

Can Machines Understand Composition? Dataset and Benchmark for Photographic Image Composition Embedding and Understanding

Zhaoran Zhao Peng Lu* Anran Zhang Peipei Li Xia Li Xuannan Liu
 Yang Hu Shiyi Chen Liwei Wang Wenhao Guo
 Beijing University of Posts and Telecommunications, Beijing, China

Abstract

With the rapid growth of social media and digital photography, visually appealing images have become essential for effective communication and emotional engagement. Among the factors influencing aesthetic appeal, composition—the arrangement of visual elements within a frame—plays a crucial role. In recent years, specialized models for photographic composition have achieved impressive results across various aesthetic tasks. Meanwhile, rapidly advancing multimodal large language models (MLLMs) have excelled in several visual perception tasks. However, their ability to embed and understand compositional information remains underexplored, primarily due to the lack of suitable evaluation datasets. To address this gap, we introduce the Photographic Image Composition Dataset (PICD), a large-scale dataset consisting of 36,857 images categorized into 24 composition categories across 355 diverse scenes. We demonstrate the advantages of PICD over existing datasets in terms of data scale, composition category, label quality, and scene diversity. Building on PICD, we establish benchmarks to evaluate the composition embedding capabilities of specialized models and the compositional understanding ability of MLLMs. To enable efficient and effective evaluation, we propose a novel Composition Discrimination Accuracy (CDA) metric. Our evaluation highlights the limitations of current models and provides insights into directions for improving their ability to embed and understand composition.

1. Introduction

With the rapid growth of social media platforms, digital photography has become a primary medium for information conveying and emotional expression [37]. Aesthetically pleasing images not only enhance information conveyance but also resonate emotionally with audiences [38]. At the heart of this is **Composition**, which entails the pro-

cess of identifying and arranging the elements to produce a coherent image [36], and is a critical factor in assessing and improving the aesthetic quality of photographic works [9, 44, 52].

Strong compositional skills can produce harmonious and visually striking effects [13, 42, 53], thereby enhancing the artistic expression of a photographic image. Therefore, as artificial intelligence technology continues to evolve, researchers are exploring the machine’s ability to appreciate and create beauty, with the understanding and representation of composition becoming an indispensable foundation for achieving aesthetic appreciation and creation.

The models employed in photographic image composition research can be broadly categorized into specialized models and multimodal large language models (MLLMs). 1) Specialized models are employed to extract embeddings that capture composition information, which are then utilized in downstream aesthetic tasks such as image aesthetic assessment, aesthetics-driven image cropping, and image composition classification [12, 16, 42]. Various network architectures have been designed to integrate compositional knowledge into the learning process, aiming to enhance performance in these tasks. Despite notable improvements in aesthetic-related applications, the effectiveness of these models in embedding composition information remains insufficiently evaluated. 2) Multimodal large language models (MLLMs), have demonstrated impressive capabilities across various visual perception tasks, including image and text comprehension, video-text analysis, and image-text generation [25]. However, their ability to understand composition information in images remains underexplored.

The lack of comprehensive evaluation for both specialized models and MLLMs stems from the limitations of current composition evaluation datasets. This gap restricts the ability to guide future research and advance machine capabilities in aesthetic evaluation and creative generation. Currently, publicly available composition datasets remain limited, and existing datasets generally suffer from several issues: 1) a limited number of composition categories,

*Corresponding author.

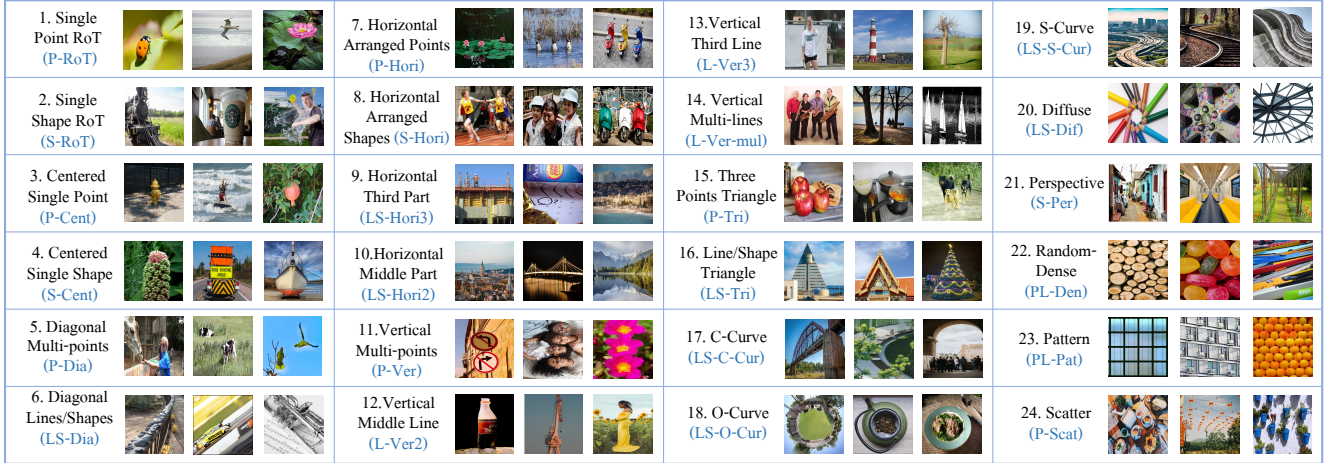


Figure 1. Sample images for 24 composition categories in PICD (blue text in parentheses represents the abbreviation)

which restricts the comprehensiveness of evaluations; 2) high noise in category labels, undermining the reliability of evaluation results; 3) small dataset sizes, leading to wider error margins and compromising statistical significance; and 4) insufficient scene diversity, which limits the assessment of model generalization capabilities.

To address these challenges, we introduce the Photographic Image Composition Dataset (PICD), a large-scale dataset consisting of 36,857 images across 24 composition categories and 355 diverse scenes. We demonstrate that PICD surpasses existing datasets in terms of data scale, composition category, label quality, and scene diversity.

Building upon PICD, we propose a comprehensive benchmark to evaluate the composition embedding ability of specialized models and the composition understanding ability of MLLMs. The benchmark includes common tasks for both model types to assess their performance in composition discrimination and robustness to semantic interference. Additionally, we design customized tasks for each model type, such as feature distribution analysis for specialized models and composition element arrangement recognition for MLLMs.

We propose a novel Composition Discrimination Accuracy (*CDA*) metric to evaluate models' ability using a simple three-choice task. Experimental results show that the *CDA* and *mAP* metrics exhibit a strong positive correlation in specialized model evaluations. However, *mAP* requires complex retrieval experiments, whereas *CDA* offers a lightweight alternative, making it especially suitable for MLLMs where composition retrieval is challenging.

Finally, we evaluate both state-of-the-art specialized models and MLLMs. The results reveal that both types face significant challenges in understanding composition. Specialized models struggle to effectively embed compositional information, while MLLMs show lower analytical accuracy.

Additionally, both model types are prone to semantic Interference.

2. Related Work

Datasets with Composition Labels: Three photographic image datasets introduce multi-category composition labels for composition assessment or classification. The KUPCP dataset [24], specifically developed for composition classification, contains 4,244 images primarily featuring outdoor scenes. However, its limited scene categories and a small number of composition categories restrict its capacity to reflect scene diversity in real-world composition learning. The CADB dataset, introduced for composition assessment tasks [54], is larger in size and provides more composition categories, yet these categories lack expert validation, leading to reliability concerns. Released in 2021, the LODB dataset [60] offers more detailed annotations for 6,263 images across 17 composition categories, covering more scenes and providing high-quality composition labels. However, it has inconsistent annotation granularity and remains limited in size. None of the existing datasets achieves fine-grained and high-quality labels, diversity scenes and large-scale simultaneously, making it challenging to accurately assess models' ability to learn composition knowledge aligned with human perception.

Specialized Models Composition Embedding: Image composition is crucial for aesthetic-related tasks like composition classification, assessment, aesthetic evaluation, and image cropping. For the first three tasks, early deep learning methods extracted composition embeddings directly using convolutional neural networks (CNNs) and trained them with composition labels[24]. More advanced approaches have been developed for composition modeling, some of which are based on saliency maps and composition rule

templates [54]; and more methods [10, 12, 40, 59] that utilize graph neural networks (GNNs) or Transformer models to learn relationships between image regions and capture composition information for downstream tasks. For the image cropping task, early works [33, 42, 53] leveraged regions of interest (ROI) or saliency information to selectively process key areas of the image and generate embeddings that assist in crop decision-making. Later studies [13] designed composition response maps based on common composition types, further improving cropping decisions by extracting composition embeddings. Further, some methods[27, 41] have incorporated graph structures to model relationships between image regions and then used GNNs for composition embedding.

MLLMs: Recent advancements in multimodal large language models (MLLMs) have significantly enhanced the ability to interpret both visual and linguistic information, with notable contributions in the field of multimodal large language models [5, 57], demonstrating impressive capabilities in understanding image-text pairs [46, 55]. Noteworthy models such as Deepseek-VL [32] and Idefics [23] have shown strong performance on multi-image tasks. Meanwhile, models like Mantis [17] and LLaVA-Next-Interleave [28] have advanced multi-image task handling with reduced computational resources. Despite these significant advancements and the introduction of benchmarks in domains such as comprehension [51] and aesthetic assessment [62], composition understanding remains an underexplored area. This gap is largely due to the absence of relevant datasets, and, to date, no dedicated benchmark has been proposed for evaluating composition understanding in multimodal models.

3. PICD Dataset

3.1. Composition Labels Design

Basic Composition Categories Design:

As shown in Fig. 2, the PICD label system are constructed along two dimensions: composition elements type and arrangement.

- **Basic Elements:** According to Kandinsky’s principles of composition [18], **points**, **lines**, and **shapes** are classic elements in graphic arts, widely employed to build visual compositions. A *point* typically refers to a small and distinct object that captures the viewer’s attention strongly. A *line* is usually formed by slender objects or a series of connected points, guiding the viewer’s gaze and creating a sense of direction within the image. A *shape* denotes a larger area with usually clear but potentially irregular boundaries [9]. These basic elements can be combined and arranged in various ways to form diverse visual compositions.
- **Arrangements:** As shown in the first green column of Fig. 2 we design 12 common arrangements inspired by

Element Arrangement	Element Type			Primary factors
	Point	Line	Shape	
Rule of Thirds	1. Single Point RoT (P-RoT)	⊗	2. Single Shape RoT(S-RoT)	Object Arrangement
Centered	3. Centered Single Point (P-Cent)	⊗	4. Centered Single Shape (S-Cent)	Object Arrangement
Diagonal	5. Diagonal Multi-points (P-Dia)	⊗	⊗	Object Arrangement Visual Guidelines
Horizontal	7. Horizontal – Arranged Points (P-Hori)	⊗	10. Horizontal Arranged Shapes (S-Hori)	Object Arrangement
	⊗	9. Horizontal Third Part (LS-Hori3)	8. Horizontal Middle Part (LS-Hori2)	Visual Guidelines
Vertical	11. Vertical Multi-points (P-Ver)	⊗	⊗	Object Arrangement
	⊗	12. Vertical Middle Line (L-Ver2)	13. Vertical Third Line (L-Ver3)	Visual Guidelines
		14. Vertical Middle -lines (L-VerMul)		
Triangle	15. Three Points Triangle (P-Tri)	⊗	⊗	Object Arrangement Visual Guidelines
C-Curve	⊗	17. C-Curve (LS-C-Cur)		Visual Guidelines
O-Curve	⊗	18. O-Curve (LS-O-Cur)		Visual Guidelines
S-Curve	⊗	19. S-Curve (LS-S-Cur)		Visual Guidelines
Radial	⊗	20. Diffuse (LS-Dif)	21. Perspective (S-Per)	Visual Guidelines
Dense	22. Random- Dense (PL-Den)	23. Pattern (PL-Pat)	⊗	Object Arrangement
Scatter	24. Scatter (P-Scat)	⊗	⊗	Object Arrangement

Figure 2. The PICD label system, which is expanded based on two dimensions: element type and arrangement. Column 1 (green) represents the element arrangement, while Columns 2, 3, and 4 correspond to three types of compositional elements. The composition categories are numbered from 1 to 24 with abbreviated labels in blue brackets. Merged categories are highlighted in red boxes, and categories excluded due to non-existence or insufficient samples are marked with blue strikethroughs. Column 5 (blue and purple) highlights the primary compositional factors.

[9, 54]: Rule of Thirds, Centered, Diagonal, Horizontal, Vertical, Triangle, C-curve, O-curve, S-curve, Radial, Dense and Scatter.

The combination of 12 arrangements and 3 basic elements theoretically results in 36 distinct composition categories.

Category Optimization: Although it is theoretically possible to generate 36 categories, we further optimized these categories in practice, as shown in Fig.2. Each arrangement method’s visual effects may be generated by two factors: object arrangement (highlighted in purple box) and vi-

sual guidelines [35] (highlighted in blue box). If guidelines mainly contribute to a composition, the guidelines may be composed of line elements or shape boundaries, producing similar visual effects. Therefore, we combine these compositions into the categories indicated by the red bounding boxes in Fig 2. To more precisely describe different visual effects, we further expand some categories. For instance, in the “Vertical” and “Horizontal” categories, the number and positioning of guidelines significantly impact visual perception. Thus, “Horizontal” is subdivided into “Horizontal Middle Part” and “Horizontal Third Part”, while “Vertical” is divided into “Vertical Middle Line”, “Vertical Thirds Line”, and “Vertical Multi-lines”. Additionally, in the “Radial” category, we distinguish between “Perspective Composition” and “Diffuse Composition” based on the presence of perspective relationships. In the “Dense” category, regular repetition creates a harmonious rhythm, so it is further divided into “Pattern” (regular repetition) and “Random Dense”. The detailed description of each composition category can be found in the Appendix.

3.2. Data Annotation Process

Data Source: The proposed PICD dataset aggregates images from multiple open-source platforms and public datasets, including open platforms like Unsplash [3], Pexels [2], Flickr [1] and publicly available image datasets (AADB [21], AVA [34], OpenImage [22], PARA [48], COCO2017 [29], TAD66k [11], VOC2007 [8], VOC2012 [7]).

Script-Assisted Candidate Image Preparation: To quantitatively define the composition constructs in Sec.3.1, we first establish quantification rules for point, line, and shape elements, based on object area ratio, aspect ratio and detection confidence. Leveraging object detection [43] and semantic line detection [58] techniques, we identify potential composition elements within images and extract the possible count, position, size, and category information of each element. Using this information, we develop scripts to automatically filter candidate images based on element category, count, and spatial arrangement.

Handling of Special Categories: For categories such as perspective, radial alignment, and geometric shapes (categories 16-21 in Fig. 2), where script-based filtering proves less efficient, we employ a keyword-based search strategy. Specifically, we use targeted keywords (e.g., “perspective,” “radial,” “C-shape”) on open platforms like Unsplash [3] to collect additional images. This process enhances the diversity of the dataset and supplements categories that are challenging to capture through automated scripts.

Expert Voting: To ensure annotation accuracy and consistency, we invite five experienced photography and visual design experts to manually annotate the candidate images. An online platform is developed to facilitate and manage

the expert review process, with the workflow outlined as follows:

1. **Expert Review Platform:** Experts independently assess whether each candidate image meets the compositional requirements of a given category. To maintain impartiality and prevent bias, experts are restricted from viewing the evaluations of their peers.
2. **Annotation Rules and Consensus Confirmation:** A given annotation is confirmed only if at least three experts reach a consensus; otherwise, the image is excluded from the final dataset.

Scene Diversity Monitoring: To encourage scene diversity across different composition categories, we utilize a scene classification network [61] to monitor and analyze the distribution of scene types within each composition category during voting. Suppose a composition category contains 50 consecutive candidate images from the same scene type. In that case, the platform prioritizes displaying candidate images from other scene types, thus maintaining a varied selection among candidate images.

3.3. Statistics and Advantage of PICD

Table 1. Comparison of Dataset Characteristics. PICD achieves the largest dataset scale, greatest diversity in composition categories, broadest scene variety, and highest label quality.

	Image Num.	Composition Cate.	Average Num.	Scene Num.	Label Quality
KUPCP[24]	5584	9	620	276	0.618
CADB[54]	9497	13	731	351	0.597
LODB[60]	6263	17	368	331	0.816
PICD	36857	24	1536	355	0.885

Dataset Size and Composition Categories: Table 1 summarizes each dataset’s Image number (2nd column), number of composition categories (3rd column), and average image number per category (4th column). The blue bar chart in Fig. 3 further illustrates the distribution of images per composition category across various datasets. Notably, PICD provides the largest dataset size, the highest number of composition categories, and the most substantial average image count per category. These characteristics ensure comprehensive evaluations, reduce error margins and enhance the statistical significance of the results.

Scene Diversity: Photographic images in practical applications span a diverse range of scenes, with each composition category encompassing multiple scene classes. We employ a scene classification network trained on the Places365 dataset [61] to classify scenes within each dataset, revealing that PICD has the most extensive scene variety, as shown in Table 1. As indicated by the sienna line in Fig. 3, PICD includes 23 composition categories with over 100 scene classes, covering 95.83% of its categories. In comparison,

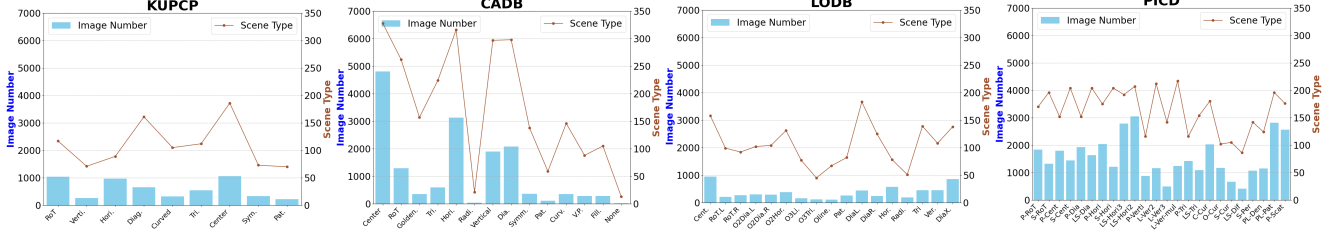


Figure 3. Comparison of image counts and scene variety across composition categories in various datasets. The blue bar chart (left axis) represents the number of images in each composition category, while the sienna line chart (right axis) indicates the scene types existing in each composition category.

LODB, CADB, and KUPCP contain 52.94%, 76.92%, and 55.56%, respectively. This high level of scene diversity in PICD underscores its suitability for robust evaluation of composition learning across a broad spectrum of scenes.

Label Quality: To evaluate label quality, we conduct a user study. Specifically, we randomly select 50 images from each of the four datasets across all composition categories, resulting in a total of 150 images. These images are shuffled, and each image is presented with four label descriptions: one correct label corresponding to the image composition, and three distractor labels from other categories within the same dataset to avoid cross-dataset interference, thereby enhancing the validity of the comparison. Users are instructed to select the label that best matches each image’s composition. A total of 18 participants independently completed the label selection task. We calculate the average selection accuracy across all users as the label quality score, as shown in Table 1, the label quality score of PICD is significantly higher than that of other datasets.

4. Benchmark

4.1. Evaluation of Specialized Models

Four tasks are designed to evaluate the effectiveness of composition information embedded in specialized models:

Task I. Composition Triplet Distinction: This task is a three-choice evaluation designed to assess the ability of embeddings obtained by specialized models to distinguish between different compositions. Specifically, given a triplet $T = \langle I_{p1}, I_{p2}, I_n \rangle$, where I_{p1} and I_{p2} are images from the same composition category in the PICD dataset and I_n is from a different composition category, the task is to identify the image with a composition distinct from the other two.

- **Evaluation Settings:** To construct the triplets, we first traverse the 24 composition categories in PICD. For each category, it is treated as the positive category, while the other 23 categories are considered negative. This results in 24×23 positive-negative combinations. For each combination, we randomly sample 12 triplets, each consisting

of two positive samples from the positive category and one negative sample from the negative category. Therefore, the total number of triplets is 6,624. To identify the negative sample within each triplet, we calculate pairwise distances among the composition embeddings generated by the model. The two images with the shortest distance are assumed to share the same composition category, while the third image is considered the negative sample. To further compare the challenges posed by existing datasets and PICD, we also generate triplets using the same approach on KUPCP [24], CADB [54] and LODB [60], and perform a dataset comparison on this task.

- **Metrics:** To assess the performance of specialized models in the Composition Triplet Distinction Task, we design Composition Discrimination Accuracy (CDA), calculated as the proportion of triplets where the model correctly identifies the negative sample:

$$CDA = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{neg}_i = neg_i), \quad (1)$$

where N denotes the total number of triplets within a specific composition category, $\hat{neg}_i$ is the predicted negative sample for the i -th triplet, neg_i represents the ground truth, and $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the prediction is correct and 0 otherwise. We calculate the average CDA across all categories to capture overall model performance, noted as $avg-CDA_1$.

Task II. Robustness towards Semantic Interference: The second part aims to evaluate whether the composition distinction ability evaluated in Task I is impacted by semantic information.

- **Evaluation Settings:** Unlike the triplets used in the Composition Triplet Distinction Task, this section includes 2,760 triplets specifically designed to introduce semantic interferences. In each triplet, the negative sample is semantically similar to one of the positive samples, while the two positive samples differ in semantic content. We define semantically similar samples as those sharing object semantic labels.

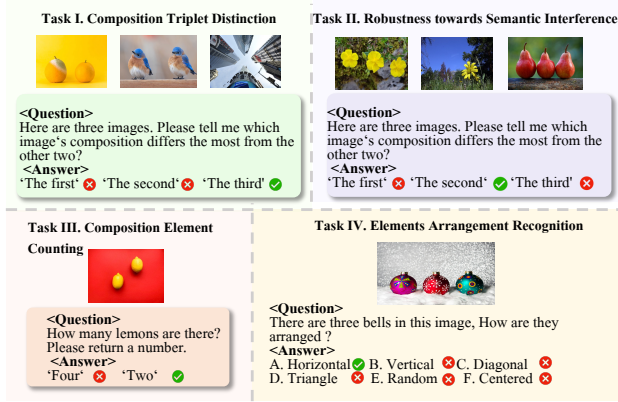


Figure 4. Sample questions for Task I-IV in evaluation of MLLMs

- Metrics: We compute average CDA the same as Task I, noted as $avg-CDA_2$.

Task III. Composition Retrieval: This task evaluates the model’s ability to distinguish and match composition categories, aligned with real-world applications, and assesses the effectiveness of embeddings from specialized models in matching composition similarities within a large-scale dataset.

- Evaluation Settings: Specifically, we randomly sampled 50 images from each of the 24 composition categories in PICD to create a query set of 1,200 images, while the remaining images were designated as the gallery set.
- Metrics: To validate the accuracy of composition retrieval, we calculated the mean Average Precision at k ($mAP@k$) [14, 20] for each composition category and compute the average, noted as $avg-mAP@k$ ($k=10, 50, 100$).

Task IV. Feature Distribution Analysis: To evaluate the effectiveness of embeddings obtained by specialized models, we analyze their distribution in high-dimensional space. Greater concentration within categories and clearer separation between them indicate that the embeddings effectively capture compositional characteristics.

- Metrics: We utilize the Silhouette Score [39] and Davies-Bouldin Index (DBI) [15] as metrics to assess the compactness of the feature distribution within each composition category and the separation between categories.

4.2. Evaluation of MLLMs

Four evaluation tasks are designed to assess the effectiveness of MLLMs’ composition understanding ability, and the examples of questions are shown in Fig.4.

Tasks I & II. Composition Triplet Distinction & Robustness towards Semantic Interference: These two tasks are designed to evaluate the MLLMs’ capability to distinguish between different composition categories, as well as to as-

sess whether the composition discriminative ability of the embeddings is influenced by semantic information.

- Evaluation Settings: Triplet construction for both tasks adheres to the same methodology used in specialized model evaluations, with the primary difference being in the identification of the negative sample. For each inference, the three images in T are presented in a randomized order with the prompt: “Which of these three images has a composition that differs from the other two?” The model is required to select one image (e.g., the first, second, or third) based on compositional difference. Question samples are shown in Fig.4.

- Metrics: Same with Task I&II for specialized models.

Task III. Composition Elements Counting: For photographic images, especially when the number of elements is small, subtle differences in element count can significantly impact the compositional characteristics. This task is designed to evaluate the MLLM’s ability to accurately count compositional elements.

- We manually annotated 1,593 images from P-Cent, S-Cent, P-Dia, P-Hori, LS-Hori2, P-Ver, and L-Ver-mul in PICD. Each image contains between 1 and 6 objects, representing varying levels of scene complexity. For each evaluation, a single image and its corresponding prompt are provided to the MLLM, which is tasked with returning the number of composition elements. Example prompts are shown in Fig. 4.
- Metrics: We calculate the average accuracy across different scene complexity as the metric for evaluating the model’s ability to count primitives, denoted as Acc_n .

Task IV. Elements Arrangement Recognition: The arrangement of elements is one of the key factors influencing composition. This task aims to assess the MLLM’s ability to recognize the arrangement of fundamental compositional elements.

- Evaluation Settings: We use the same images selected in Task III, specifying the categories and quantities of elements in the prompt. Based on the given image, the model is required to have the MLLM select one of six predefined arrangements. Example questions are shown in Fig. 4.
- Metrics: We calculate the average accuracy across different arrangement methods as the metric for evaluating the model’s ability for arrangement recognition, denoted as Acc_t

5. Results and Analysis

5.1. Specialized Models

Baselines: We developed composition feature extractors based on ten publicly available network architectures with pre-trained weights, spanning four task categories: composition classification (PCPNet [24]), composition assessment (SAMPNet [54]), aesthetic evaluation (MUSIQ [19],

EAT [12], HLGCN [40], BAID [50]), and cropping methods (CAC [13], CGS [27], S2C [41], GANC [53]). We utilized the network architecture up to the classification layer as the composition feature extractor for composition classification, composition assessment, and image aesthetic assessment. For cropping networks, we used the full image as the target region and extracted features from the network layers preceding the cropping quality prediction module.

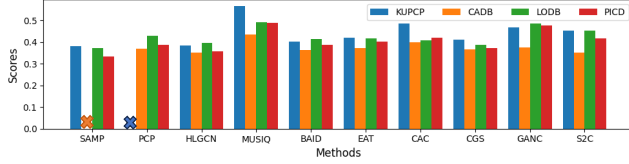


Figure 5. $avg-CDA_1$ of all specialized models on Task I across different datasets.

Task I. Composition Triplet Distinction: As shown in Fig. 5, significant differences exist in the challenges posed by different datasets. The KUPCP dataset has limited scene diversity, resulting in an overestimation of the performance. The average $avg-CDA_1$ of all methods is 0.44. This makes it unsuitable for evaluating robustness against semantic interference. The CADB dataset contains considerable noise, reducing data reliability. The results show that the scores of the methods are relatively concentrated, with a lower score variance compared to other datasets. The LODB dataset is smaller in scale, with an average score of 0.43, which may overestimate the methods’ performance and reduce the results’ reliability. The PICD dataset, being larger in scale, with diverse scenes, more composition categories, and high-quality annotations, provides a more accurate reflection of algorithm performance, with an average score of 0.40. Furthermore, as shown in Table 2, the $avg-CDA_1$ values for all specialized methods do not exceed 0.5, indicating that the composition embeddings generated by current models have limited ability to distinguish between different composition categories.

Task II. Robustness towards Semantic Interference: The $avg-CDA_2$ in Table 2 shows a reduction compared to $avg-CDA_1$, indicating varying levels of dependence on semantic information within the composition embeddings across these models. This suggests that directly leveraging features from deep networks for composition embeddings often introduces substantial semantic content, causing a coupling between semantic and composition information within the embedding vectors. As a result, even when there are significant compositional differences between positive and negative samples, the network may still incorrectly classify them under the same composition if their semantic content is similar. This coupling can impair the generalization of the composition embeddings.

Task III. Composition Retrieval: As shown in Fig. 6, All

Table 2. Performance of the specialized models in Task I, II & IV.

Model	Task I	Task II	Task IV	
	$avg-CDA_1$	$avg-CDA_2$	<i>Silhouette</i>	<i>DBI</i>
SAMP [54]	0.334	0.332	-0.014	42.385
PCP [24]	0.388	0.354	-0.116	10.894
HLGCN [40]	0.357	0.343	-0.234	56.929
MUSIQ [19]	0.488	0.367	-0.010	10.624
BAID [50]	0.387	0.374	-0.100	13.124
EAT [12]	0.401	0.338	-0.066	10.340
CAC [13]	0.419	0.405	-0.126	11.877
CGS [27]	0.371	0.352	-0.195	11.578
GANC [53]	0.478	0.445	-0.076	6.493
S2C [41]	0.417	0.383	-0.059	11.669

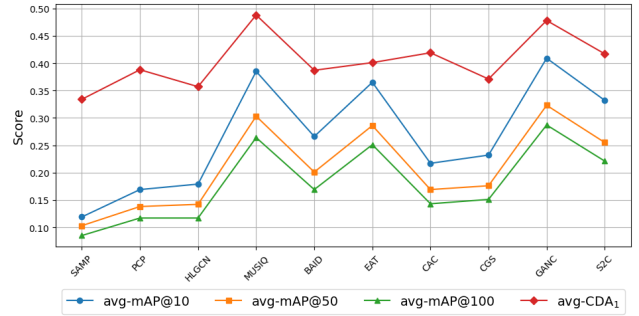


Figure 6. $avg-mAP@k$ and $avg-CDA_1$ across all specialized models in Task III, showing a clear positive correlation between the two metrics.

methods yield $avg-mAP@10$ below 0.5, $avg-mAP@50$ and $avg-mAP@100$ below 0.35, indicating limited retrieval capability on the large-scale dataset. Among them, aesthetic evaluation methods (MUSIQ, EAT) and cropping-based methods (GANC, S2C) perform relatively better due to their use of larger training datasets and more effective supervision tasks respectively.

Notably, We computed the Spearman correlation coefficient and p-value to assess the relationship between $avg-mAP@10$ and $avg-CDA_1$. We observe a significant positive correlation with a Spearman correlation [45] of 0.758 and a p-value [47] of 0.011, as shown qualitatively in Fig. 6. This finding suggests that the CDA metric can serve as a lightweight proxy for retrieval performance. This makes it particularly suitable for MLLMs where composition retrieval is challenging.

Task IV. Feature Distribution Analysis: Table 2 shows that all methods exhibit a Silhouette Score below 0 (with a range from -1 to 1, where values closer to 1 indicate tighter intra-cluster cohesion), suggesting that the sample points have low cohesion within their respective clusters and are prone to being misclassified into other categories. Additionally, the DBI (ranging from 0 to positive infinity, with values closer to 0 indicating tighter intra-cluster cohesion) is greater than 10, indicating poor separation between categories with minimal distinguishability. These results high-

light that the embeddings learned by the specialized models fail to effectively capture the composition information, as the clustering exhibits poor internal consistency and insufficient category separation.

5.2. MLLMs

Baselines: We selected several representative models, including both open-source and proprietary options: Qwen-VL-max [4], Qwen-VL-plus [4], InterVL2 [6], InternVL2-AWQ [6], InternLM-XComposer-2.5(XComposer2.5) [56], LLaVA-Onevision [26], LLaVA-v1.6-mistral [31], LLaVA-v1.5 [30], Idefics2 [17], MiniCPM [49].

Table 3. Performance of MLLMs in four evaluation tasks.

Model	Task I	Task II	Task III	Task IV
	$avg-CDA_1$	$avg-CDA_2$	Acc_n	Acc_t
XComposer2.5 [56]	0.263	0.250	0.504	0.364
InternVL2 [6]	0.340	0.349	0.131	0.327
InternVL2-AWQ [6]	0.344	0.350	0.118	0.277
LLaVA-Onevision [26]	0.402	0.343	0.224	0.313
LLaVA-v1.6-mistral [28]	0.220	0.218	0.430	0.231
LLaVA-v1.5 [30]	0.332	0.325	0.496	0.363
Idefics2 [17]	0.458	0.387	0.498	0.252
MiniCPM [49]	0.332	0.325	0.472	0.421
Qwen-VL-plus [4]	0.382	0.352	0.440	0.271
Qwen-VL-max [4]	0.462	0.372	0.303	0.422

Task I. Composition Triplet Distinction: The $avg-CDA_1$ results in Table 3 show consistently low accuracy, with most values clustering around 0.33, close to random chance. This suggests that existing open-source and proprietary MLLMs exhibit limited capability in distinguishing the differences between different compositions.

Task II. Robustness towards Semantic Interference: From Table 3, the $avg-CDA_2$ for various methods is consistently lower than $avg-CDA_1$, suggesting that these models struggle to disentangle compositional information from semantic content and are susceptible to semantic interference, which in turn hampers their effectiveness in composition understanding. This can be attributed to the lack of composition-specific training in large models, which limits their ability to distinguish purely compositional aspects from semantic interference.

Task III & IV: Element Number Counting & Arrangement Recognition: As shown by Acc_n and Acc_t in Table 3, all models achieved below 50% accuracy on element counting and topology recognition, indicating that even with known composition elements, the models struggle to accurately identify both the quantity and spatial arrangement of these elements. In addition to the potential lack of targeted training, composition recognition tasks inherently require an understanding of the relative positions of elements, which demands reasoning and inductive capabilities. Furthermore, humans can easily reason about the meaning of different compositions through long-term visual

experience and prior knowledge, whereas models typically fail to effectively utilize such prior knowledge for reasoning and judgment.

6. Discussion

Our evaluation of specialized models highlights key challenges in embedding compositional information. These models often struggle to distinguish composition categories due to the tight coupling of semantic and compositional information, which limits their generalization ability. Additionally, these models rely on implicit composition learning, which requires large labeled datasets; however, datasets specifically annotated for compositional details are both scarce and costly. In contrast, humans distinguish compositions by explicitly recognizing compositional elements and arrangements. Therefore, developing a network architecture that explicitly recognizes compositional elements and their topology, aligned with human cognition, can not only enhance interpretability but also improve efficiency, offering significant potential and warranting further exploration.

Similarly, MLLMs struggle with composition understanding, particularly in accurately perceiving compositional elements and recognizing topological relationships. The main obstacle here is the lack of sufficient training data to support effective composition understanding. While large-scale compositional datasets for MLLMs are a feasible solution, they remain practically challenging. Therefore, we propose that a promising future direction is to develop specialized, interpretable models. Through knowledge distillation, these specialized models could enhance the compositional understanding capabilities of MLLMs, combining the strengths of both approaches.

7. Conclusion

In this study, we introduce PICD, a large-scale photographic composition dataset with 36,857 images across 24 composition categories and 355 scenes. We show that PICD outperforms existing datasets in data scale, composition category, label quality, and scene variability. We also propose a benchmark based on PICD to evaluate the composition embedding and understanding abilities of specialized models and MLLMs. To facilitate effective and efficient evaluation, we introduce a novel Composition Discrimination Accuracy (CDA) metric, which correlates strongly with mAP and offers a lightweight alternative, especially for MLLMs, where composition retrieval is challenging. Our evaluation highlights current model limitations and suggests directions for improvement.

Acknowledgement This research is sponsored by National Natural Science Foundation of China (Grant No. 62306041) and Beijing Nova Program (Grant No. 20230484488).

References

- [1] Flickr: Image and video hosting service. <https://www.flickr.com>. 4
- [2] Pexels: Free stock photos videos. <https://www.pexels.com>. 4
- [3] Unsplash: Free high-resolution photos. <https://unsplash.com>. 4
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 8
- [5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024. 3
- [6] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 8
- [7] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep.*, 2007(1-45):5, 2012. 4
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 4
- [9] Michael Freeman. *The Photographer’s Eye Digitally Remastered 10th Anniversary Edition: Composition and Design for Better Digital Photos*. Routledge, 2017. 1, 3
- [10] Koustav Ghosal and Aljosa Smolic. Image aesthetics assessment using graph attention network. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3160–3167. IEEE, 2022. 3
- [11] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 942–948, 2022. 4
- [12] Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. Eat: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1023–1032, 2023. 1, 3, 7
- [13] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7057–7066, 2021. 1, 3, 7
- [14] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 12147–12157, 2021. 6
- [15] Ali Idrus. Distance analysis measuring for clustering using k-means and davies bouldin index algorithm. *TEM Journal*, 11(4):1871–1876, 2022. 6
- [16] Gengyun Jia, Peipei Li, and Ran He. Theme-aware aesthetic distribution prediction with full-resolution photographs. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8654–8668, 2022. 1
- [17] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 3, 8
- [18] Wassily Kandinsky. *Point and line to plane*. Dover Publications, 1979. 3
- [19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6, 7
- [20] Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2585–2595, 2023. 6
- [21] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charles Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 662–679. Springer, 2016. 4
- [22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 4
- [23] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 3
- [24] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *Journal of Visual Communication and Image Representation*, 55:91–105, 2018. 2, 4, 5, 6, 7
- [25] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 1
- [26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 8
- [27] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4213–4222, 2020. 3, 7

- [28] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3, 8
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 8
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 8
- [32] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3
- [33] Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin. An end-to-end neural network for image cropping by learning composition from aesthetic photos. *arXiv preprint arXiv:1907.01432*, 2019. 3
- [34] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 4
- [35] Bryan Peterson. *Learning to see creatively: Design, color, and composition in photography*. Amphoto Books, 2015. 4
- [36] David Präkel. *Basics photography 01: composition*. A&C Black, 2012. 1
- [37] Jorge Reyna, Jose Hanham, and Peter Meier. The internet explosion, digital media principles and implications to communicate effectively in the digital space. *E-learning and Digital Media*, 15(1):36–52, 2018. 1
- [38] Maria Schreiber. Audiences, aesthetics and affordances analysing practices of visual communication on social media. *Digital Culture & Society*, 3(2):143–164, 2017. 1
- [39] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020. 6
- [40] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8475–8484, 2021. 3, 7
- [41] Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-semantic collaborative cropping for user generated content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4988–4997, 2024. 3, 7
- [42] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12104–12111, 2020. 1, 3
- [43] Ultralytics. Ultralytics yolov8, 2023. Accessed: 2024-11-03. 4
- [44] Jihong Wang and Steve Langer. A brief review of human perception factors in digital displays for picture archiving and communications systems. *Journal of Digital Imaging*, 10: 158–168, 1997. 1
- [45] Clark Wissler. The spearman correlation formula. *Science*, 22(558):309–311, 1905. 7
- [46] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023. 3
- [47] Chengwei Xiao, Jiaqi Ye, Rui Máximo Esteves, and Chunming Rong. Using spearman’s correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14):3866–3878, 2016. 7
- [48] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19861–19869, 2022. 4
- [49] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 8
- [50] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22388–22397, 2023. 7
- [51] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3
- [52] Richard D Zakia and David Page. *Photographic composition: A visual guide*. Routledge, 2012. 1
- [53] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1304–1319, 2020. 1, 3, 7
- [54] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. *arXiv preprint arXiv:2104.03133*, 2021. 2, 3, 4, 5, 6, 7
- [55] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3
- [56] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 8

- [57] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024. 3
- [58] Kai Zhao, Qi Han, Chang-Bin Zhang, Jun Xu, and Ming-Ming Cheng. Deep hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4793–4806, 2021. 4
- [59] Lin Zhao, Meimei Shang, Fei Gao, Rongsheng Li, Fei Huang, and Jun Yu. Representation learning of image composition for aesthetic prediction. *Computer Vision and Image Understanding*, 199:103024, 2020. 3
- [60] Zhaoran Zhao, Peng Lu, Xujun Peng, and Wenhao Guo. Image layout learning, 2024. Accessed: 2024-08-09. 2, 4, 5
- [61] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. <https://github.com/CSAILVision/places365>, 2017. Accessed: 2024-11-04. 4
- [62] Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024. 3