Appendix of *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

The supplementary material is organized into three main sections. The first section provides additional experimental results for Task I, including evaluations of specialized models and MLLMs. The second section offers an in-depth introduction to point, line, and shape elements and a detailed explanation of the selection criteria for the 24 composition categories. The third section provides a detailed description of the multi-label data in the PICD dataset.

1. More Experimental Results

While the main text reports only the average CDA across the 24 layout categories in Evaluation Task I for specialized models and MLLMs, denoted as $avg\text{-}CDA_1$, the detailed CDA results for each layout category are presented in this section to provide a more comprehensive supplement to the main text. The CAD_1 of Task I for each composition category in PICD is recorded in Table 1 for specialized models and Table 2 for MLLMs. We also evaluate several large models for embedding features, such as CLIP [15] and DinoV2 [14]. The results can be found in Table 3.

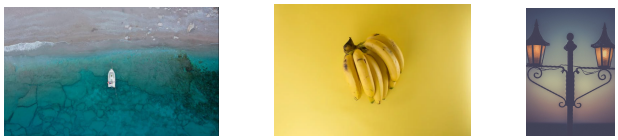
2. Details for 24 Composition Categories in PICD

In the following, we present the qualitative and quantitative selection criteria for point, line, and shape elements used during the dataset construction process, as well as the corresponding criteria for each composition category. These criteria serve as essential references for expert voting and form the foundation for script development.

2.1. Composition Element Definition

Point Element A small and well-defined target is considered a point element [3, 7]. Specifically, during the dataset construction process, we adhered to the following rules:

1) The size of the target must exceed 1/90 of the image to ensure sufficient visual significance, as shown in Fig. 1a.
2) The size of the target must be less than 1/6 of the image to distinguish it from shape elements, as shown in Fig. 1b.
3) Objects must be clear and prominent to avoid confusion with the background.
4) If a target consists of multiple parts, each part may be defined as a point element. Images of this type are discarded to avoid ambiguity. A Negative example is shown in Fig. 1c.



(a) (b) (c)

Figure 1. (a) and (b) represent the lower and upper area bounds, respectively, for defining point elements in the image. Condition (c) serves as a counterexample, where point elements are part of an object rather than standalone, leading to potential confusion and thus not being selected as valid point elements.

Line Element Elongated objects are considered line elements, which are often used to guide the viewer's visual flow[3, 7, 17, 21]. During the dataset construction process, the line elements in an image must meet the following criteria:

1) Aspect Ratio: To ensure the element is elongated, the ratio of its long side to its short side must satisfy long side: short side $\leq 1/3$, qualifying it as a line element.
2) Length of the Long Side: The element must be visually significant within the image. As shown in 2a, the ratio of the line element's long side to the corresponding side of the image must be $\geq 1/2$, serving as a boundary criterion.

*Corresponding author.

Table 1. Supplementary Results of Evaluation Task I for Specialized Models: The CDA_1 Scores Across Various Composition Categories in PICD.

	SAMP [23]	PCP [9]	HLGCN [16]	MUSIQ [8]	BAID [20]	EAT [4]	CAC [5]	CGS [11]	GANC [22]	S2C [18]
P-RoT	0.388	0.460	0.388	0.299	0.406	0.369	0.457	0.333	0.330	0.290
S-RoT	0.367	0.335	0.342	0.455	0.313	0.365	0.433	0.296	0.331	0.313
P-Cent	0.396	0.400	0.415	0.328	0.513	0.396	0.513	0.546	0.615	0.462
S-Cent	0.398	0.453	0.387	0.372	0.504	0.453	0.431	0.355	0.431	0.288
P-Dia	0.345	0.275	0.409	0.447	0.364	0.302	0.418	0.434	0.276	0.411
LS-Dia	0.356	0.388	0.322	0.413	0.295	0.514	0.396	0.326	0.458	0.338
P-Hori	0.349	0.407	0.451	0.393	0.447	0.379	0.520	0.442	0.338	0.465
S-Hori	0.280	0.342	0.309	0.444	0.378	0.411	0.400	0.283	0.404	0.389
LS-Hori3	0.333	0.486	0.355	0.652	0.609	0.412	0.507	0.431	0.638	0.543
LS-Hori2	0.382	0.491	0.338	0.639	0.629	0.429	0.531	0.409	0.669	0.578
P-Ver	0.362	0.308	0.373	0.417	0.399	0.263	0.337	0.358	0.312	0.399
L-Ver2	0.382	0.367	0.364	0.440	0.455	0.402	0.353	0.447	0.418	0.342
L-Ver3	0.370	0.388	0.333	0.451	0.440	0.413	0.451	0.225	0.355	0.311
L-Ver-mul	0.297	0.293	0.366	0.504	0.283	0.435	0.395	0.250	0.428	0.312
P-Tri	0.327	0.370	0.424	0.487	0.415	0.449	0.433	0.401	0.262	0.415
LS-Tri	0.354	0.370	0.313	0.515	0.315	0.492	0.297	0.398	0.440	0.451
LS-C-Cur	0.337	0.391	0.362	0.380	0.275	0.327	0.395	0.397	0.489	0.500
LS-O-Cur	0.293	0.493	0.341	0.471	0.312	0.529	0.464	0.224	0.594	0.301
LS-S-Cur	0.255	0.406	0.344	0.674	0.538	0.421	0.516	0.548	0.585	0.462
LS-Dif	0.261	0.304	0.275	0.558	0.203	0.267	0.359	0.235	0.478	0.337
S-Per	0.283	0.449	0.308	0.617	0.536	0.433	0.514	0.409	0.670	0.413
PL-Den	0.286	0.366	0.319	0.656	0.243	0.400	0.370	0.439	0.812	0.630
PL-pat	0.316	0.389	0.305	0.547	0.109	0.399	0.185	0.260	0.745	0.535
P-Scat	0.309	0.384	0.420	0.553	0.309	0.367	0.373	0.469	0.388	0.513

Table 2. Supplementary Results of Evaluation Task I for MLLMs: The CDA_1 Scores Across Various Composition Categories in PICD.

Methods	Qwen-VL-plus [1]	Qwen-VL-max [1]	Intern-VL2 [2]	InternVL2-AWQ [2]	X-Compo-ser2.5 [24]	LLaVA-Onevision [10]	LLaVA-v1.6mistral [12]	LLaVA-v1.5 [13]	Idefics2 [6]	Mini-CPM [19]
P-RoT	0.405	0.486	0.275	0.319	0.296	0.428	0.264	0.377	0.511	0.377
S-RoT	0.286	0.420	0.330	0.341	0.243	0.327	0.218	0.308	0.545	0.308
P-Cent	0.413	0.547	0.308	0.326	0.241	0.396	0.207	0.326	0.527	0.326
S-Cent	0.298	0.308	0.312	0.290	0.285	0.380	0.266	0.373	0.493	0.373
P-Dia	0.429	0.446	0.402	0.395	0.251	0.415	0.211	0.348	0.658	0.348
LS-Dia	0.322	0.337	0.304	0.283	0.306	0.320	0.196	0.337	0.447	0.337
P-Hori	0.421	0.522	0.355	0.315	0.215	0.382	0.204	0.348	0.593	0.348
S-Hori	0.323	0.399	0.362	0.377	0.274	0.364	0.200	0.312	0.625	0.312
LS-Hori3	0.561	0.634	0.431	0.402	0.279	0.514	0.268	0.373	0.438	0.373
LS-Hori2	0.513	0.630	0.388	0.380	0.234	0.487	0.255	0.359	0.447	0.359
P-Ver	0.349	0.435	0.351	0.388	0.285	0.351	0.210	0.319	0.525	0.319
L-Ver2	0.396	0.420	0.319	0.377	0.305	0.331	0.251	0.333	0.407	0.333
L-Ver3	0.346	0.442	0.370	0.344	0.265	0.370	0.234	0.326	0.399	0.326
L-Ver-mul	0.306	0.395	0.370	0.337	0.223	0.344	0.246	0.304	0.366	0.304
P-Tri	0.426	0.522	0.391	0.446	0.237	0.371	0.200	0.275	0.687	0.275
LS-Tri	0.308	0.348	0.243	0.236	0.320	0.409	0.191	0.359	0.214	0.359
LS-C-Cur	0.311	0.308	0.348	0.359	0.243	0.333	0.192	0.312	0.283	0.312
LS-O-Cur	0.335	0.348	0.297	0.297	0.303	0.391	0.203	0.333	0.478	0.333
LS-S-Cur	0.625	0.739	0.424	0.409	0.222	0.571	0.222	0.286	0.382	0.286
LS-Dif	0.339	0.435	0.236	0.315	0.298	0.471	0.145	0.333	0.254	0.333
S-Per	0.440	0.598	0.366	0.348	0.271	0.464	0.308	0.337	0.283	0.337
PL-Den	0.353	0.489	0.293	0.297	0.239	0.489	0.246	0.366	0.591	0.366
PL-pat	0.258	0.344	0.366	0.362	0.210	0.331	0.109	0.330	0.393	0.330
P-Scat	0.399	0.536	0.326	0.304	0.278	0.418	0.222	0.297	0.447	0.297

Table 3. Comparison of CLIP and DinoV2 on PICD.

Model	$avg\text{-}CDA_1$	$avg\text{-}CDA_2$	$map@100$	$Silhouette$	DBI
CLIP	0.571	0.404	0.322	-0.008	8.556
DinoV2	0.509	0.390	0.316	0.004	12.489

3) Consistency in Line Width: The line width should not vary dramatically. The width difference between the two ends must be minimal. For example, as shown in 2b, the narrow side of the line should be less than $1/4$ of the narrow side of the image.

4) Minimum Thickness: The line element should not be overly thin. Its width must be at least $1/30$ of the length of the image.

5) No Interference from Other Shape Elements: The overall shape of the line must not be disrupted by other shape elements, though combinations of points and lines are allowed.

6) Formed by Multiple Point Elements: A line element can be formed by multiple point elements arranged in a linear manner. Specifically, four or more point elements are required to constitute a valid line element.

(a) (b) (c)

Figure 2. Example images illustrating criteria 2), 3), and 6) for Line Elements: (a) an example of Criterion 2, demonstrating that the long side of the line element should be sufficiently long; (b) an example of Criterion 3, showing that the width of the line element should not vary significantly; (c) an example of Criterion 6, indicating that multiple point elements can form a line element.

Shape Element Objects with large areas or clearly segmented regions are treated as shape elements[3, 7, 17, 21].

1) The aspect ratio of the element satisfies that the short side is more than one-third of the long side in length.

2) The subject occupies more than $1/6$ of the image area.

2.2. Composition Categories

1. Single Point RoT(P-RoT): A single point element is positioned at one of the intersection points of the Rule of Thirds Grid, as shown in Fig. 5a, emphasizing the concentration of the visual focus. Sample images can be seen in Fig. 3.

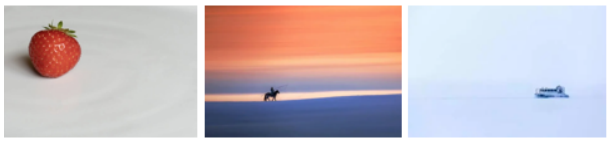
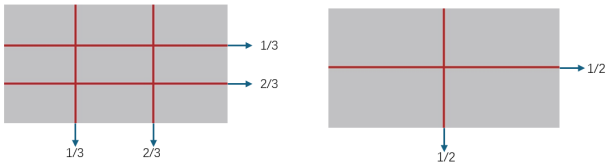


Figure 3. Sample images from Single Point RoT(P-RoT) Category

2. Single Shape RoT (S-RoT): A large, prominent object is positioned on either the left or right side of the frame, with its vertical bisector aligned with one of the vertical lines of the Rule of Thirds Grid in Fig. 5a. More samples can be seen in Fig. 4.

Figure 4. Sample images from Single Shape RoT (S-RoT) Category.

3. Centered Single Point (P-Cent): A prominent object is positioned at the center point of Symmetry Grid, as shown in Fig. 5b, forming a clear focal point and drawing the viewer's attention. This composition often highlights a single subject with a strong sense of dominance. Sample images are in Fig. 6.



(a) Rule of Thirds Grid (b) Symmetry Grid

Figure 5. Sketch map for The Rule of Thirds Grid and the Symmetry Grid.

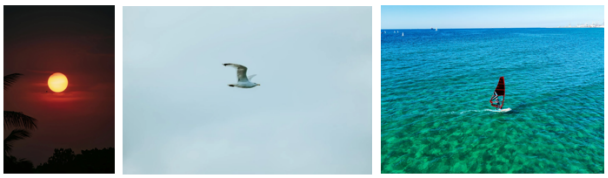


Figure 6. Sample images from Centered Single Point(P-Cent) Category

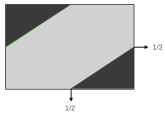
4. Centered Single Shape (S-Cent): A shape element is located at the center of the image, where its vertical bisector intersects with the vertical line of the symmetry grid, ensuring a balanced composition. Sample images are shown in Fig. 7.

Figure 7. Sample images from Centered Single Shape (S-Cent) Category

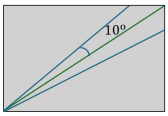
5. Diagonal Multi-points (P-Dia): A few points (up to three) are arranged along the diagonal, guiding the viewer's gaze and enhancing the dynamic feel of the image. The diagonal formed by these elements exceeds half the length of the screen's diagonal, with the direction limited to a single axis (either top-left to bottom-right or vice versa).

Multiple diagonals in the same direction are allowed. Similarly, point elements should fall within the light gray area shown in Fig. 8a, positioned near the center of the frame. The arrangement of points must align with the screen's diagonal direction, with an allowable deviation of

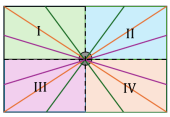
no more than 10°, as illustrated in Fig. 8b. Other points or lines should not interfere with the diagonal's flow, though multiple diagonals in the same direction are permissible. Sample images can be seen in Fig.9.



(a) Illustration of element location ranges for LS-Dia and P-Dia.



(b) Illustration of element Angle range for LS-Dia and P-Dia.



(c) Four divisional quadrants for LS-Dia and S-Per

Figure 8. An illustration of the image selection rules in the LS-Dia, P-DIA, and S-Per categories. (a) and (b) is valid under vertical-axis symmetry.

Figure 9. Sample images from Diagonal Multi-points (P-Dia) Category

6. Diagonal Lines/Shapes (LS-Dia): The composition divides the image into two distinct sections using a diagonal line or shape element, evoking a sense of motion and tension while directing the viewer's gaze along the diagonal trajectory. To maintain the integrity of the diagonal composition, no additional lines, shapes, or points should disrupt the primary diagonal trend. Multiple diagonals aligned in the same direction are allowed, provided they maintain consistent orientation.

If the lines exhibit multiple directions, they must only appear within either the II & IV quadrant or the I & III quadrant, as illustrated in Fig. 8c. Configurations spanning other combinations, such as I&II, I&II&III are excluded from diagonal composition, distinguishing it from perspective composition. Sample images are provided in Fig. 10.

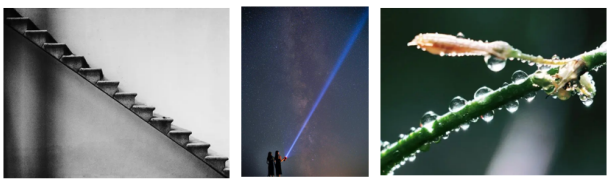
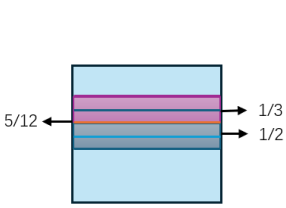


Figure 10. Sample images of Diagonal Lines/Shapes(LS-Dia)

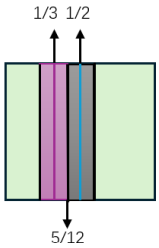
7. Horizontal–Arranged Points (P-Hori): A few points (less than or equal to three) are arranged horizontally, guiding the viewer's gaze to move horizontally. Sample images can be seen in Fig. 11.

Figure 11. Sample images from Horizontal–Arranged Points(P-Hori) Category

8. Horizontal Middle Part (LS-Hori2): A distinct horizontal line is located in the grey region in Fig. 12a and divides the image into two parts, emphasizing the symmetry between the top and bottom sections. Sample images can be seen in Fig. 13.



(a) Horizontal Division Grid. The grey area represents the intended position for the horizontal line in the LS-Hori2 category, while the pink area represents the intended position for the horizontal line in the LS-Hori3 category.



(b) Vertical Division Grid. The grey area represents the intended position for the vertical line in the L-Ver2 category, while the pink area represents the intended position for the vertical line in the L-Ver3 category.

Figure 12. Horizontal and Vertical Division Grid of the image

Figure 13. Sample images of Horizontal Middle Part (LS-Hori2)

9. Horizontal Third Part (LS-Hori3): A horizontal line positioned within the pink region of Fig. 12a creates a harmonious visual composition. Sample images can be seen in Fig. 14.

10. Horizontal Arranged Shapes (S-Hori): Multiple prominent objects (number of object $\leq$ 6) are evenly dis-

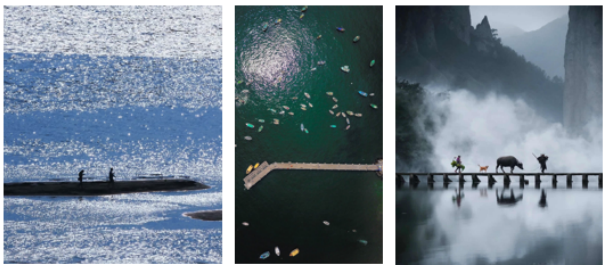


Figure 14. Sample images of Horizontal Third Part (LS-Hori3)

tributed along the horizontal direction. Sample images can be seen in Fig. 15.

Figure 15. Sample images from Horizontal Arranged Shapes(S-Hori) Category

11. Vertical Multi-points (P-Ver): The point element serves as the main subject of the image, and the points must not be placed too close to one another. The line connecting the centroids of the point elements must pass through the vertical line, allowing for some deviation, and form an angle $\leq 10^{\circ}$ with the vertical line.Sample images can be seen in Fig. 16.

Figure 16. Sample images from Vertical Multi-points(P-Ver) Category

12. Vertical Middle Line (L-Ver2): The line element is located within the grey region in Fig. 12b. Sample images can be seen in Fig. 17.

13. Vertical Third Line (L-Ver3): At least one line element is located in the pink region in Fig. 12b. Sample images can be seen in Fig. 18.

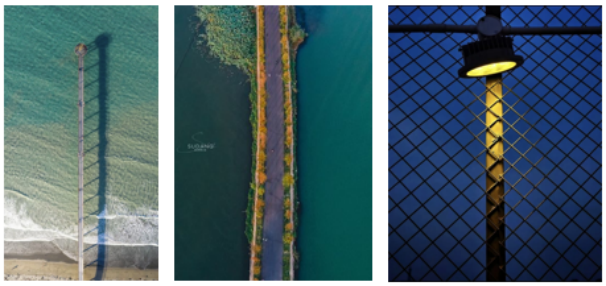


Figure 17. Sample images from L-Ver2 Category

Figure 18. Sample images from Vertical Middle Line (L-Ver3) Category

14. Vertical Multi-lines (L-VerMul): Multiple vertical lines (usually more than three but fewer than six) are evenly or unevenly distributed within the image, dividing it into multiple vertical sections and forming a complex, structured composition. Sample images can be seen in Fig. 19.

Figure 19. Sample images of Vertical Multi-lines(L-VerMul)

15. Three Points Triangle (P-Tri): Three points form a triangular arrangement, creating a sense of visual balance. The boundaries between points are not fuzzy and can be clearly distinguished. Sample images can be seen in Fig. 20.

Figure 20. Sample images of Three Points Triangle (P-Tri)

16. Line/Shape-Triangle (LS-Tri): The outer outline of the element presents a prominent triangle, and the area of

the triangle is greater than 1/6 of the screen. Sample images can be seen in Fig. 21.

Figure 21. Sample images from Line/Shape-Triangle (LS-Tri) Category

17. C-Curve (LS-C-Cur): The outer outline of the main body of the picture is circular, and the main body area accounts for $\geq 1/3$ of the size of the picture, (significant enough) as follows.

And the body is $\leq 3/4$ of the whole circle (sufficiently incomplete). The arc of the c shape should be complete, sample images can be seen in Fig. 22.

Figure 22. Sample images from C-Curve (LS-C-Cur) Category

18. O-Curve (LS-O-Cur): The outer contour of the main subject in the image is circular, with the subject area occupying $\geq 1/3$ of the image size (ensuring sufficient prominence). Additionally, the subject forms $> 3/4$ of a complete circle (ensuring adequate completeness). Radiological compositions are excluded.Sample images can be seen in Fig. 23.

Figure 23. Sample images from O-Curve (LS-O-Cur) Category

19. S-Curve (LS-S-Cur): The outer outline of the main body of the picture is S-shaped, and the obvious bending times of the curve are less than or equal to 8 and greater than or equal to 2. The horizontal distance between the leftmost and rightmost points of the "S" shape is greater than or equal to 1/3 of the corresponding side of the frame. Sample images can be seen in Fig. 24.

Figure 24. Sample images from S-Curve(LS-S-Cur) Category.

20. Diffuse (LS-Dif): The elements in the picture take the form of radiating outward from a point, the radiating subject is larger than 1/3 of the picture and the radiating center is inside the picture (including the edge). Sample images can be seen in Fig. 25.

Figure 25. Sample images from Diffuse(LS-Dif) Category

21. Perspective (S-Per): Using linear perspective to create a sense of spatial depth, usually including a vanishing point. Elements in the image shrink or deform according to perspective rules, enhancing depth perception and guiding the viewer's gaze toward the vanishing point. It occupies at least 3 areas out of the four regions a, b, c, and d of Fig. 8c. Sample images can be seen in Fig. 26.

Figure 26. Sample images from Perspective (S-Per) Category.

22. Random-Dense(PL-Den): Multiple small objects are randomly but densely distributed in the image, enhancing its visual complexity and diversity. Individual objects are identifiable. Sample images can be seen in Fig. 27.

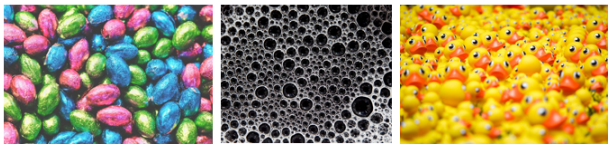


Figure 27. Sample images from Random-Dense (PL-Den) Category

23. Pattern (PL-Pat) : Multiple similar small objects or lines are arranged in a regular pattern, producing a dense yet orderly visual effect. The same element is repeated across the image, with the number of elements satisfying ≥ 6 (applicable to both points and lines). The arrangement is regular, with clear boundaries between elements, ensuring that multiple objects are not mistakenly perceived as a single entity. Sample images can be seen in Fig. 28.

Figure 28. Sample images from Pattern(PL-Pat) Category.

24. Scatter (P-Scat): Multiple point elements are sparsely and irregularly distributed across the image, creating a loose yet harmonious visual structure. The number of point elements is ≥ 6, with the image maintaining a balanced yet non-uniform distribution. Adequate spacing must be preserved between elements. Example images are shown in Fig. 29.

Figure 29. Sample images from Scatter(P-Scat) Category.

3. Images with multiple labels in PICD

Despite the efforts to ensure unique label design in PICD, which aims to maximize the distinctiveness between composition categories, some images inevitably belong to multiple categories. In PICD, the number of labels for such images does not exceed 2. Table 4 summarizes the specific multi-label categories and their corresponding counts.

Table 4. Distribution of multi-label data in the PICD dataset.

Multi-labels	Number
LS-Hori2 + L-Ver2	33
LS-Hori2 + L-Ver3	35
LS-Hori2 + P-Cent	16
LS-Hori3 + L-Ver3	69
L-Ver3 + P-RoT	5
Total	**158**

References

[1] 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.

[2] Zhe Chen, Jiannan Wu, Wenhai 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.

[3] Michael Freeman. *The Photographer's Eye Digitally Remastered 10th Anniversary Edition: Composition and Design for Better Digital Photos*. Routledge, 2017.

[4] 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.

[5] 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.

[6] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

[7] Wassily Kandinsky. *Point and line to plane*. Dover Publications, 1979.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al.

Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

[16] 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.

[17] Yvonne Spielmann. The visual flow: Fixity and transformation in photo-and videographic imagery. *Heterogeneous Objects: Intermedia and Photography after Modernism*, pages 106–107, 2013.

[18] 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.

[19] 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.

[20] 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.

[21] Richard D Zakia. Photography and visual perception. *Journal of Aesthetic Education*, 27(4):67–81, 1993.

[22] 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.

[23] Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. *arXiv preprint arXiv:2104.03133*, 2021.

[24] 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.