Cropper: Vision-Language Model for Image Cropping through In-Context Learning

Supplementary Material

This supplementary material provides:

- Sec. A: We present the implementation details, additional comparison results, additional qualitative results.
- Sec. B: We present the implementation details, additional ablation study, and additional qualitative results for subject-aware cropping task.
- Sec. C: We present the implementation details, additional ablation study, and additional qualitative results for aspect ratio-aware cropping task.
- Sec. D: We present details about user study.

A. Free-form Cropping

A.1. Implementation details

We show the prompt for zero-shot cropping using Gemini 1.5 Pro [8] in Tab. 1.

Prompt & Output	**Instruction**
Initial Prompt	Localize the aesthetic part of the image. (x_1, y_1, x_2, y_2) represents the region. x_1 and x_2 are the left and right most positions, normalized into 1 to 1000, where 1 is the left and 1000 is the right. y_1 and y_2 are the top and bottom positions, normalized into 1 to 1000 where 1 is the top and 1000 is the bottom. Please propose a new region (x_1, y_1, x_2, y_2)
Output	$(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$

Table 1. Prompt for zero-shot cropping with Gemini 1.5 Pro [8].

Tab. 2 shows further comparison on the FCDB [1] dataset for free-form cropping.

A.2. Additional qualitative results

Iterative update. We showcase some intermediate results of the iterative refinement in Fig. 1. Our method progressively refines the predicted crops, achieving increasing accuracy and better overlap with the ground-truth cropping box in each iteration.

Qualitative comparison. We present additional results in Fig. 2. Our method generates better visual pleasing crops.

Model	Training-Free	Training Set	IoU ↑	Disp ↓
A2RL [3]	✗	AVA	0.663	0.089
A3RL [4]	✗	AVA	0.696	0.077
VPN [12]	✗	CPC	0.711	0.073
VEN [12]	✗	CPC	0.735	0.072
ASM [9]	✗	CPC	0.749	0.068
GAIC [13]	✗	GAICD	0.672	0.084
CGS [5]	✗	GAICD	0.685	0.079
TransView [6]	✗	GAICD	0.682	0.080
Chao et al. [11]	✗	GAICD	0.695	0.075
Cropper (Ours)	✓	GAICD	0.667	0.087

Table 2. Quantitative comparison among different methods for free-form image cropping on the FCDB [1] dataset. Cropper shows competitive performance as a *training-free* approach.

B. Subject-aware Cropping

B.1. Prompts

We show the details of the prompts for the subject-aware cropping in Tab. 3. The goal is to get accurate coordinates of the crop $(\hat{\hat{x}}_1, \hat{\hat{y}}_1, \hat{\hat{x}}_2, \hat{\hat{y}}_2)$. In the initial prompt, we use 30 in-context (ICL) examples for image cropping for 10 iterations. 10 examples are ranked by the scorer and we use top-5 crops for our task, the format of image i's j-th crop is defined as $(x_1^{i,j}, y_1^{i,j}, x_2^{i,j}, y_2^{i,j})$. Intermediate results of initial prompt are coordinates of 5 crops. Subsequently, the crop is iteratively refined by accumulating the context into prompts, using refinement prompt. Note that scorer is "VILA+Area".

B.2. Ablation study of scores

We show the ablation study of scorer on the subject-aware cropping in Tab. 4. With "VILA+Area", our proposed method achieves the best performance.

B.3. Additional qualitative results

We showcase more results in Fig. 3. Our method demonstrates subject awareness, enabling the generation of high-quality cropped images.

C. Aspect-ratio aware Cropping

C.1. Prompts

We show the details of the prompts for the aspect ratio-aware cropping in Tab. 5. For this task, we use the following hyperparameter: number of in-context learning

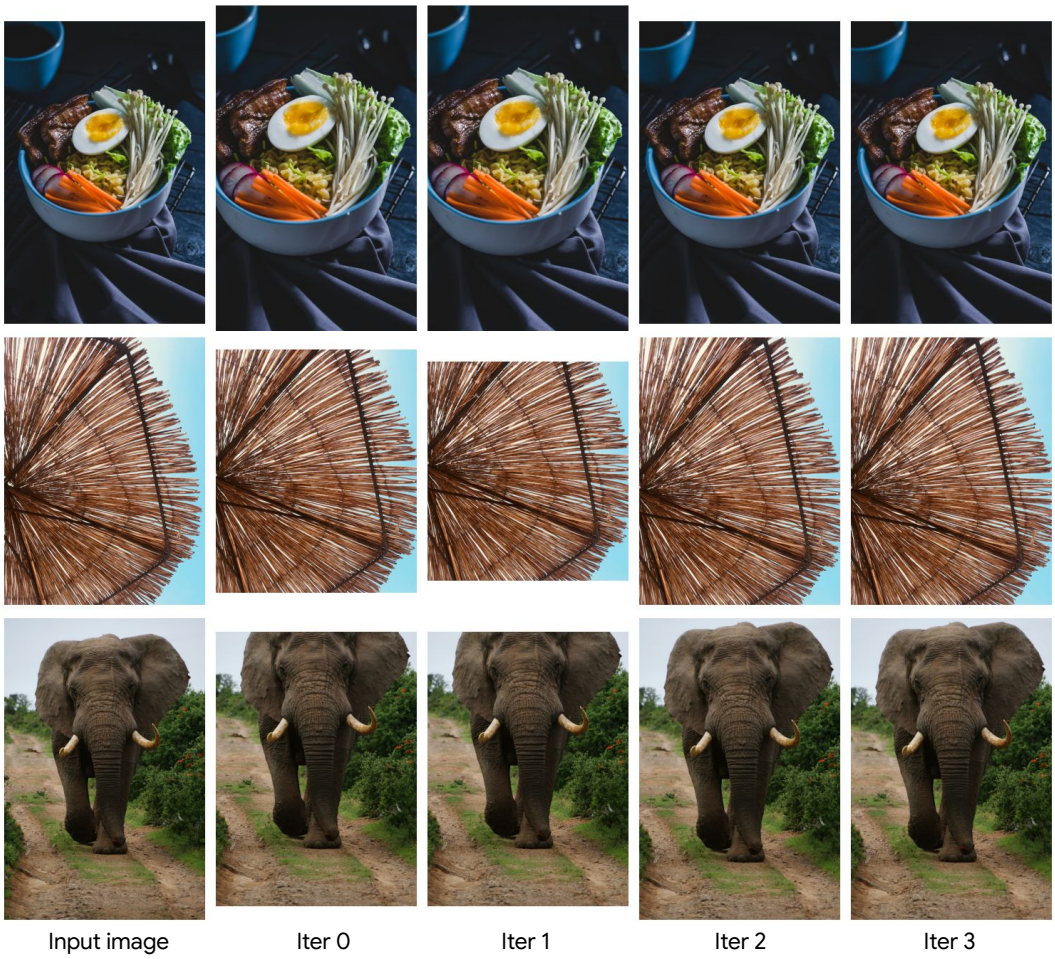

Figure 1. Results from each iteration for free-form cropping using Cropper. The iteration process demonstrated progressive convergence, resulting in improved crop quality. All input images are from Unsplash [10].

examples$S = 10$, number of crops $R = 6$, number of iteration $L = 2$, temperature = 0.05.

C.2. Ablation study of scores

We show the ablation study of scores on the aspect ratio-aware cropping in Tab. 6. With CLIP score only, our proposed method achieves the best performance.

D. User study

We include the instructions for users as follows:

- Your Task: Carefully analyze the source image and the two output images and SELECT one output.
- Content: This refers to the key elements and objects in the image, such as people, buildings, or other recognizable features. The output should keep the important details of these objects as close to the original as possible.
- Aesthetics: The output has a sense of aesthetics. It follows common human natures with proper layout. Select the output that not only preserves the content best and fits the aesthetics.
- Your Goal: Select the image that, overall, looks the most natural and visually appealing to the source image.

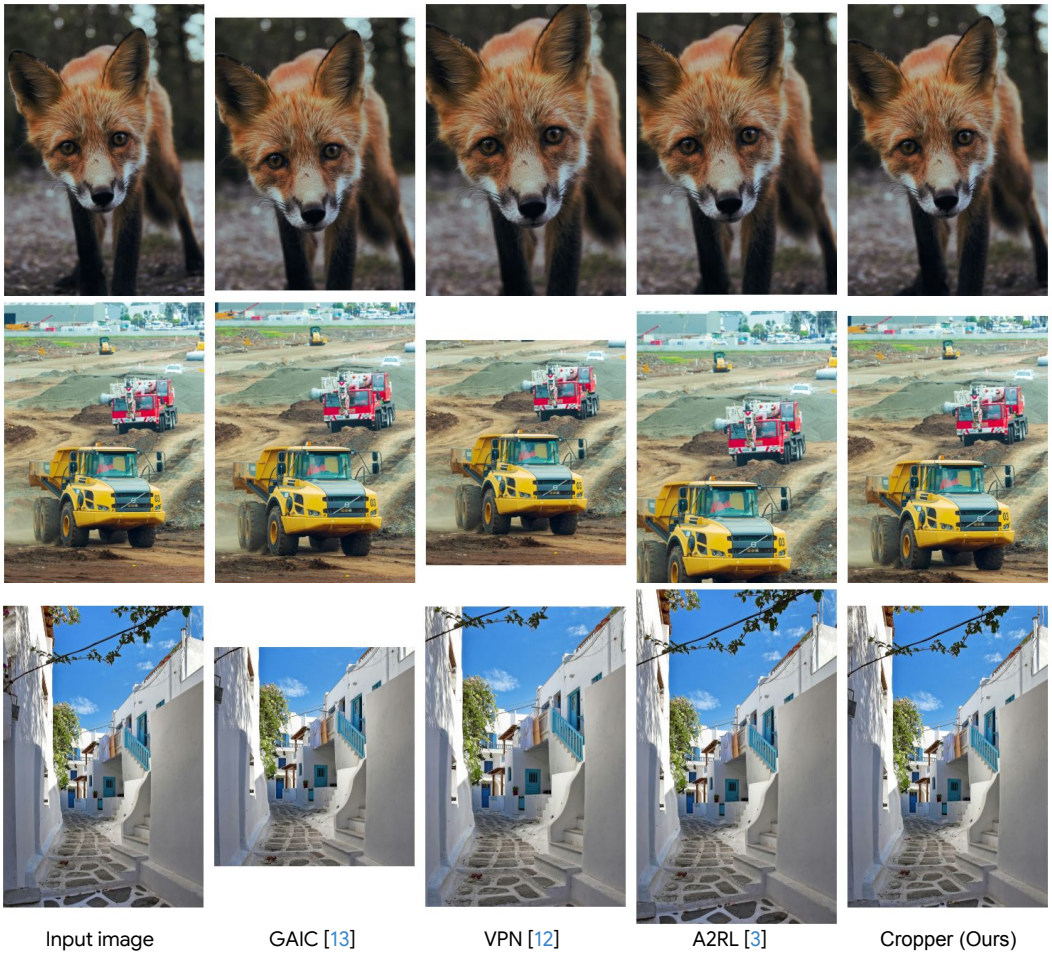

Figure 2. Comparing with GAIC [13], VPN [12] and A2RL [3] for free-form cropping on images from Unsplash [10].

Figure 3. Qualitative results of subject-aware cropping. The result shows that our method can generate crops on different subjects. The input image is from Unsplash [10].

Prompt & Output	Instruction
Initial Prompt	Find visually appealing crop. Each region is represented by (x_1, y_1, x_2, y_2) coordinates. x_1, x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1, y_2 are the top and bottom positions, normalized into 0 to 1 where 0 is the top and 1 is the bottom. {image 1} $((c_x^1, c_y^1), x_1^1, y_1^1, x_2^1, y_2^1)$, {image 2}, $((c_x^2, c_y^2), x_1^2, y_1^2, x_2^2, y_2^2)$, ... {image S}, $((c_x^S, c_y^S), x_1^S, y_1^S, x_2^S, y_2^S)$, {Query image}, (c_x, c_y)
Output	$(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$
Iterative Crop Refinement Prompt	Localize aesthetic part of image. The region is represented by (x_1,y_1,x_2,y_2). x_1, x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1, y_2 are the top and bottom positions, normalized into 0 to 1 where 0 is the top and 1 is the bottom. We provide several images here. {Cropped image 1} Output: $(\hat{x}_1^1, \hat{y}_1^1, \hat{x}_2^1, \hat{y}_2^1)$ {Cropped image 2} Output: $(\hat{x}_1^2, \hat{y}_1^2, \hat{x}_2^2, \hat{y}_2^2)$ ... {Cropped image R} Output: $(\hat{x}_1^R, \hat{y}_1^R, \hat{x}_2^R, \hat{y}_2^R)$ Propose different crop. The region should be represented by (x_1,y_1,x_2,y_2). Output:
Output	$(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$

Table 3. VLM prompt used for subject-aware cropping.

Prompt & Output	Instruction
Initial Prompt	Find visually appealing crop. Give the best crop in the form of a crop box and make sure the crop has certain width:height. Box is a 4-tuple defining the left, upper, right, and lower pixel coordinate in the form of (x_1, y_1, x_2, y_2). Here are some example images, its size, and crop w:h triplets and their corresponding crops. {image 1}, size (w_1, h_1), crop ratio (r_1), output $(x_1^1, y_1^1, x_2^1, y_2^1)$, { image 2 }, size (w_2, h_2), crop ratio (r_2), output $(x_1^2, y_1^2, x_2^2, y_2^2)$, ... {image S}, size (w_S, h_S), crop ratio (r_S), output $(x_1^S, y_1^S, x_2^S, y_2^S)$, {Now Give the best crop in the form of a crop box for the following image. Give R possible best crops.} {Query image}, size (w, h), crop ratio (r)
Output	$(\hat{x}_1^1, \hat{y}_1^1, \hat{x}_2^1, \hat{y}_2^1)$,$(\hat{x}_1^2, \hat{y}_1^2, \hat{x}_2^2, \hat{y}_2^2)$, ..., $(\hat{x}_1^T, \hat{y}_1^T, \hat{x}_2^T, \hat{y}_2^T)$
Iterative Crop Refinement Prompt	Initial Prompt + Example Image: {Query image}; Crop ratio: r; Example output: {Cropped image 1} $(\hat{x}_1^1, \hat{y}_1^1, \hat{x}_2^1, \hat{y}_2^1)$, {Cropped image 2} $(\hat{x}_1^2, \hat{y}_1^2, \hat{x}_2^2, \hat{y}_2^2)$ ... {Cropped image R} $(\hat{x}_1^R, \hat{y}_1^R, \hat{x}_2^R, \hat{y}_2^R)$ Propose a different better crop with the given ratio. Output:
Output	$(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$

Table 5. VLM prompt used for aspect ratio-aware cropping.

VILA [2]	Area	CLIP [7]	IoU ↑	Disp ↓
✓	✗	✗	0.753	0.0413
✗	✓	✗	0.755	0.0402
✗	✗	✓	0.749	0.0417
✓	✓	✗	**0.769**	**0.0372**
✓	✗	✓	0.751	0.0401
✗	✓	✓	0.754	0.0394
✓	✓	✓	0.766	0.0379

Table 4. Ablation study for scores on the subject-aware cropping. Cropper achieves the best performance with VILA [2] + Area score.

VILA [2]	Area	CLIP [7]	IoU ↑	Disp ↓
✓	✗	✗	0.718	0.0631
✗	✓	✗	0.713	0.0630
✗	✗	✓	**0.756**	**0.0529**
✓	✓	✗	0.716	0.0630
✗	✓	✓	0.741	0.0562
✓	✗	✓	0.742	0.0560
✓	✓	✓	0.729	0.0588

Table 6. Ablation study for aspect ratio-aware cropping task. Comparison of combinations of VILA [2], Area, and CLIP [7] components shows that the CLIP-only configuration achieves the best IoU and Disp values.

References

[1] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *WACV*, 2017. 1

[2] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *CVPR*, 2023. 4

[3] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *CVPR*, 2018. 1, 3

[4] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *TIP*, 28(10), 2019. 1

[5] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *CVPR*, 2020. 1

[6] Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong. Transview: Inside, outside, and across the cropping view boundaries. In *ICCV*, 2021. 1

[7] 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 *ICML*, 2021. 4

[8] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[9] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *AAAI*, 2020. 1

[10] Unsplash Website. Unsplash. Accessed: March 21, 2025, URL: https://unsplash.com/. 2, 3

[11] Chao Wang, Li Niu, Bo Zhang, and Liqing Zhang. Image cropping with spatial-aware feature and rank consistency. In *CVPR*, 2023. 1

[12] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *CVPR*, 2018. 1, 3

[13] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *TPAMI*, 44(3), 2020. 1, 3