Decomposition of Graphic Design with Unified Multimodal Model

Hui Nie [†1] **Zhao Zhang** [‡2] **Yutao Cheng** [2] **Maoke Yang** [2]
Gonglei Shi [2] **Qingsong Xie** [3] **Jie Shao** [2] **Xinglong Wu** [2]

Abstract

We propose **L**ayer **D**ecomposition of **G**raphic **D**esigns (**LDGD**), a novel vision task that converts composite graphic design (e.g., posters) into structured representations comprising ordered RGB-A layers and metadata. By transforming visual content into structured data, LDGD facilitates precise image editing and offers significant advantages for digital content creation, management, and reuse. This task presents two core challenges: (1) predicting the attribute information (metadata) of each layer, and (2) recovering the occluded regions within overlapping layers to enable high-fidelity image reconstruction. To address this, we present the **De**compose L**a**yer **M**odel (**DeaM**), a large unified multimodal model that integrates a conjoined visual encoder, a language model, and a condition-aware RGB-A decoder. DeaM adopts a two-stage processing pipeline: first generates layer-specific metadata containing information such as spatial coordinates and quantized encodings, and then reconstructs pixel-accurate layer images using a condition-aware RGB-A decoder. Beyond full decomposition, the model supports interactive decomposition via textual or point-based prompts. Extensive experiments demonstrate the effectiveness of the proposed method. The code is accessed at https://github.com/witnessai/DeaM.

1. Introduction

In the era of digital media production, multi-layer composite images serve as a fundamental structure in visual design, enabling the conveyance of rich and complex information through the integration of visual elements and text. In digital image editing, layers are employed to isolate different components of an image, functioning analogously to transparent sheets that can be stacked to achieve various visual effects or construct new compositions. Understanding the underlying layer structure of such images is essential for a wide range of applications, including content editing, material archiving, and image reconstruction. However, despite significant progress in computer vision and computer graphics, automatically decomposing a composite image into its constituent layers remains a challenging and largely unsolved problem.

This paper introduces **L**ayer **D**ecomposition of **G**raphic **D**esigns (**LDGD**), as shown in Figure 1. The goal is to decompose a composite image into a set of semantically meaningful and individually discernible layers with a well-defined order—akin to peeling back the layers of an onion. Each layer, whether it represents the background, primary imagery, atmospheric elements, text, or other visual components, is annotated with structured metadata, including attributes such as position, size, and text color. The primary challenge of this task lies in the intricate visual overlap and interdependence among layers, which makes single-purpose vision models such as standard image segmentation inadequate for achieving accurate and structured layer decomposition.

We introduce **De**compose L**a**yer **M**odel (**DeaM**), a purpose-built framework for this challenging task, inspired by the success of large unified multimodal models in understanding complex visual patterns and generating multimodal outputs. The DeaM incorporates three critical components: a conjoined visual encoder, a comprehensive large language model, and a sophisticated condition-aware RGB-A decoder. The conjoined visual encoder is designed to perceive and integrate visual information across different semantic hierarchies. The large language model then processes this visual information along with textual instructions to execute layer decomposition tasks. Following this, the condition-aware RGB-A decoder takes the tokens associated with the image layers derived from this decomposition, producing the relevant visual outputs. DeaM operates by producing metadata for each dissected layer, represented as a JSON format. In Figure 2, DeaM demonstrates the ability to forecast a wide variety of details for the textual layers, such as font

[†]Work done during internship at ByteDance. [‡]Equal Contribution & Project Lead. [1]University of Chinese Academy of Sciences. [2]ByteDance Intelligent Creation, China. [3]OPPO AI Center. Correspondence to: Zhao Zhang <zzhang@mail.nankai.edu.cn>, Jie Shao <shaojie.mail@bytedance.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

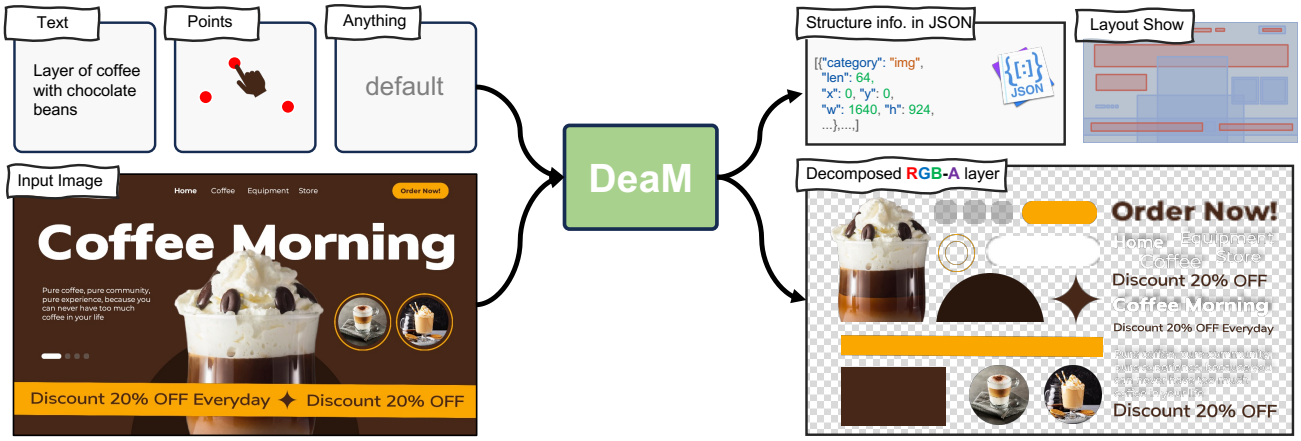

Figure 1. We introduce a new task, Layer Decomposition of Graphic Designs (LDGD), which involves decomposing a composite image into multiple layers (including the background, primary imagery, atmosphere graphics, text, etc.) in a logical layer order. To address this task, we develop a decomposition model, DeaM, which supports both full-layer decomposition and selective layer extraction guided by diverse forms of input, including textual prompts and point-based instructions.

style, color palette, textual content, alignment, and so forth, thereby facilitating text layer reconstruction and secondary edits. For the image layers, DeaM predicts an additional field, capturing the visual tokens' indices essential for the decoder to regenerate RGB-A images.

In addition to fully automated layer decomposition, DeaM supports interactive prompts, including both text instructions and point-based inputs, allowing users to selectively guide the decomposition of specific layers.

To support the LDGD task, we construct an in-house dataset named CreatiLD, comprising over 200,000 poster images annotated with rich information necessary for layer decomposition. To further enhance the model's capabilities, we leverage the linguistic proficiency of GPT-4 to generate open-ended instruction-tuning data, thereby broadening the model's understanding of diverse user queries.

Our contributions are multifold:

- We formalize the novel and practical task, LDGD, which aims to decompose a composite image into individual and complete RGB-A layers with metadata, even if the layers are obstructed by each other.
- We present the DeaM, a large unified multimodal model that performs end-to-end decomposition of an image. DeaM also accommodates user interactions, allowing for the targeted decomposition of layers specified through text or point-based prompts.
- Our DeaM achieves consistently superior results compared to the baseline and a variety of existing approaches, highlighting its effectiveness in the challenging task of layer decomposition.

2. Related Work

2.1. Layered Image Decomposition

Earlier work (Monnier et al., 2021; Sbai et al., 2018; Du et al., 2023) explored similar tasks of decomposing images into layers by learning object prototypes along with parameters for occlusion and transformation, combining them to reconstruct complete images. While these studies have made notable progress, their problem formulations fall short of capturing the complexity and diversity inherent in real-world scenarios. To address these limitations, we extend their task definition and propose a new formulation that is more aligned with practical applications.

Our task setting differs in several key aspects: (1) Our input images contain text, which is typically not regarded as an object. We explicitly decompose the text as a separate visual layer. (2) Our output includes not only the image of each individual layer but also associated metadata for each layer.

Overall, our formulation represents a unified understanding and generation task, making it inherently more challenging than previous approaches.

In addition, existing derendering methods (Ma et al., 2022; Rodriguez et al., 2025) typically convert images into SVG format. However, SVG representations struggle to capture complex visual details, and these methods are generally limited to decomposing simple graphical content while lacking the ability to accurately decompose the text.

2.2. Large Multimodal Models

Over the recent year, there has been a notable surge in interest surrounding large language models (LLMs) (Devlin

et al., 2019; Radford et al., 2019; Raffel et al., 2020; Ouyang et al., 2022; Zhang et al., 2022; Zeng et al., 2023; Chiang et al., 2023) in the realms of natural language processing and computer vision. This surge is attributable to the superior capabilities of LLMs, which have demonstrated excellence in multifaceted applications, notably their comprehensively detailed global knowledge base and multifunctional utilities. Large multimodal models (LMMs) (Team et al., 2023; Zhang et al., 2024a; OpenAI, 2023; Wang et al., 2023; Alayrac et al., 2022; Li et al., 2025; Gong et al., 2023; Huang et al., 2023; Dai et al., 2023; Zhu et al., 2024; Driess et al., 2023; Ye et al., 2023; Bai et al., 2023; Wang et al., 2024a) integrate the powerful linguistic capabilities of LLMs and extends its multimodal processing abilities, breaking down barriers between visual and linguistic modalities, and even more beyond. Two notable contributions in vision-language learning are BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2023). BLIP-2 integrates a frozen image encoder with large language models using a lightweight Q-Former for efficient cross-modal alignment. LLaVA pioneers the use of GPT-4 (OpenAI, 2023) to generate instruction tuning data for multimodal tasks. In multimodal localization, Shikra (Chen et al., 2023) excels at spatial reasoning, while Graphist (Cheng et al., 2025) extends localization to poster layout design, showcasing strong structural understanding. Recently, emerging unified multimodal models (UMM) (Zheng et al., 2023; Zhan et al., 2024; Sun et al., 2024; Zhou et al., 2025; Xie et al., 2025; Wu et al., 2024c; Wang et al., 2024b; Fang et al., 2024; Wu et al., 2024a; Ma et al., 2024; Qu et al., 2024; Wu et al., 2024b; Li et al., 2024; Shi et al., 2024; Tong et al., 2024) have demonstrated the ability to jointly generate text and images, enabling unified multimodal generation. Inspired by these advances, we explore leveraging a UMM for layer decomposition of graphic designs, aiming to simultaneously generate textual elements and multiple visual layers.

2.3. Image Editing

Although contemporary studies have not yet concentrated on the concurrent decomposing of multiple image layers, a subset of prevalent image generation and editing techniques demonstrates partial capability in isolating and decomposing individual layers. Instruction-based image editing: InstructPix2Pix (Brooks et al., 2023), HIVE (Zhang et al., 2024b), MGIE (Fu et al., 2024), SmartEdit (Huang et al., 2024), and SEED-X (Ge et al., 2024) can sometimes achieve layer decomposition effects by facilitating editing of designated layers. However, these methods are unable to well decompose multiple elements of an image. Several inpaiting methods (Liu et al., 2024; Lugmayr et al., 2022; Yang et al., 2023b; Corneanu et al., 2024) can modify the specified regions in image with extra mask condition. These methods only sample the masked regions from a pre-trained diffusion model, while keeping the unmasked areas the same in each denoising step. Another kinds of inpainting methods (Xie et al., 2023a;b; Yang et al., 2023a; Yu et al., 2023) fine-tune a specially designed image inpainting model to integrate corrupted image and mask. However, these methods cannot decompose multiple layers simultaneously and struggle to handle cases where layers occlude each other.

3. Task Formulation

Given a composite image $\mathbf{x} \in \mathcal{X}$ that includes various image elements (such as the background, primary imagery, decorations, and text), the goal of **L**ayer **D**ecomposition of **G**raphic **D**esigns (**LDGD**) is to decompose the image into an **ordered** series of RGB-A layers $\mathcal{I} = \{I_i \in \mathbb{R}^{h_i \times w_i \times 4}\}_{i=1}^{n}$ with their metadata (such as position, text color and so on) from the original image. We hope that each layer is as close as possible to the corresponding part in the original image and that re-rendering these layers in accordance with their metadata will yield the original picture. For the sake of simplicity, we treat text elements as RGB-A images, because once our model recognizes the text, color, font, and size, we can render them into RGB-A images. The position information for each layer contains five numerical elements $(x_i, y_i, w_i, h_i, l_i)$, where x_i and y_i denote the coordinates of the upper left corner of the layer in relation to the original image, w_i and h_i represent the width and height of the layer within the original image, and l_i indicates the layer order. If the ordering of the layers is illogical, then re-rendering the image using the obtained layers would result in implausible occlusions (Cheng et al., 2025).

4. Baseline

Layer decomposition of graphic designs is a novel and underexplored task for which no existing methods provide a satisfactory solution. For example, traditional image segmentation can separate foreground objects but fails to recover occluded background content, and often treats text as non-editable image regions. Similarly, image inpainting is typically limited to generating a single-layered output, such as simply erasing text, without structural decomposition.

To create a basic approach to this task, we develop a baseline combining image inpainting, matting, and OCR. We first extract text data, font attributes, and bounding boxes using OCR, then form a mask for inpainting. Image matting separates the primary imagery, whose mask aids in further inpainting to restore the background via the Volcano Engine API. This method **can only decompose the input image into three layers: background, primary imagery, and text**. It cannot decompose the image into more layers.

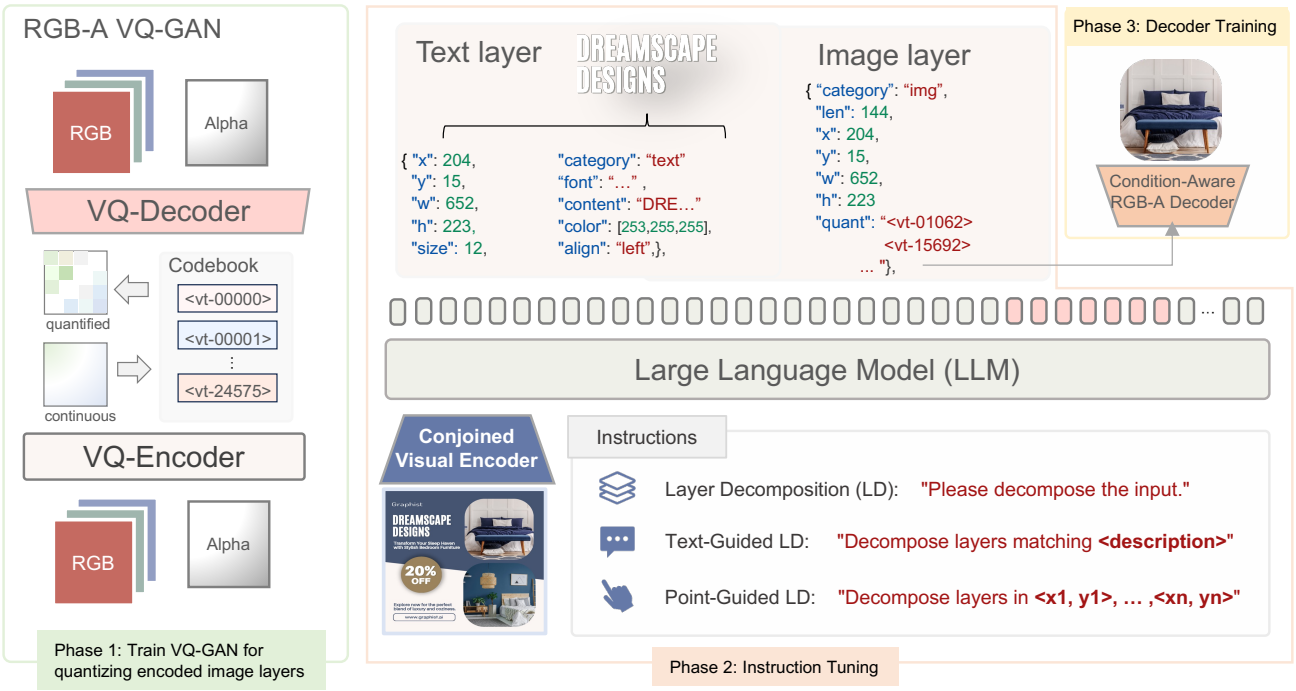

Figure 2. **DeaM Pipeline.** The training process of DeaM consists of three phases: (1) **VQ-GAN Training**: Train an RGB-A VQ-GAN, then use it to encode all image layers with "quant" annotations. (2) **Instruction Tuning**: Train DeaM to generate detailed image descriptions (metadata), performing layer decomposition tasks including element positions, layer order, indices in the VQ-GAN codebook, and so on. (3) **Decoder Training**: Train a condition-aware RGB-A decoder using ResNet as the condition encoder, keeping the VQ encoder frozen. This phase is decoupled from the instruction tuning.

5. Proposed DeaM

As shown in Figure 2, **De**compose **La**yer **M**odel (**DeaM**) comprises three key components: a conjoined visual encoder, a large language model backbone, and a condition-aware RGB-A decoder. The conjoined visual encoder is capable of encoding semantically multi-layered visual information. The large language model processes the input visual and textual instruction information and generates layer decomposition results. The condition-aware RGB-A decoder then decodes the tokens corresponding to the image layers in the decomposition results to produce the respective images.

5.1. RGB-A VQ-GAN.

We modify the RGB VQ-GAN into an RGB-A VQ-GAN by changing the number of channels in the convolutional kernels of the first and last layers from 3 to 4 to accommodate the alpha channel. When using the default training strategy (Esser et al., 2021), there is ambiguity in the model's encoding of RGB-A images. For instance, the model produces the same encoding for some black lines that are similar but not identical. We found that this issue arises because the model treats the reconstruction of the alpha channel and the RGB channels as equally important during training, which is not the case in reality. The RGB information is actually more important. To eliminate the model's encoding ambiguity, we adjusted the VQ-GAN reconstruction loss by applying different weights to the RGB and the alpha channel. The training data of RGB-A VQ-GAN includes image layers from all poster images, the ImageNet training set (Deng et al., 2009), and a subset of several million images from LAION (Schuhmann et al., 2021; 2022).

5.2. Conjoined Visual Encoder.

Visual images contain visual elements with different levels of semantic information, such as object categories with pronounced semantic information (e.g., people, watches) and decorative elements with less semantic information (e.g., geometric element patterns). Existing LMMs (Liu et al., 2023) typically employ CLIP as the sole visual encoder, encoding generally higher-level semantic information. To enable the model to focus on both high-level semantic object categories and lower-level visual elements, we design a conjoined visual encoder. And we utilize two different types of visual encoders: the CLIP Vision Encoder (Radford et al., 2021), and DINO v2 (Oquab et al., 2024). After encoding, we concatenate the visual features along the channel dimen-

sion, thereby obtaining visual features with dimensions of $N \times (1024 \times 2)$. Both visual encoders here use an input resolution of 336×336 to ensure a consistent number of tokens after encoding.

5.3. Large Language Model Backbone.

Vocabulary Expansion. In the architectural design of our model, a crucial component involves the utilization of semantic tokens transformed by the VQ-GAN model (Esser et al., 2021). This architecture integrates an encoder and a decoder, augmented by a specialized quantization layer whose role is to translate image information into a sequence of tokens predefined in a codebook. Both the encoder and the decoder adopt convolutional layer structures to effectively process images of varying resolutions. Specifically, the encoder compresses the spatial dimensions of the input image through a series of downsampling operations, while the decoder reconstructs the image back to its original size via matching upsampling operations. To purely represent the image information of each layer and avoid the generated layer images containing meaningless background information, the model training utilizes images with four channels, namely RGB-A. That is, areas not associated with this layer will be transparent. This RGB-A image format (Zhang & Agrawala, 2024) can support the re-rendering of decomposed layers and facilitate a range of subsequent applications such as more sophisticated image editing and creation. We trained the VQ-GAN model with a downsampling ratio of $f = 16$. Higher image resolutions lead to clearer reconstructions, but they also make the training sequences of the model much longer and significantly increase computational cost. Given the limitations on the length of the model's output tokens, when VQ-GAN generates tokens for each image layer, we set the input resolution for semantically rich natural images to 192×192 and for semantically sparse decorative elements to 128×128. In the end, we obtain either 144 or 64 tokens. In our method, the VQ-GAN encoder is employed to encode the **image layers** (i.e., not text layers), and the obtained codebook IDs serve as the ground truth for training. Subsequently, we incorporate an equivalent number of special tokens like '<vt-{number}>' to the LLM, reflecting the size of the VQ-GAN's codebook. The **number** represents the index of the VQ-GAN codebook embedding. In this work, we use InternLM2-7B (Team, 2023) as the LLM backbone.

Enhancing Prediction Regularity. We have discovered that predicting the output image VQ-GAN codebook index '<vt-{number}>' can be lengthy (for instance, semantically rich images or graphic elements may have 144 tokens, while semantically sparse decorative elements might have only 64). This poses a substantial challenge for unified multimodal models, which struggle to accurately infer the precise number of indices in a single forward pass. Without the correct number of indices, these index sequences cannot be reshaped into a two-dimensional matrix (12×12 or 8×8), and thus, we would be unable to decode and generate images using the decoder. To address this issue, when constructing training image token data for the layer, after encoding with VQ encoder to get a two-dimensional matrix of indices, we append a learnable newline token ('\n') at the end of each row of indices before flattening the sequence. This approach introduces more regularity to the sequence, thereby reducing the prediction difficulty for the unified multimodal model.

5.4. Condition-Aware RGB-A Decoder

After obtaining the output of the decomposing results, for image layers containing the image tokens '<vt-{number}>', we reshape them into a 12×12 or 8×8 two-dimensional array based on the length of the image tokens. We can retrieve the corresponding index's VQ-GAN codebook embedding and use the VQ decoder to decode and restore them back into images. However, we found that directly using the pretrained VQ decoder to decode the generated images results in less clarity. To further enhance the quality of decoded images, we designed a Condition-Aware RGB-A Decoder, as shown in Figure 3. Motivated by the observation that the input images are of high clarity, but not utilized during decoding, we aim for the decoder to leverage the input images as conditional information to enhance image generation quality as much as possible. The decoder takes two types of inputs: one is the original layers, and the other is the corresponding regions of the input images associated with these layers. We noted that the unobscured parts of the layer images are relatively easy to restore, while the obscured parts are more challenging. To focus the model's attention on the obscured regions, we force the model to learn the mask of the obscured areas. Simultaneously, to reduce the impact of unnecessary foreground information in the occluded regions on the final reconstruction quality, we multiply the predicted mask with the features of the conditional information, thereby erasing the information of the occluded areas. Finally, these modified data are concatenated with the latent features to serve as the input for the next part of the network. The training loss for the decoder is the same as that for VQ-GAN(Esser et al., 2021). During inference, we use the features indexed by the codebook from the special token predicted by the unified multimodal model as the features encoded by the VQ-Encoder, combined with the images of the crop areas corresponding to the modified layer as input.

5.5. Training Strategy

The training process of DeaM is divided into three phases: VQ-GAN training, instruction tuning, and decoder training. In the **first phase**, we train an RGB-A VQ-GAN. After the training is complete, we use this VQ-GAN to encode all im-

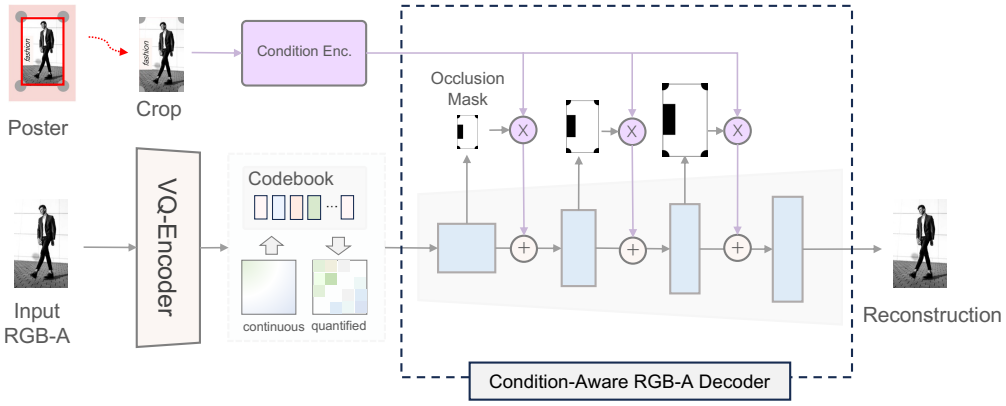

Figure 3. **Condition-Aware RGB-A Decoder.** During the training, the decoder utilizes the VQ-GAN decoder for initialization and accepts images corresponding to the poster areas of the layers as conditional information to further enhance the image generation quality. To ensure that the decoder concentrates on generating the occluded regions, the model is compelled to predict the mask of the occluded areas. During inference, the features encoded by the VQ-Encoder are replaced with the quantized encodings predicted by the DeaM.

age layers, adding annotations for the "quant" field. In the **second phase**, we focus on training DeaM to generate detailed descriptions of images (metadata), that is, to perform the layer decomposition task, which includes the positions, layer order, corresponding indices of image elements in the VQ-GAN codebook and so on. By providing the model with specific task instructions and images, DeaM learns to autoregressively generate the JSON structure information for each layer, encompassing all the details necessary to render the different image elements. We also include text-guided, point-guided, and GPT-4 generated instruction-tuning data for training. In the **third phase**, we train a condition-aware RGB-A decoder. Here, we use ResNet as the condition encoder and keep the VQ encoder frozen during training, only training the decoder and condition encoder. Additionally, the training of the decoder module is decoupled from the training of the unified multimodal model. We use 16 NVIDIA A800 GPUs for training.

6. Training Datasets

6.1. Human-Annotated Data

We curate a large-scale in-house dataset of over 200,000 multi-layer structured poster images, containing layer sequences, category labels, and coordinate information. This dataset, named CreatiLD, is collected from the internet and comprises 224,054 graphic designs with complete layer-level annotations. The majority of samples are posters covering domains such as holiday events, retail, dining, and corporate communications. On average, each sample contains 10.30 layers, with an approximate image-to-text ratio of 6.3:3.7. To support the LDGD task, we enrich the dataset with additional annotations. For image layers, a captioning model is employed to generate descriptive captions, and color statistics are extracted. For text layers, we apply Optical Character Recognition (OCR) to extract textual content, use a font classification model to identify typefaces, and compute font sizes based on the layer coordinates. During training, each input consists of a single-layer poster image, and the model is trained to output detailed metadata and a structured representation for each corresponding layer.

6.2. Instruction Data

The construction of instruction data aimed to enable the DeaM to adhere to user instructions and accurately decompose one or more layers in alignment with user requirements. We contemplate how to design user instructions with greater freedom to support a broader range of application scenarios. We crafted two types of instruction-tuning data with considerable flexibility, and in addition, we employed GPT-4 to collect a large corpus of open instruction-tuning data.

Text-Guided Instruction-Tuning Data. We curated a set of instruction data for each layer endowed with a caption. Adhering to the specified format, each instruction read: "Please decompose the layer that manifests as follows: [caption].". The '[caption]' was replaced with a description pertinent to a specific layer, which the model was then tasked to decompose. Ultimately, we collected 200,000 annotated data entries.

Point-Guided Instruction-Tuning Data. To facilitate a wider range of interactive modalities, we have also gathered instruction-tuning data based on point interactions. Users input points at various locations on an image, and the model then interprets which layer corresponds to each specified point. Expanding upon this, our approach allows for a more intuitive user experience, enabling individuals to engage with the visual content directly and receive immediate,

context-aware feedback from the model. In the end, we collected 100,000 annotated data entries.

GPT4-Generated Data. Drawing inspiration from the LLaVA (Liu et al., 2023) and Shikra (Chen et al., 2023) models, we employed the GPT-4 model to generate a more versatile collection of instructions, catered to address open-ended user inquiries. By inputting the JSON format information (including information on the types and positions of each layer) of posters into GPT-4, the model autonomously devised five question-answer pairs relevant to each poster's content. Model-generated questions were constrained to those answerable with the metadata, ensuring the identification of related layers was possible. We collected 15,000 question-answer pairs through GPT-4 for the fine-tuning of the model.

7. Experiments

7.1. Test Datasets

To facilitate a more open and transparent comparison with other methods, we utilize a publicly available academic dataset Crello for evaluation.

Crello dataset[1]: Crello (Yamaguchi, 2021) is now referred to as VistaCreate[2], provides a collection of visual designs originating from an online design tool. This comprehensive compilation encompasses a wide variety of design types, tailored for multiple uses such as infographics for social platforms, digital advertising banners, headers for blogs, and templates for print posters. Within this dataset, each individual design is accompanied by intricate details specifying the order of layers, the precise location of each element within the space, and the classifications of the various design components. The test set of this dataset contains over 2,000 images.

7.2. Evaluation Metrics

In order to systematically evaluate the effectiveness of layer decomposition, we have designed a comprehensive set of evaluation metrics. We focus primarily on two aspects: (1) the quality of the image reconstruction, (2) the prediction accuracy of the layer position (that is, the planar position of the layer on the original image and the order of the layers). For assessing the quality of layer reconstruction, we refer to the evaluation methods used for image reconstruction quality. We primarily utilize the image reconstruction quality metric FID (Fréchet Inception Distance). Motivated by the premise that if the method excels in decomposing all layers with high fidelity, the composite image generated by merging all the layers would exhibit a high degree of similarity to the original image. As for the prediction accuracy of bounding boxes and the accuracy of layer order prediction, we draw on the Hungarian matching algorithm used in object detection (Carion et al., 2020) to view the problem as one of predicting the accuracy between two sets of bounding boxes. We denote the ground truth set of objects as b, and the collection of N predictions as $\hat{b}_i = \{\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i\}$. In order to establish a bipartite matching between these two groups, we look for a permutation of $\alpha \in \mathcal{P}_N$ items that yield the minimal cost:

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{P}_N} \sum_{i=1}^{N} \mathcal{L}_{box}(b, \hat{b}_{\alpha(i)}), \tag{1}$$

where we define $\mathcal{L}_{box}(\cdot,\cdot)$ as $\mathcal{L}_{iou}(b, \hat{b}_{\alpha(i)})$, $\mathcal{L}_{iou}$ is IoU loss (Zhou et al., 2019) in this work. We use the mean IoU (Intersection over Union) of the matched bounding boxes as a metric Loc_α to measure the accuracy of the bounding box prediction. Considering that there might be various ways to decompose some decorative elements in an image, such as two lines that could be decomposed either as two separate layers or as a single layer, it is inconvenient to perform box calculations. Therefore, we only focus on the box detection capabilities of the core elements (such as primary imagery).

Con. Vis. Enc.	Enh. Pre. Reg.	Con. Dec.	FID↓
			105.524
✓			101.132
✓	✓		95.488
✓	✓	✓	70.629

Table 1. **Ablation Study.** Here, Con. Vis. Enc. denotes the conjoined visual encoder, Enh. Pre. Reg. stands for enhancing prediction regularity, and Con. Dec. refers to the condition-aware RGB-A encoder.

method	FID↓	$\text{Loc}_{\hat{\alpha}}$ ↑
Baseline	99.603	0.7069
DeaM(ours)	70.629	0.7128

Table 2. Comparison with the baseline.

7.3. Quantitative Results

Ablation Study. We perform ablation studies on key components and strategies of our model design, with results summarized in Table 1. Using image reconstruction quality as the evaluation metric, we observe consistent performance gains contributed by each component. Notably, the condition-aware RGB-A decoder yields the most substantial improvement, primarily by significantly enhancing image clarity. Additionally, the conjoined visual encoder strengthens the model's detection capability and reduces missed detections, further boosting reconstruction quality.

[1] https://huggingface.co/datasets/cyberagent/crello

[2] https://create.vista.com/

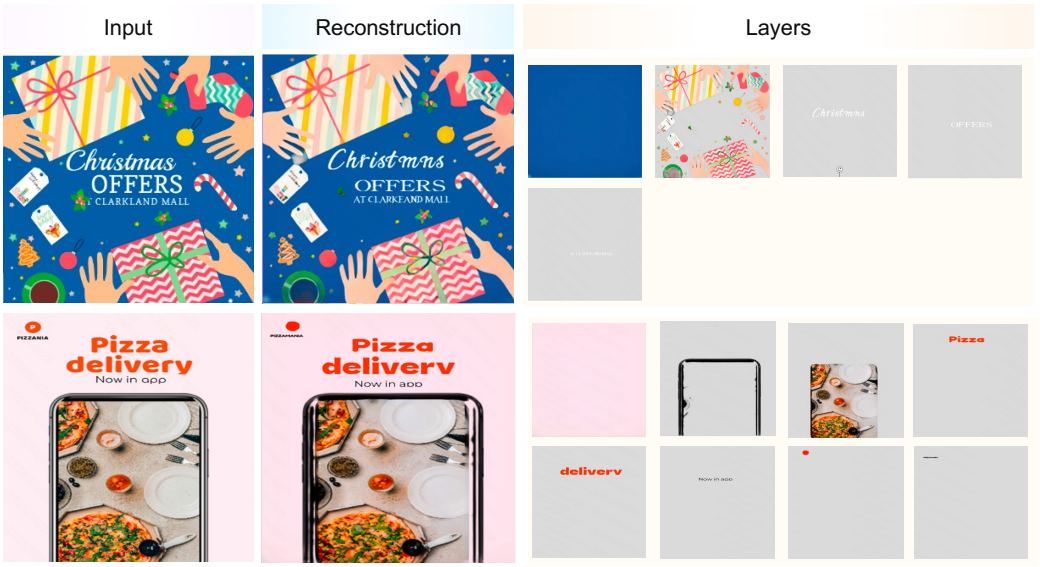

Figure 4. Layer decomposition results of DeaM on the Crello dataset.

Comparison of Baseline. As shown in Table 2, in terms of image reconstruction quality, our method significantly outperforms the baseline. Moreover, unlike the baseline which only parses images into **three layers** (background, main subject, and text), our method can parse to more layers, enhancing its capability.

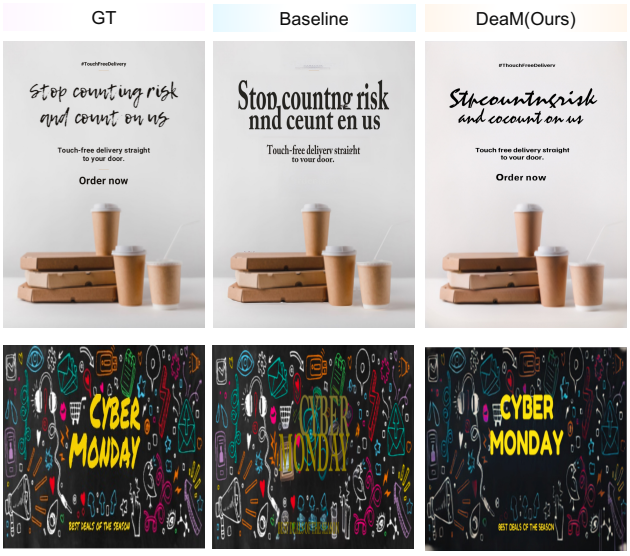

Figure 5. Comparison with the reconstruction results of the baseline.

7.4. Qualitative Results

Layer Decomposition. We present the results of layer decomposition and composition as shown in Figure 4, indicating that the reconstructed image is close to the original image in terms of image quality. For each row in the figure, the leftmost image is the input, the middle column shows the reconstructed image based on the predicted layer order, and the rightmost section displays the decomposed layers of the input. These layers are arranged from left to right in hierarchical order, where layers further to the left represent lower levels in the visual hierarchy, with the leftmost layer typically corresponding to the background. From these figures, it can be seen that the bounding box predictions for each layer are quite accurate, and the text content is largely correct, although the accuracy of the font prediction is not exemplary. The example at the top of the Figure 4 is quite interesting: when there are many decorative elements and their spatial relationships lack a clear hierarchical order, the model tends to predict them as being on the same layer. Moreover, there is still room for improvement in our image decoding and generation quality, which calls for a more powerful image tokenizer.

Comparison of Baseline. As illustrated in Figure 5, we compare the reconstruction quality between the baseline and our proposed DeaM. DeaM exhibits strong performance in reconstructing text regions, effectively recovering key details such as content, font, size, and color in most cases. In contrast to the baseline's inpainting-based approach, which often introduces artifacts or unnatural patterns, DeaM substantially reduces these issues, resulting in cleaner and more faithful reconstructions.

Text-Guided and Point-Guided Layer Decomposition. As shown in Figure 6, we showcase the capabilities of our model through the lens of point-guided and text-guided layer decomposition, revealing its versatility and robust under-

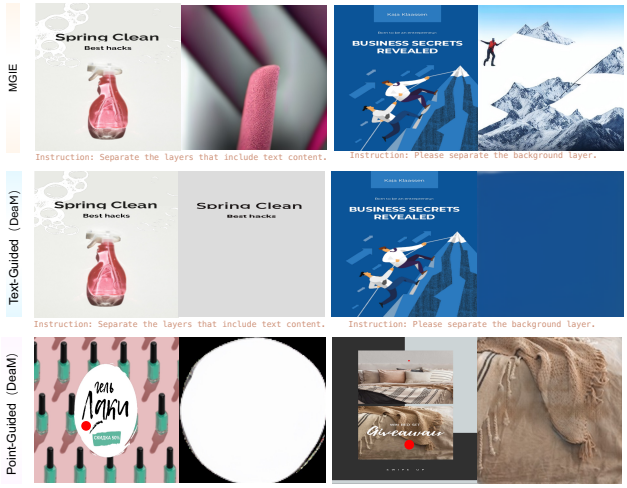

Figure 6. Results of point-guided and text-guided layer decomposition. For each column, the left image represents the input picture, and the right image shows the results decomposed by the model.

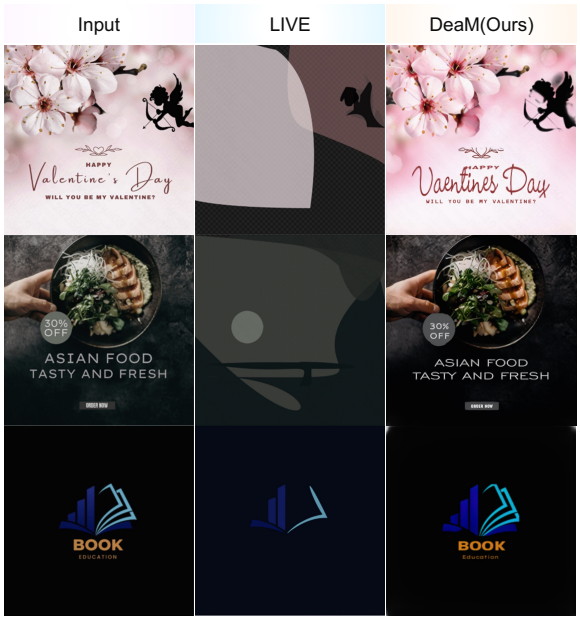

Figure 7. Comparison with LIVE. The first column shows the input images, while the second and third columns present the reconstruction results from LIVE and our method DeaM, respectively.

standing of complex image structures. The first row presents the inference results of the instruction-based image editing method MGIE. In the second row, corresponding textual prompts for text-guided layer decomposition are shown beneath each image. The third row visualizes the prompt points used for point-guided layer decomposition on the input images, indicated by red dots. The text-guided layer decomposition outcomes illustrate the model's adeptness in differentiating between various layer types. It does not merely recognize the presence of text or imagery; it understands their distinct roles within a composite image. This distinction is crucial for applications that demand selective editing or targeted adjustments within an image. The illustrative examples provided, particularly the image in the third row, not only validate the model's precision in pinpointing locations denoted by user-specified points but also highlight its predictive prowess in envisioning the resulting image once certain layers are omitted. This is evident in the model's ability to infer what the image would look like without the textual elements that overlay the picture, a task that requires a nuanced understanding of both spatial relationships and the hierarchy of visual elements.

Comparison with Image Vectorization Method. We compare our approach with an image vectorization method, LIVE (Ma et al., 2022), which generates compact SVG representations featuring layer-wise structures that align semantically with human perception. As shown in Figure 7, LIVE tends to perform well on simple logo images but struggles with more complex natural images.

8. Conclusion

In summary, we introduce the novel task of layer decomposition of graphic design, and present DeaM, a dedicated framework that achieves significant progress in automatic layer decomposition, fine-grained image understanding, and manipulation. Leveraging a novel RGB-A VQ-GAN, DeaM effectively decomposes composite images into distinct layers, substantially improving both encoding and decoding efficiency—crucial for high-fidelity image editing. The model's decomposing capability is further enhanced by the integration of multiple carefully designed components and strategies. DeaM also supports a more user-friendly interaction paradigm, allowing control via both point-based and text-based instructions. In addition, we construct a new dataset specifically for the layer decomposition task, which facilitates future research in this area. Extensive experiments demonstrate the effectiveness of DeaM, highlighting its potential to inspire a wide range of computer vision applications.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.

Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:23716–23736, 2022.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Cheng, Y., Zhang, Z., Yang, M., Nie, H., Li, C., Wu, X., and Shao, J. Graphic design with large multimodal model. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2473–2481, 2025.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Corneanu, C., Gadde, R., and Martinez, A. M. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 4334–4343, 2024.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. C. H. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P. Palm-e: An embodied multimodal language model. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 202, pp. 8469–8488, 2023.

Du, Z., Kang, L., Tan, J., Gingold, Y. I., and Xu, K. Image vectorization and editing via linear gradient layer decomposition. *ACM Trans. Graph.*, 42(4):97:1–97:13, 2023.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, 2021.

Fang, R., Duan, C., Wang, K., Li, H., Tian, H., Zeng, X., Zhao, R., Dai, J., Li, H., and Liu, X. Puma: Empowering unified mllm with multi-granular visual generation. *arXiv preprint arXiv:2410.13861*, 2024.

Fu, T., Hu, W., Du, X., Wang, W. Y., Yang, Y., and Gan, Z. Guiding instruction-based image editing via multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., and Shan, Y. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., and Chen, K. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, N. J. B., Chaudhary, V., Som, S., Song, X., and Wei, F. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., and Shan, Y. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371, 2024.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Cahyono, J. A., Yang, J., Li, C., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2025.

Li, H., Tian, C., Shao, J., Zhu, X., Wang, Z., Zhu, J., Dou, W., Wang, X., Li, H., Lu, L., et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

Liu, A., Niepert, M., and den Broeck, G. V. Image inpainting via tractable steering of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Gool, L. V. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11451–11461, 2022.

Ma, X., Zhou, Y., Xu, X., Sun, B., Filev, V., Orlov, N., Fu, Y., and Shi, H. Towards layer-wise image vectorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16293–16302, 2022.

Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Zhao, L., et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

Monnier, T., Vincent, E., Ponce, J., and Aubry, M. Unsupervised layered image decomposition into object prototypes. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 8640–8650, 2021.

OpenAI. Gpt-4v(ision) system card, 2023. URL https://openai.com/research/gpt-4v-system-card.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D. K., Yuan, Z., and Wu, X. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Rodriguez, J. A., Puri, A., Agarwal, S., Laradji, I. H., Rajeswar, S., Vázquez, D., Pal, C., and Pedersoli, M. Starvector: Generating scalable vector graphics code from images and text. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pp. 29691–29693, 2025.

Sbai, O., Couprie, C., and Aubry, M. Vector image generation by learning parametric layer decomposition. *arXiv preprint arXiv:1812.05484*, 2018.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022.

Shi, W., Han, X., Zhou, C., Liang, W., Lin, X. V., Zettlemoyer, L., and Yu, L. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Emu: Generative pretraining in multimodality. In *International Conference on Learning Representations (ICLR)*, 2024.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Team, I. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., and Liu, Z. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., and Dai, J. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Chen, K., Xu, B., Li, J., Dong, Y., Ding, M., and Tang, J. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024a.

Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., and Bai, X. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024b.

Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024c.

Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q., Gu, Y., Chen, Z., Yang, Z., and Shou, M. Z. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Xie, S., Zhang, Z., Lin, Z., Hinz, T., and Zhang, K. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22428–22437, 2023a.

Xie, S., Zhao, Y., Xiao, Z., Chan, K. C., Li, Y., Xu, Y., Zhang, K., and Hou, T. Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. *arXiv preprint arXiv:2312.03771*, 2023b.

Yamaguchi, K. Canvasvae: Learning to generate vector graphic documents. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 5481–5489, 2021.

Yang, S., Chen, X., and Liao, J. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *ACM International Conference on Multimedia (ACM MM)*, pp. 3190–3199, 2023a.

Yang, S., Zhang, L., Ma, L., Liu, Y., Fu, J., and He, Y. Magicremover: Tuning-free text-guided image inpainting with diffusion models. *arXiv preprint arXiv:2310.02848*, 2023b.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., and Chen, Z. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y., and Tang, J. GLM-130B: an open bilingual pre-trained model. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., et al. Anygpt: Unified multimodal llm with discrete sequence modeling. In *Annual Meeting of the Association for Computational Linguistics.*, 2024.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., and Yu, D. Mm-llms: Recent advances in multimodal large language models. In *Annual Meeting of the Association for Computational Linguistics.*, pp. 12401–12430, 2024a.

Zhang, L. and Agrawala, M. Transparent image layer diffusion using latent transparency. *ACM Trans. Graph.*, 43(4):1–15, 2024.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.-C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9026–9036, 2024b.

Zheng, K., He, X., and Wang, X. E. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.

Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., and Levy, O. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., and Yang, R. Iou loss for 2d/3d object detection. In *International Conference on 3D Vision (3DV)*, pp. 85–94, 2019.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*, 2024.