COMPASS: Grounding Composition-Intent Guidance in Unified Multimodal Models

Ziqi Zhou[1,2], Weize Quan[2,3]*, Mining Tan[2,3], Zhihan Chen[2,3], Dandan Zheng[4], Jingdong Chen[4], Jun Zhou[4], Weiming Dong[2,3], and Dong-Ming Yan[2,3]

[1] University of Edinburgh, United Kingdom
[2] MAIS, Institute of Automation, Chinese Academy of Sciences, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[4] Ant Group, China

Abstract. Composition is a high-level visual intent that governs where subjects are placed and how a scene is organized, yet current unified multimodal models remain unreliable at fine-grained composition recognition and struggle to turn such intent into controllable generation. We present COMPASS, the first unified multimodal framework that grounds composition-intent control in a single system spanning both composition perception and composition-guided generation, with a shared expert token τ_c as the central intent anchor. On the perception side, COMPASS injects composition expertise into an MoE backbone in a minimally invasive manner and distills the inferred intent into τ_c. On the generation side, COMPASS reuses τ_c as a global conditioning signal that steers the denoising trajectory, effectively converting passive composition analysis into explicit layout control. To support systematic instruction-following composition learning and evaluation at scale, we construct COMP-11, a large-scale dataset with an 11-class taxonomy and reasoning-augmented annotations. Extensive experiments show that COMPASS substantially improves category-level composition understanding and delivers more composition-consistent, prompt-faithful generation than strong baselines. The code and dataset for this work will be released here.

1 Introduction

Visual composition is the grammar of photography, providing the structural scaffold upon which aesthetic and emotional narratives are built. In the field of aesthetic intelligence, this necessitates two core capabilities working in tandem: expert-level perception, the ability to decode the intentional geometric organization of a scene, and layout-guided creation, the ability to synthesize new content that respects a specific compositional schema. Recently, Large Multimodal Models (LMMs) [5,9,11,28,29,32,42] have trended towards a unified paradigm, aiming to integrate visual understanding and generation into a single, cohesive framework. However, while these models excel at general reasoning, they struggle to represent composition as an explicit, controllable structure. Existing aesthetic LMMs [3, 4, 17, 43] primarily emphasize passive assessment and critique, treating composition as one attribute among many rather than a structured object

* Corresponding author: *qweizework@gmail.com*

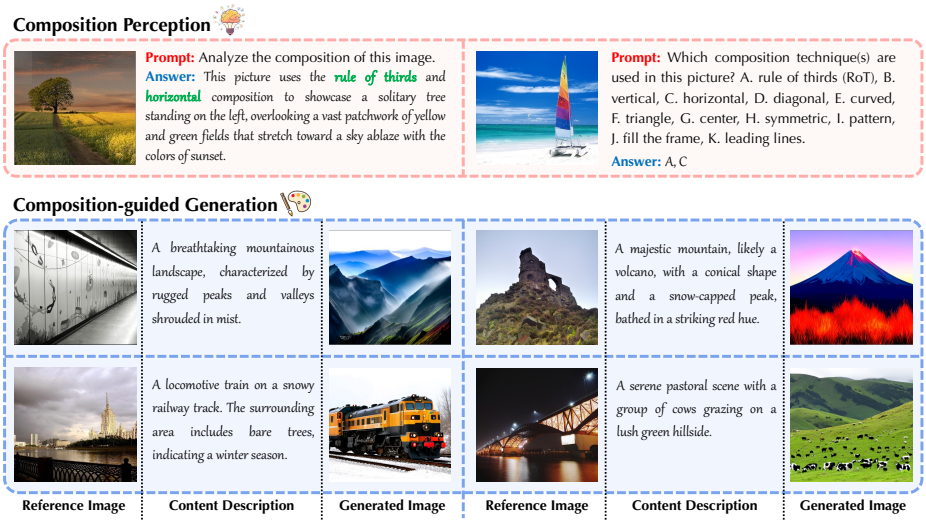

Fig. 1: Performance of the proposed COMPASS: the unified system handles various composition understanding tasks, and further uses a reference image as layout intent to synthesize new content that remains faithful to the text prompt.

that can be parsed and acted upon. Meanwhile, the broader literature on photographic composition has proposed rule-based taxonomies and datasets [19,39,41], which underscore that composition understanding is a distinct capability and remains underexplored for general LMMs.

In this work, we make the first attempt to unify expert composition perception and reference-guided generation in a cohesive framework. On the generation side, we target a challenging setting: given a reference image that conveys compositional intent and a text prompt for new semantic content, we seek to generate an image that inherits the reference layout style while changing the content. This pursuit is motivated by the inherent limitation of text-only prompting—language efficiently describes semantics but is a low-bandwidth channel for precise spatial geometry. In contrast, a reference photo provides a high-bandwidth spatial blueprint. Related controllable generation methods often rely on explicit low-level controls such as edges, depth, or bounding boxes [20,38,40], which provide strong spatial constraints but do not directly encode high-level composition rules. Closer in spirit, image-prompt editing and style-transfer methods [8,27,36] also condition on a reference image and aim to transfer a specific factor (e.g., style) to a new image; however, they often suffer from content leakage from the reference ones. In our setting, the target factor—composition—is even more entangled with image content, making layout transfer without semantic contamination substantially harder.

Crucially, reference-guided compositional generation presupposes expert perception; the model must infer the structural intent of a reference before it can re-

produce it. However, achieving this dual intelligence in a unified model raises two domain-specific hurdles. On the perception side, the challenge is twofold. First, there is a severe scarcity of systematic compositional data: existing aesthetic datasets for LMM [3,4,17,18,43] often treat composition as a coarse-grained attribute within a broad aesthetic analysis, lacking a comprehensive and consistent taxonomy to cover the diverse layout patterns used in professional photography. Second, injecting such specialized expertise into a pre-trained LMM through full-parameter fine-tuning or task-agnostic adapters like LoRA [16] can easily bias or overwrite general reasoning unless the expertise is explicitly anchored and routed. On the generation side, the task faces an entanglement trap where models struggle to disentangle the high-level compositional intent from low-level content features. This is especially difficult because paired data—images sharing the same layout but different semantics—rarely exists at scale, making conventional supervised disentanglement infeasible.

To address these challenges, we introduce COMPASS, a framework that empowers unified LMMs with specialized compositional intelligence through an expert-anchor paradigm. To overcome the scarcity of systematic data, we first build Comp-11, a comprehensive composition-focused instruction dataset. Inspired by prior composition datasets [19, 39] and recent benchmarks highlighting LMMs' compositional limitations [41], Comp-11 is designed specifically for instruction-following and actionable composition understanding. In terms of architecture, we propose a Composition-specific Mixture-of-Experts (C-MoE) strategy to achieve expert-level composition perception. Leveraging an existing MoE-based backbone [11], we instantiate a dedicated set of new expert Multilayer Perceptrons (MLPs) and optimize only these parameters during training. This approach infuses the model with compositional expertise while preserving its foundational multimodal competence. Complementing this architecture, we introduce a single learnable expert token τ_c as a prefix to the model's response. Supervised by a hard composition classification objective, τ_c serves as an explicit intent anchor that activates specialized knowledge and surfaces compositional decisions.

To break entanglement in generation without requiring paired data, we pioneer a self-supervised structural bottlenecking strategy. We apply explicit physical decoupling via pixelization and grayscaling to reference images, creating a structural invariant that suppresses semantic appearance while preserving global layout cues. Beyond the input-side bottleneck, we introduce model-side mechanisms to suppress reference leakage, including a customized attention mask that prevents learnable queries from attending to perturbed image tokens and a cross-attention refinement to ensure text faithfulness. Reusing the expert token τ_c as a global conditioning signal further bridges the gap between passive composition analysis and active, layout-controllable generation within a single unified system.

Our main contributions are summarized as follows: (1) We present the first systematic study of unified LMMs for compositional intelligence. (2) We build Comp-11, the first large-scale comprehensive composition-focused instruction dataset with an 11-class taxonomy to enable actionable understanding and rea-

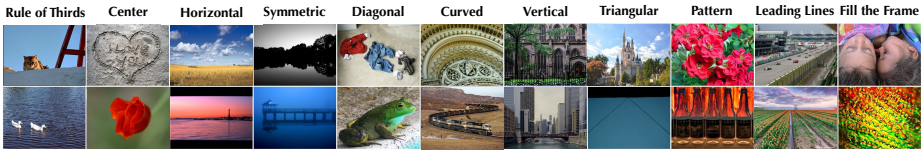

Fig. 2: Visual examples of our COMP-11 taxonomy.

soning. (3) We propose COMPASS, featuring a learnable prefix token τ_c as a shared intent anchor for both understanding and generation tasks, and a self-supervised structural bottlenecking framework that enables reference-guided generation without the need for paired training images.

2 Related Work

Composition understanding and optimization. Photographic composition has primarily been studied from two complementary perspectives: composition understanding and composition optimization. Together, these lines of research provide essential foundations for controllable composition generation.

The first line of research focuses on composition understanding, aiming to model and quantify structural layouts. Early approaches typically evaluated composition quality through regression-based models. By utilizing datasets with dedicated composition-aware annotations, such as CADB [39], these methods isolated spatial arrangement patterns rather than predicting holistic aesthetic scores. Progressing toward a more structured formulation, subsequent studies treated composition as a rule-based classification task. For instance, KU-PCP [19] categorizes images based on predefined compositional rules while explicitly localizing key geometric elements. Together, these works established quantitative and interpretable paradigms for composition recognition.

Building upon composition evaluation, composition optimization focuses on active enhancement, primarily through image cropping. Existing methods are largely supported by two types of datasets: densely annotated datasets with multiple candidate crops (e.g., SACD [35], UGCrop5K [24]) and sparsely annotated datasets providing a single optimal crop (e.g., FCDB [7]). Recent advances incorporate generative modeling to improve robustness, as exemplified by GenCrop [15], which leverages diffusion-based augmentation. Multimodal large models have also been introduced into composition tasks; for instance, PhotoFramer [37] employs multimodal instruction tuning to generate composition guidance alongside exemplar images.

Despite these advances, existing optimization methods—particularly cropping—are inherently restricted to being post-processing steps for existing images. In contrast, our work transcends this limitation by providing real-time guidance, directly generating photographs with ideal compositions.

LMM aesthetic perception. More recently, multi-dimensional aesthetic evaluation has increasingly incorporated Large Multimodal Models (LMMs) to en-

hance semantic reasoning and interpretability [1]. Representative works—such as AesExpert [17], UNIAA [43], AesBench [18], and UniPercept [3]—simultaneously generate aesthetic scores and natural language explanations. This approach facilitates unified reasoning across various aesthetic attributes, including color, lighting, subject emphasis, and composition.

However, within this paradigm, composition is primarily treated as a descriptive attribute embedded within a holistic aesthetic assessment. Even when compositional aspects are explicitly evaluated [4,17], they are subsumed into the broader semantic reasoning process rather than being modeled as explicit, controllable composition categories. Consequently, existing LMM-based aesthetic perception frameworks remain predominantly evaluative, falling short of directly supporting composition-conditioned generation.

Unified multimodal models. Unified multimodal models seeks to integrate perception and generation within a shared architecture. Existing approaches can be broadly grouped into three paradigms. Modular coordination frameworks, such as NExT-GPT [30], employ a large language model to orchestrate external modality-specific generators. While flexible, these designs rely on loosely coupled components with limited cross-modal fusion. Token-level fusion models, including SEED-X [10] and Chameleon [5], move toward tighter integration by jointly modeling visual and textual tokens within autoregressive transformers, enabling unified understanding and generation. However, purely autoregressive modeling often struggles with high-fidelity image synthesis. To address this limitation, hybrid unified frameworks incorporate diffusion mechanisms into the modeling pipeline. Transfusion [42] integrates diffusion into unified training to balance semantic controllability and synthesis quality. Subsequent models, such as Show-o [32], Emu3 [28], and the Janus series [6, 22, 29], further extend unified architectures to support multi-task perception and high-quality generation. More recent efforts emphasize scalability and system-level refinement. VILA-U [31] improves multimodal alignment efficiency for large-scale training, while BAGEL [9] and Ming-Lite-Uni [11] explore scalable and lightweight unified modeling strategies.

Despite architectural unification, existing models primarily optimize modality alignment, generation quality, and training scalability [34]. Fine-grained structural factors—such as photographic composition—are rarely treated as explicit, controllable modeling targets. Consequently, while unified multimodal architectures provide a strong backbone for integrated perception and generation, they lack dedicated mechanisms for structure-aware and composition-conditioned synthesis.

3 Dataset

We introduce Comp-11, a large-scale dataset for composition perception and generation in unified LMMs. Comp-11 is built upon an 11-class composition taxonomy and further augmented with instruction-following and reasoning-oriented multimodal annotations.

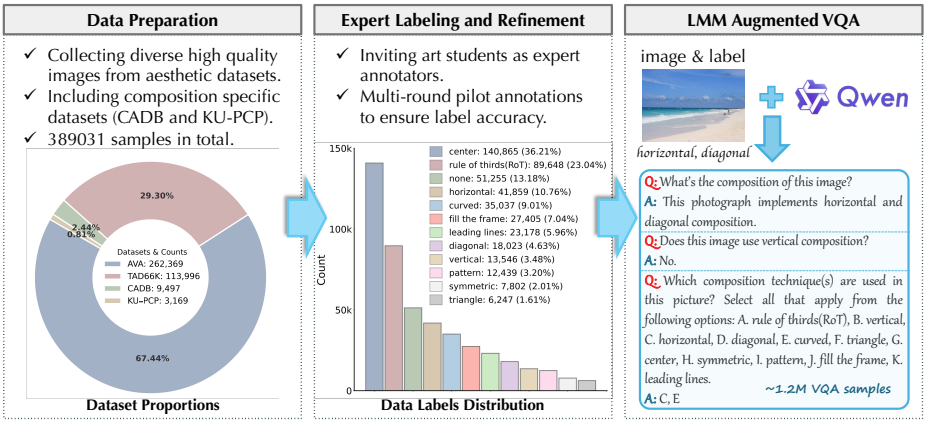

Fig. 3: Data construction pipeline of COMP-11.

Taxonomy of Composition. We consolidate composition concepts from existing composition classification datasets KU-PCP [19] and CADB [39] into a unified taxonomy with C=11 commonly used categories, which we adopt as the label set of our dataset. Specifically, the taxonomy includes Rule of Thirds (RoT), Center, Horizontal, Symmetric, Diagonal, Curved, Vertical, Triangle, Pattern, Leading Lines, and Fill the Frame. We provide qualitative visualization examples for all 11 categories in Figure 2. Precise operational definitions and annotation guidelines are deferred to the supplementary material.

Data Construction Pipeline Building upon our 11-class taxonomy, we develop a rigorous pipeline comprising three stages: data preparation, expert labeling and refinement, and LMM augmentation (Figure 3).

Data preparation. We aggregate 389,031 images from four public aesthetic benchmarks: AVA [23], TAD66K [12], CADB [39], and KU-PCP [19]. Among them, CADB and KU-PCP provide composition labels that can be directly reused, but their scale is insufficient for training a unified LMM. To build a large and systematic training set, we further incorporate high-quality images from AVA and TAD66K as unlabeled aesthetic data and annotate them with our 11-category taxonomy in the next stage. The proportion of images contributed by each source can be seen from the left side of Figure 3.

Expert labeling and refinement. To ensure expert-level annotation quality, we recruited students from photography and fine arts programs and refined labels in two rounds: (i) verifying and correcting inherited annotations from source datasets, and (ii) exhaustively annotating newly collected data under our 11-class taxonomy. To mitigate individual bias, each image is independently reviewed by multiple annotators and disagreements are resolved by consensus voting, yielding a unified and reliable label space. The middle part of Figure 3 further indicates a long-tailed, highly imbalanced label distribution that naturally arises in real-world composition patterns. Notably, we also allow a none-label assignment when

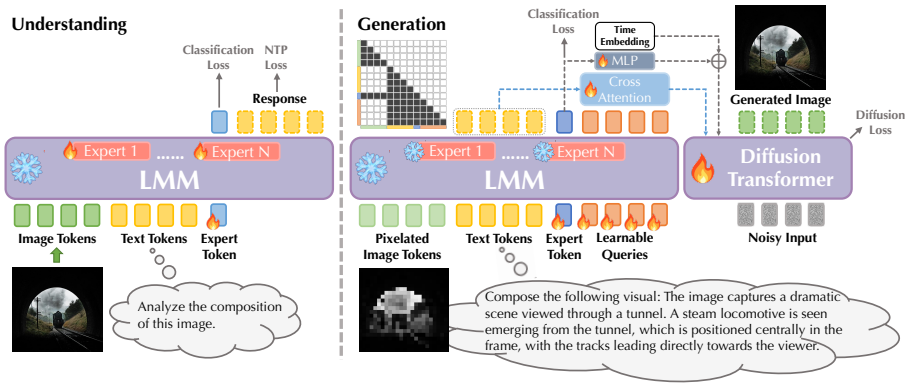

Fig. 4: Overview of COMPASS. The model first performs composition perception with N new experts (C-MoE) and an expert token τ_c to summarize layout cues. For composition-guided generation, a structurally bottlenecked (pixelated) reference is used to suppress appearance leakage, while a customized attention mask (demonstrated under the bold "Generation", where the dark squares represent "allow to attend", while the white ones indicate the opposite) and a cross-attention refinement improve controllability and text faithfulness. The expert token representation is further injected into the diffusion denoiser as a global conditioning signal to steer the generation toward the intended composition.

an image is visually cluttered or its composition is indiscernible even for experts, avoiding forced noisy labels and keeping supervision reliable.

LMM augmentation. Using the expert-verified labels as gold-standard prompts, we leverage a strong LMM [25] to generate reasoning-enhanced metadata, including (i) question answering on compositional rationale, (ii) multiple-choice layout discrimination, and (iii) judgment tasks. In total, this augmentation yields on the order of 1.2M VQA-style pairs, substantially scaling up composition supervision for unified LMM training.

4 Methodology

With the dataset and supervision described in Section 3, we study compositional intelligence in unified large multimodal models through two tightly coupled tasks. In composition perception, given an input image I and an instruction, the model produces an instruction-following response and predicts a multi-label composition vector $y \in \{0,1\}^C$ (with $C = 11$ in our setting). In reference-guided compositional generation, given a reference image I^{ref} that provides layout intent and a text prompt t that specifies new semantic content, we aim to sample an output image I^{gen} that preserves the reference composition while remaining faithful to the prompt:

$$I^{\text{gen}} \sim \mathcal{G}\big(I^{\text{ref}}, t\big), \quad \text{s.t. } \text{Comp}(I^{\text{gen}}) \approx \text{Comp}(I^{\text{ref}}),\ \text{Sem}(I^{\text{gen}}) \approx t. \tag{1}$$

To support these objectives within a single system, we introduce a minimally invasive perception expertization mechanism (C-MoE with a learnable expert token) and a self-supervised generation design that avoids paired same-layout-different-content supervision. We describe the perception branch, the generation branch, and our training strategy as follows.

4.1 Composition Perception

Compositional cues are subtle, geometry-centric, and poorly aligned with generic pretraining objectives, making naïve fine-tuning prone to either underfitting or overwriting general multimodal competence. To this end, we design a perception branch with two complementary components: C-MoE, which injects composition expertise into the MoE backbone in a minimally invasive manner, and a learnable expert token τ_c, which serves as an explicit intent anchor whose representation can be directly supervised for composition prediction.

Composition expertization via C-MoE. Our original unified LMM backbone [11] adopts MoE-based feed-forward networks (FFNs) in a subset of transformer blocks. Given latent token features $h \in \mathbb{R}^d$, the router first produces logits $z = g(h) \in \mathbb{R}^N$ over N experts and selects the index set $\mathcal{K} = \mathrm{TopK}(z)$ of the K largest entries in z (Top-K expert selection). The combining weights are then computed by a softmax restricted to $\mathcal{K}$:

$$\alpha_k(h) = \begin{cases} \dfrac{\exp(z_k)}{\sum_{j\in\mathcal{K}} \exp(z_j)}, & k \in \mathcal{K}, \\ 0, & k \notin \mathcal{K}, \end{cases} \tag{2}$$

and the MoE-FFN output is a weighted sum of the selected experts:

$$\mathrm{MoE}(h) = \sum_{k=1}^{N} \alpha_k(h)\, E_k(h) = \sum_{k\in\mathcal{K}} \alpha_k(h)\, E_k(h), \tag{3}$$

where $E_k(\cdot)$ denotes the k-th expert MLP.

To inject composition expertise without overwriting the foundation capabilities, we propose C-MoE (Composition-specific Mixture-of-Experts), which allocates a new MoE-FFN branch in each MoE block. Concretely, for every sparse FFN layer we instantiate an additional MoE module, named $\mathrm{MoE}^{\mathrm{comp}}(\cdot)$, with its own router and expert parameters, in parallel to the original backbone MoE, noted as $\mathrm{MoE}^{\mathrm{base}}(\cdot)$.

At runtime, we switch between the two branches using a task flag $s \in \{0, 1\}$:

$$\mathrm{FFN}(h) = (1 - s)\,\mathrm{MoE}^{\mathrm{base}}(h) \;+\; s\,\mathrm{MoE}^{\mathrm{comp}}(h), \tag{4}$$

where $s = 1$ activates the composition expert branch. Importantly, $\mathrm{MoE}^{\mathrm{comp}}$ may adopt a different routing policy (router type) from the original backbone, enabling task-specific routing while keeping the backbone intact.

During training, we freeze the original backbone parameters and update only the newly introduced parameters in MoE^{comp} (its router and expert MLPs) to add compositional knowledge with minimal interference to the model's general multimodal abilities.

Learnable expert token as an intent anchor. To turn composition from an implicit aesthetic attribute into an explicit and actionable decision in unified LMM, we introduce a single learnable expert token τ_c as an intent anchor for composition understanding. Concretely, given a dialogue context $\mathbf{x}$ formatted by a chat template, the assistant turn is constructed as $\mathbf{x} \,\|\, \textit{prefix}_{ans} \,\|\, \tau_c \,\|\, \mathbf{s}$, where $\textit{prefix}_{ans}$ denotes the fixed answer prefix produced by the template (e.g., role/format tokens) and $\mathbf{s} = s_{1:T}$ is the response with T tokens. We insert τ_c right after $\textit{prefix}_{ans}$ and before the first response token, so that it functions as the answer-leading token and concentrates compositional cues into a dedicated representation for downstream supervision.

We supervise τ_c with a composition classification objective. Let $H \in \mathbb{R}^{L\times d}$ be the final hidden states of the unified LMM for an input sequence of length L. We extract the hidden representation at the expert-token position, denoted as $h_c \in \mathbb{R}^d$, and apply a lightweight projection head $\phi(\cdot)$ and a classifier W to obtain logits $o \in \mathbb{R}^C$ over C composition classes:

$$o = W\,\phi(h_c). \tag{5}$$

Since an image may exhibit multiple composition techniques, we formulate composition prediction as multi-label classification and optimize a weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = \sum_{i=1}^{C} \left(w_i^{+}\, y_i \log \sigma(o_i) + w_i^{-}\, (1-y_i) \log(1-\sigma(o_i)) \right), \tag{6}$$

where $y \in \{0,1\}^C$ is the multi-hot label vector, $\sigma(\cdot)$ is the sigmoid function, and (w_i^{+}, w_i^{-}) are per-class weights to counter class imbalance (computed from the label distribution; see Section 3). In practice, we optionally apply a pooling operator when multiple τ_c instances appear (e.g., in multi-round dialogues), but Eq. (6) remains unchanged.

4.2 Composition-guided Generation

Physical decoupling via structural bottlenecking. In reference-guided compositional generation, the reference image only provides layout intent but not content. Ideally, one would learn this factorized transfer from paired supervision—images sharing the same composition but different semantics—yet such paired data is scarce in practice. Motivated by common self-supervised learning principles, we instead create a training signal by transforming each sample into a structural proxy that approximates "same layout, different appearance" without explicit pairs.

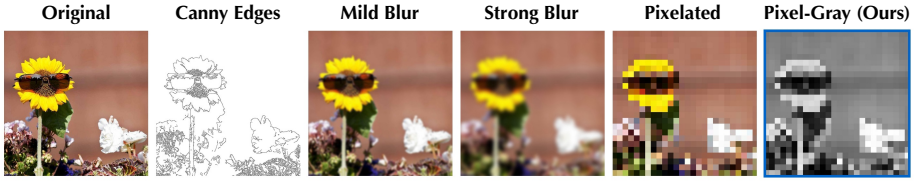

Fig. 5: Comparison of structural bottlenecks for reference-guided generation. Pixelated grayscale references suppress semantic identifiability by disrupting contour-level cues while preserving coarse spatial layout, leading to better layout transfer with reduced content leakage compared to edge- or blur-based alternatives.

Concretely, we introduce a structural bottleneck before the model: we transform the reference image I^{ref} into a grayscale, pixelated image $\tilde{I}$ and condition generation on $\tilde{I}$ instead of the raw reference. Grayscaling suppresses color as a strong content cue, while pixelization aggressively removes recognizable details yet preserves coarse spatial massing and dominant layout lines. Alternative structuralizations do not reliably eliminate semantics. Edge-based representations preserve fine contours that directly encode object identity, while Gaussian blurring mainly smooths textures but often retains coherent silhouettes, allowing viewers (and models) to infer plausible semantics from shape cues even when details are unclear. Pixelization offers a better compromise: it disrupts contour-level recognizability and thus weakens semantic inference, yet largely preserves coarse spatial layout and region massing. Accordingly, we adopt grayscale pixelization as the structural bottleneck for reference-guided generation (Figure 5).

Customized attention mask for leakage suppression. On top of the above input-side bottleneck, we further introduce model-side mechanisms to prevent reference content leakage and to ensure text faithfulness during generation. Following existing unified model architectures [11,26,29], when processing $\tilde{I}$, our unified LMM forms an autoregressive token sequence that contains (i) perturbed image tokens $\mathbf{v} = v_{1:M}$ extracted from $\tilde{I}$, (ii) text tokens $\mathbf{t} = t_{1:L}$, and (iii) a set of learnable queries $\mathbf{q} = q_{1:Q}$ used as the downstream generation interface. Although $\tilde{I}$ already removes most semantic details, allowing text tokens or learnable queries to attend to $\mathbf{v}$ can still create shortcut copying. We therefore use a customized causal mask $A \in \{0,1\}^{N\times N}$ (N is the model length) such that

$$A_{ij} = \mathbb{1}[j \le i] \cdot \mathbb{1}\big[(i \notin \mathcal{T} \cup \mathcal{Q} \setminus \{\tau_c\}) \ \vee \ (j \notin \mathcal{I})\big], \tag{7}$$

where $\mathcal{I}$ indexes the reference-image token span, $\mathcal{Q}$ indexes the learnable-query span, and $\mathcal{T}$ indexes all text-token positions after the reference image. Thus, standard causality ($j \le i$) is preserved, while any position $i \in (\mathcal{T} \cup \mathcal{Q}) \setminus \{\tau_c\}$ is forbidden to attend to reference-image tokens $j \in \mathcal{I}$, leaving τ_c as the only post-image token that may read $\mathbf{v}$ for composition-related summarization. We visualize the allowed and blocked regions of A in Figure 4 (under the bold "Generation") for more intuitive demonstration.

Cross-attention refinement for text faithfulness. Reference-guided generation is prone to content leakage unless the model is simultaneously discouraged from copying the reference and encouraged to follow the prompt [33]. Our customized attention mask implements the former by preventing post-image tokens and learnable queries from attending to reference-image tokens. To complement this, we add a lightweight cross-attention refinement module that directly ties the learnable queries to the prompt representation, strengthening text faithfulness in the diffusion conditioning stream. Concretely, after the LMM forward pass, let $H \in \mathbb{R}^{N\times d}$ be the resulting hidden states. We take the subsequences for learnable queries and prompt tokens, denoted as $H_{\text{queries}} = \text{LMM}(\mathbf{q}) \in \mathbb{R}^{Q\times d}$ and $H_{\text{prompt}} = \text{LMM}(\mathbf{t}) \in \mathbb{R}^{L\times d}$, and apply

$$\hat{\mathbf{q}} = \text{CrossAttn}(H_{\text{queries}}, H_{\text{prompt}}), \tag{8}$$

where H_{queries} acts as queries and H_{prompt} provides keys/values. This process anchors the learnable queries to the prompt representation and thereby improving text content faithfulness.

Expert token as a global conditioning signal. Finally, we reuse the expert token τ_c as a layout capsule that bridges composition perception and controllable generation. In our generation setting, the reference-image pathway is deliberately weakened, so the model cannot reliably exploit residual appearance cues from reference tokens; instead, transferable layout information is encouraged to be summarized into the representation at τ_c. Formally, let $h_c \in \mathbb{R}^d$ be the final hidden state at τ_c, and let $c = \psi(h_c)$ be a projection into the diffusion conditioning space. We inject c as a global bias via timestep modulation, $\tilde{e}_t = e_t + c$, where e_t is the standard timestep embedding. The modulated embedding $\tilde{e}_t$ is broadcast within the diffusion transformer and used to modulate denoising blocks in an AdaLN-style manner, supplying a scene-level control signal that steers the denoising trajectory toward the intended composition. Importantly, we keep the composition classification supervision on τ_c enabled during generation training to handle the domain shift introduced by the perturbed reference images. This encourages τ_c to remain a stable carrier of compositional cues and reduces the tendency to fall back to shortcut copying.

4.3 Stage-wise Training Objective and Strategy

We train composition perception and reference-guided generation in two stages, rather than jointly, to stabilize optimization and to better match their heterogeneous input domains.

Stage I: composition perception. We optimize the unified LMM on the perception data with a standard instruction-following objective together with the composition supervision on the expert token:

$$\mathcal{L}_{\text{perc}} = \mathcal{L}_{\text{ntp}} \; + \; \lambda_{\text{cls}}\, \mathcal{L}_{\text{cls}}, \tag{9}$$

where $\mathcal{L}_{\text{ntp}}$ denotes the next-token prediction loss for autoregressive language models (details in the supplementary material) and $\mathcal{L}_{\text{cls}}$ is the weighted multi-

Table 1: Category-wise AP (%) for composition understanding on the COMP-11 perception test split. † denotes understanding-only models, and ‡ denotes unified models.

Method	RoT	Center	Hori.	Vert.	Diag.	Curved	Tri.	Sym.	Patt.	Lead.	Fill	mAP ↑
AesExpert† [17]	42.3	50.6	53.0	50.2	47.8	50.4	63.9	56.3	54.4	58.1	50.1	49.5
Qwen3-VL† [2]	56.7	70.8	78.7	76.9	54.1	54.6	87.2	86.0	61.8	72.5	66.2	66.4
InternVL3† [44]	50.6	65.7	57.5	55.9	55.8	64.3	82.8	78.3	65.5	62.3	52.6	60.0
Janus-Pro‡ [6]	50.2	55.1	56.7	66.8	52.7	51.5	71.0	69.5	55.5	51.5	52.3	54.2
BAGEL‡ [9]	51.1	66.4	80.2	68.2	54.2	57.9	87.4	83.2	70.7	74.0	52.8	63.2
Ming-Lite-Uni‡ [11]	31.4	68.1	67.4	63.2	63.9	57.3	84.2	82.0	69.8	65.8	62.2	58.2
COMPASS (ours)‡	**90.1**	**86.2**	**96.1**	**97.5**	**96.2**	**92.9**	**98.4**	**98.1**	**95.3**	**96.2**	**87.2**	**90.6**

label classification loss defined in Eq. (6). During this stage, we keep the backbone frozen and update only the composition-specific modules, including the C-MoE branch, the learnable expert token τ_c and its lightweight prediction head.

Stage II: reference-guided generation. We then train the generation pipeline under structural bottlenecking, where the model must follow the text prompt while transferring only layout cues from the perturbed reference. The objective combines the diffusion training loss with the same $\mathcal{L}_{\text{cls}}$ on τ_c:

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{diff}} \; + \; \lambda'_{\text{cls}} \, \mathcal{L}_{\text{cls}}. \tag{10}$$

Here $\mathcal{L}_{\text{diff}}$ is the standard diffusion objective used by the denoiser (detailed in the supplementary material). In this stage, we freeze the C-MoE composition experts learned in Stage I to preserve the acquired perception capability and to maintain a single shared expertization module across tasks. Meanwhile, we train all generation-specific parameters and re-train the expert token τ_c (and its lightweight head) to account for the domain shift of reference inputs. Although τ_c is trained separately across stages, it is lightweight and can be packaged with the unified system as task-specific parameters.

5 Experiments

Experimental setup. With the constructed COMP-11 dataset, we evaluate COMPASS on both composition perception and composition-guided generation. For perception, we hold out 20% of the full dataset as a test split; the class distribution of the test set is provided in the supplementary material. For generation, we further sample 10,000 evaluation groups from the perception test split with a uniform coverage over composition categories. Each group consists of (i) a reference image that provides layout intent and (ii) a content reference image whose caption is used as the semantic prompt to the model. Additional implementation details are included in the supplementary material.

Table 2: Quantitative results on composition-guided generation. † denotes generation-only models, and ‡ denotes unified models.

Method	FID ↓	CLIPScore ↑	Comp-Cons. ↑	$\cos(h_c^{\text{ref}}, h_c^{\text{gen}})$ ↑
Step1X-Edit† [21]	**7.5**	0.72	0.60	0.51
BAGEL‡ [9]	7.8	0.76	0.59	0.48
Ming-Lite-Uni‡ [11]	9.6	0.64	0.53	0.44
COMPASS (ours)‡	8.2	**0.83**	**0.74**	**0.62**

Table 3: Ablation study on composition-guided generation. ⋆ denotes "same as the previous row" with additional removals. $^{\text{P}}$ and $^{\text{G}}$ indicate perception-stage and generation-stage ablations, respectively.

Variant	mAP ↑	FID ↓	CLIPScore ↑	Comp-Cons. ↑	$\cos(h_c^{\text{ref}}, h_c^{\text{gen}})$ ↑
Full (COMPASS)	**90.6**	8.2	**0.83**	**0.74**	**0.62**
w/o C-MoE$^{\text{P}}$	78.2	–	–	–	–
w/o $\tau_c{}^{\text{P}}$	71.1	–	–	–	–
w/o cross attention$^{\text{G}}$	–	**8.0**	0.79	0.64	0.56
⋆ + w/o mask$^{\text{G}}$	–	8.6	0.72	0.59	0.48
⋆ + w/o $\tau_c{}^{\text{G}}$	–	8.3	0.75	0.53	0.42

5.1 Quantitative results

Metrics. We report results on three aspects. (i) *Category-level composition understanding*: we evaluate each composition category as a binary decision and report average precision (AP), which directly reflects the model's discriminative capability under class imbalance. (ii) *Composition-guided generation*: we measure general quality by FID [14] and CLIPScore [13], and measure composition transfer by the agreement (Comp-Cons.) between predicted composition labels (by COMPASS) of the reference and generated images and the cosine similarity ($\cos(h_c^{\text{ref}}, h_c^{\text{gen}})$) between their expert-token hidden representations. (iii) *Head-only classification* (in supplementary material): we compute multi-class accuracy using the lightweight head attached to the expert token. Detailed definitions can be found in the supplementary material.

Baselines. For perception, we compare against strong multimodal models, including an open-source aesthetic finetuned expert model: AesExpert [17], state-of-the-art VLMs: Qwen3-VL (8B) [2] and InternVL3 (8B) [44], and unified models: Janus-Pro (7B) [6], BAGEL (7B) [9], and our backbone Ming-Lite-Uni (8B) [11]. For generation, we compare against image editing/unified models that support reference-guided generation, including Step1X-Edit [21], BAGEL [9], and Ming-Lite-Uni [11]. When a baseline does not support pixelated grayscale references, we feed the original reference image and prepend the prompt with *"Generate an image with the composition style of the reference image and the following content: "* for a fair comparison.

Main results. Table 1 reports category-level AP for composition understanding. We observe that even strong contemporary LMMs still struggle to reliably identify fine-grained composition types, with performance often varying substantially across categories, highlighting that composition-centric reasoning remains underdeveloped in general-purpose multimodal models. In contrast, COMPASS consistently improves recognition across all categories, demonstrating that explicit composition expertization and direct supervision on the expert token can effectively elicit composition knowledge. Table 2 summarizes quantitative generation results, where the substantial gains in composition-specific metrics and CLIPScore confirm that our framework effectively inherits the reference layout intent while successfully mitigating semantic leakage from the reference image. More importantly, by grounding generation on explicit composition understanding, COMPASS enables composition-intent controllable generation—an expert capability that is largely absent from current unified models.

Ablations. We ablate key components in both perception and generation. The results are presented in Table 3. On perception, removing either C-MoE or the expert token causes a pronounced drop in mAP, suggesting that composition recognition benefits from both (i) explicit capacity allocation for composition features (C-MoE) and (ii) a dedicated, directly supervised intent carrier (τ_c). On generation, removing the cross-attention refinement and the customized attention mask mainly harms text faithfulness, reflected by a decrease in CLIPScore. Finally, disabling the expert-token conditioning degrades composition-transfer metrics, confirming that τ_c serves as the global layout capsule that stabilizes composition control throughout the denoising trajectory. Due to space limitations, detailed numerical results are provided in the supplementary material.

5.2 Qualitative results

Figure 6 presents representative qualitative comparisons for composition-guided generation. Rather than directly copying the reference layout, COMPASS first infers the compositional intent implied by the reference and then integrates this intent into the semantic content, yielding generations that are more harmonious and self-consistent while remaining faithful to the prompt semantics. Additional qualitative examples for both composition perception and composition-guided generation, together with failure-case analysis, are provided in Figure 1 and the supplementary material.

6 Conclusion

In this work, we present COMPASS, the first unified framework to bridge expert-level composition perception and reference-guided creation within a single LMM architecture. Supported by our systematic COMP-11 dataset, COMPASS utilizes an expert-anchor paradigm and a self-supervised structural bottleneck to effectively decouple compositional intent from semantic content without requiring paired training data. This methodology transforms unified models from

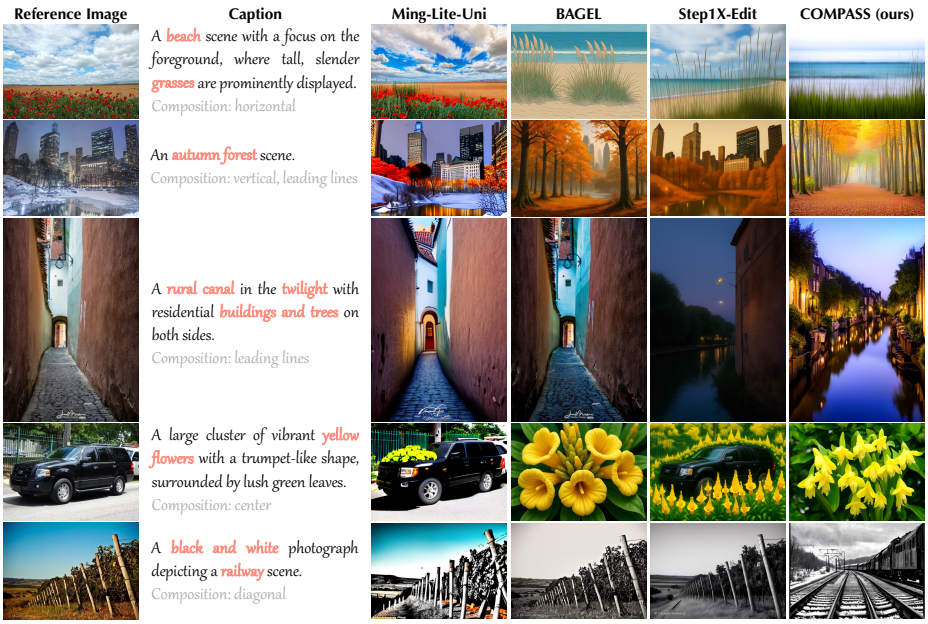

Fig. 6: Qualitative comparison on composition-guided generation. Gray annotations under the captions are for intuitive illustration only and are not used as model inputs.

composition-blind observers into expert-level aesthetic creators, establishing a generalizable blueprint for domain-specific multimodal intelligence.

Acknowledgment. This work was partially supported by the National Natural Science Foundation of China (62572469, 62572458) and the Ant Group. Z. Zhou acknowledges the funding support from the CDT in Machine Learning Systems at the University of Edinburgh, and would like to thank Laura Sevilla-Lara for her valuable support.

References

1. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.J.: Artemis: Affective language for visual art. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11569–11579 (2021)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Cao, S., Li, J., Li, X., Pu, Y., Zhu, K., Gao, Y., Luo, S., Xin, Y., Qin, Q., Zhou, Y., Chen, X., Zhang, W., Fu, B., Qiao, Y., Liu, Y.: UniPercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. arXiv preprint arXiv:2512.21675 (2025)
4. Cao, S., Ma, N., Liu, Y., et al.: ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15313–15322 (2026)

5. Chameleon Team: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-Pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
7. Chen, Y.L., Huang, T.W., Chang, K.H., Tsai, Y.C., Chen, H.T., Chen, B.Y.: Quantitative analysis of automatic image cropping algorithms:a dataset and comparative study. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2017)
8. Chung, J., Hyun, S., Heo, J.P.: Style Injection in Diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8795–8805 (2024)
9. Deng, C., Xie, J., Zhang, D.J., Gu, Y., Mao, W., Bai, Z., Wang, W., Lin, K.Q., Chen, Z., Yang, Z., Shou, M.Z.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
10. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: SEED-X: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)
11. Gong, B., Zou, C., Zheng, D., Yu, H., Chen, J., Sun, J., Zhao, J., Zhou, J., Ji, K., Ru, L., et al.: Ming-Lite-Uni: Advancements in unified architecture for natural multimodal interaction. arXiv preprint arXiv:2505.02471 (2025)
12. He, S., Zhang, Y., Xie, R., Jiang, D., Ming, A.: Rethinking image aesthetics assessment: Models, datasets and benchmarks. In: International Joint Conference on Artificial Intelligence (IJCAI). vol. 6, p. 22 (2022)
13. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528 (2021)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: International Conference on Neural Information Processing Systems (NIPS). vol. 30 (2017)
15. Hong, J., Yuan, L., Gharbi, M., Fisher, M., Fatahalian, K.: Learning subject-aware cropping by outpainting professional photos. In: AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 2175–2183 (2024)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
17. Huang, Y., Cao, Z., Li, Q., Zhao, C., Gao, C., Yan, Z., Zuo, Y., Yang, F., Cheng, D., Tan, P., et al.: AesExpert: Towards multi-modality foundation model for image aesthetics perception. In: ACM International Conference on Multimedia. pp. 5925–5936 (2024)
18. Huang, Y., Yuan, Q., Sheng, X., Yang, Z., Wu, H., Chen, P., Yang, Y., Li, L., Lin, W.: AesBench: An expert benchmark for multimodal large language models on image aesthetics perception. arXiv preprint arXiv:2401.08276 (2024)
19. Lee, J.T., Kim, H.U., Lee, C., Kim, C.S.: Photographic composition classification and dominant geometric element detection for outdoor scenes. Journal of Visual Communication and Image Representation **55**, 91–105 (2018)
20. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: GLIGEN: Open-set grounded text-to-image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22511–22521 (2023)

21. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., Jiang, D.: Step1X-Edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
22. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., yu, X., Zhao, L., Wang, Y., Liu, J., Ruan, C.: JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. arXiv preprint arXiv:2411.07975 (2024)
23. Murray, N., Marchesotti, L., Perronnin, F.: AVA: A large-scale database for aesthetic visual analysis. In: IEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2408–2415 (2012)
24. Su, Y., Cao, Y., Deng, J., Rao, F., Wu, Q.: Spatial-semantic collaborative cropping for user generated content. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 4988 – 4997 (2024)
25. Team, Q.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
26. Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: MetaMorph: Multimodal understanding and generation via instruction tuning. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17001–17012 (2025)
27. Wang, H., Wang, Q., Bai, X., Qin, Z., Chen, A.: InstantStyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 (2024)
28. Wang, X., Zhang, Z., Wang, W., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
29. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., Luo, P.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12966–12977 (2025)
30. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: NExT-GPT: Any-to-any multimodal LLM. In: International Conference on Machine Learning (ICML) (2024)
31. 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 (2024)
32. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: International Conference on Learning Representations (ICLR) (2025)
33. Xu, R., Xi, W., Wang, X., Mao, Y., Cheng, Z.: Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
34. Xu, Y., Wei, H., Lin, M., Deng, Y., Sheng, K., Zhang, M., Tang, F., Dong, W., Huang, F., Xu, C.: Transformers in computational visual media: A survey. Computational Visual Media **8**(1), 33–62 (2022)
35. Yang, G.Y., Zhou, W.Y., Cai, Y., Zhang, S.H., Zhang, F.L.: Focusing on your subject: Deep subject-aware image composition recommendation networks. Computational Visual Media **9**(1), 87–107 (2023)
36. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)

37. You, Z., Wang, K., Zhang, H., Cai, X., Gu, J., Xue, T., Dong, C., Zhang, Z.: PhotoFramer: Multi-modal image composition instruction pp. 10197–10207 (2026)
38. Zhang, J., Guo, J., Sun, S., Lou, J.G., Zhang, D.: LayoutDiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22490–22499 (2023)
39. Zhang, L., Niu, Y., Liu, W., et al.: Image composition assessment with saliency-augmented multi-pattern pooling (2021)
40. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (2023)
41. Zhao, Z., Lu, P., Hu, Y., Zhang, A., Chen, S., Li, P., Li, X., Liu, X., Wang, L., Guo, W.: Can machines understand composition? dataset and benchmark for photographic image composition embedding and understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14411–14421 (2025)
42. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. In: International Conference on Learning Representations (ICLR) (2025)
43. Zhou, Z., Wang, Q., Lin, B., Su, Y., Chen, R., Tao, X., Zheng, A., Yuan, L., Wan, P., Zhang, D.: UNIAA: A unified multi-modal image aesthetic assessment baseline and benchmark. arXiv preprint arXiv:2404.09619 (2024)
44. Zhu, J., Tian, H., Gao, Z., Liu, Y., Li, H., Chen, D., Jiang, T., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)