1] Lama Affara, Liangliang Nan, Bernard Ghanem, and Peter Wonka. Large scale asset extraction for urban images. In *European Conference on Computer Vision (ECCV)*, pages 437–452, 2016. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv.2502.13923*, 2025. 4

[3] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:25005–25017, 2022. 2, 8

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[5] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics (TOG)*, 43(1):1–24, 2023. 1, 7

[6] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024. 1, 2, 7

[7] Lan Chen, Qi Mao, Yuchao Gu, and Mike Zheng Shou. Edit transfer: Learning image editing via vision in-context relations. *arXiv preprint arXiv.2503.13327*, 7, 2025. 2

[8] Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In *European Conference on Computer Vision (ECCV)*, pages 450–467, 2024. 7

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 4, 5, 6

[10] Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *International Conference on Machine Learning (ICML)*, 2025. 8

[11] Junhao Dong, Raoof Zare Moayedi, Yew-Soon Ong, and Seyed-Mohsen Moosavi-Dezfooli. Allies teach better than enemies: Inverse adversaries for robust knowledge distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026. 8

[12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1107–1128, 2024. 2

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. 3

[14] Diego Gomez, Julien Philip, Adrien Kaiser, and Élie Michel. Rrm: Relightable assets using radiance guided material extraction. In *Computer Graphics International Conference (CGI)*, pages 17–41, 2024. 3

[15] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 1 (2):3, 2022. 2, 3

[18] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv.2410.23775*, 2024. 2, 4

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4

[20] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5198–5208, 2024. 7

[21] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:22199–22213, 2022. 2

[22] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 2, 3, 6, 7

[23] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv.2506.15742*, 2025. 2, 4, 5, 6, 7

[24] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1952–1961, 2023. 2

[25] Lanjiong Li, Guanhua Zhao, Lingting Zhu, Zeyu Cai, Lequan Yu, Jian Zhang, and Zeyu Wang. Assetdropper: Asset extraction via diffusion models with reward-driven optimization. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 2, 3, 4, 5, 6

[26] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. *arXiv preprint arXiv.2504.07960*, 2025. 2

[27] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, pages 366–384, 2024. 5

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pages 38–55, 2024. 4, 7

[29] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021. 2

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. 7

[32] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10912–10922, 2021. 7

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[34] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Conference on Neural Information Processing Systems (NeurIPS)*, 32, 2019. 2

[35] Jocelin Su, Nan Liu, Yanbo Wang, Joshua B Tenenbaum, and Yilun Du. Compositional image decomposition with diffusion models. In *International Conference on Machine Learning (ICML)*, pages 46823–46842. PMLR, 2024. 8

[36] Tomoyuki Suzuki, Kang-Jun Liu, Naoto Inoue, and Kota Yamaguchi. Layerd: Decomposing raster graphic designs into layers. *arXiv preprint arXiv.2509.25134*, 2025. 1, 2

[37] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6830–6839, 2023. 2

[38] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv.2304.03284*, 2023. 2

[39] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Conference on Neural Information Processing Systems (NeurIPS)*, 36: 8542–8562, 2023. 2

[40] Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xiangtai Li, and Yiren Song. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv.2505.21541*, 2025. 1

[41] Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. Measured albedo in the wild: Filling the gap in intrinsics evaluation. In *International Conference on Computational Photography (ICCP)*, pages 1–12, 2023. 7

[42] Jiarui Xu, Yossi Gandelsman, Amir Bar, Jianwei Yang, Jianfeng Gao, Trevor Darrell, and Xiaolong Wang. Improv: Inpainting-based multimodal prompting for computer vision tasks. *Transactions on Machine Learning Research (TMLR)*, 2023. 2

[43] Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7643–7653, 2025. 1

[44] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgbx: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference (SIGGRAPH)*, 2024. 1

[45] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 1, 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2

[47] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Conference on Neural Information Processing Systems (NeurIPS)*, 36: 17773–17794, 2023. 2

[48] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv.2504.20690*, 2025. 2, 4, 5, 6, 8