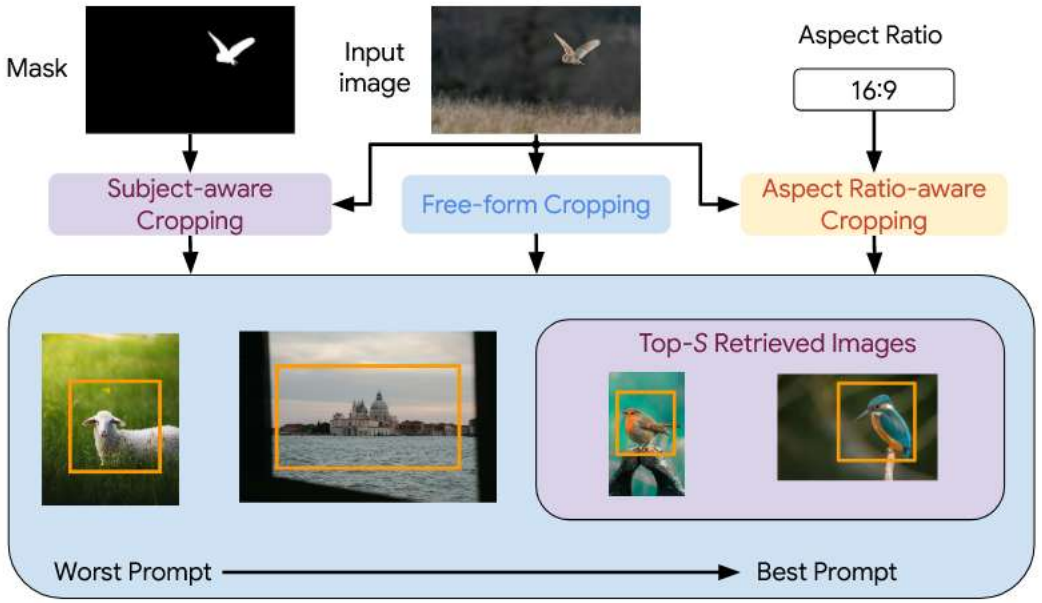
Cropper: Vision-Language Model for Image Cropping through In-Context Learning

Seung Hyun Lee, Jijun Jiang, Yiran Xu, Zhuofang Li, Junjie Ke, Yinxiao Li, Junfeng He, Steven Hickson, Katie Datsenko, Sangpil Kim, Ming-Hsuan Yang, Irfan Essa, Feng Yang
Google DeepMind, Google Research, Google, University of Michigan, University of Maryland, Korea University

arXiv preprint arXiv:2408.07790 (2024)

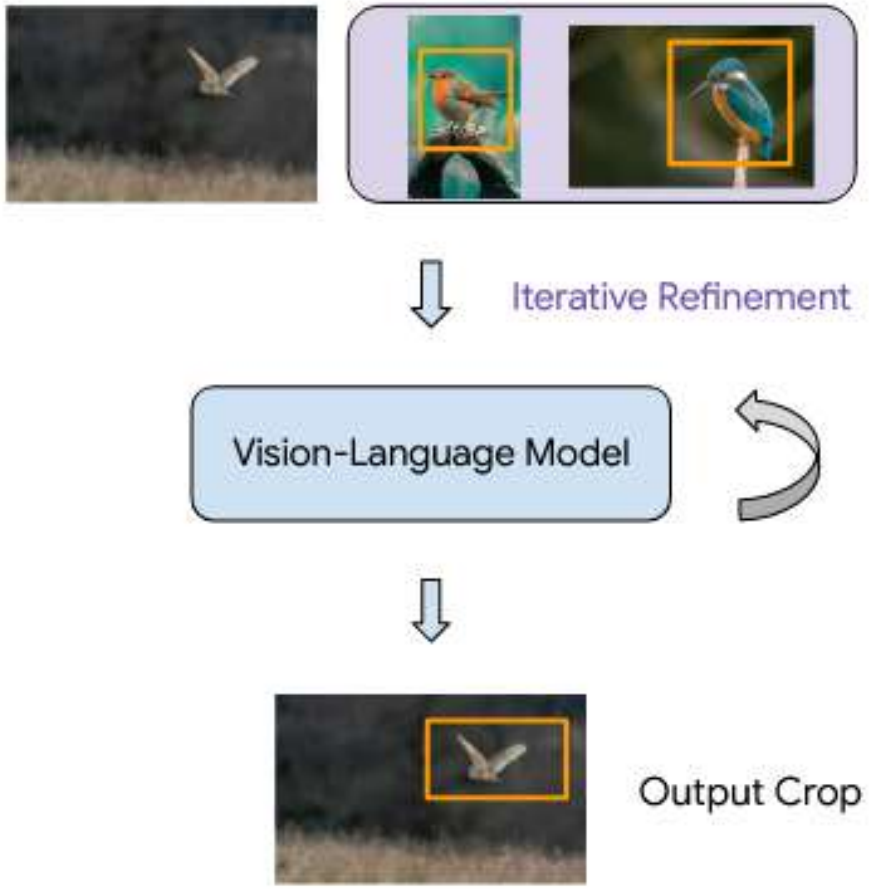
Introduction

- The existing cropping methods usually train neural networks on images and real cropping, but most of these methods rely on specially designed networks and features, which are difficult to generalize effectively.
- For specialized cropping tasks, such as using subject masks for subject-aware cropping or using target aspect ratios for aspect ratio-aware cropping, it is usually necessary to redevelop and train specialized networks
- Recent advancements in large vision-language models (VLM), such as GPT-4o and Gemini, have unlocked new potential for various vision tasks, although users are often unable to fine-tune the VLM for downstream tasks.
- Although effectively adapting large blackbox models for downstream tasks is very difficult, in-context learning (ICL) ability is observed in large models.



Introduction

- We introduce a unified visual in-context-learning framework Cropper for image cropping tasks, including freeform, subject-aware, and aspect ratio-aware cropping.
- Our prompt retrieval strategy automates the effective selection of ICL examples for cropping tasks.
- The proposed iterative refinement strategy enables the model to progressively enhance the output crop.
- With a few in-context examples and no explicit training, Cropper surpasses the existing supervised learning methods across various benchmarks.



Cropper Overview

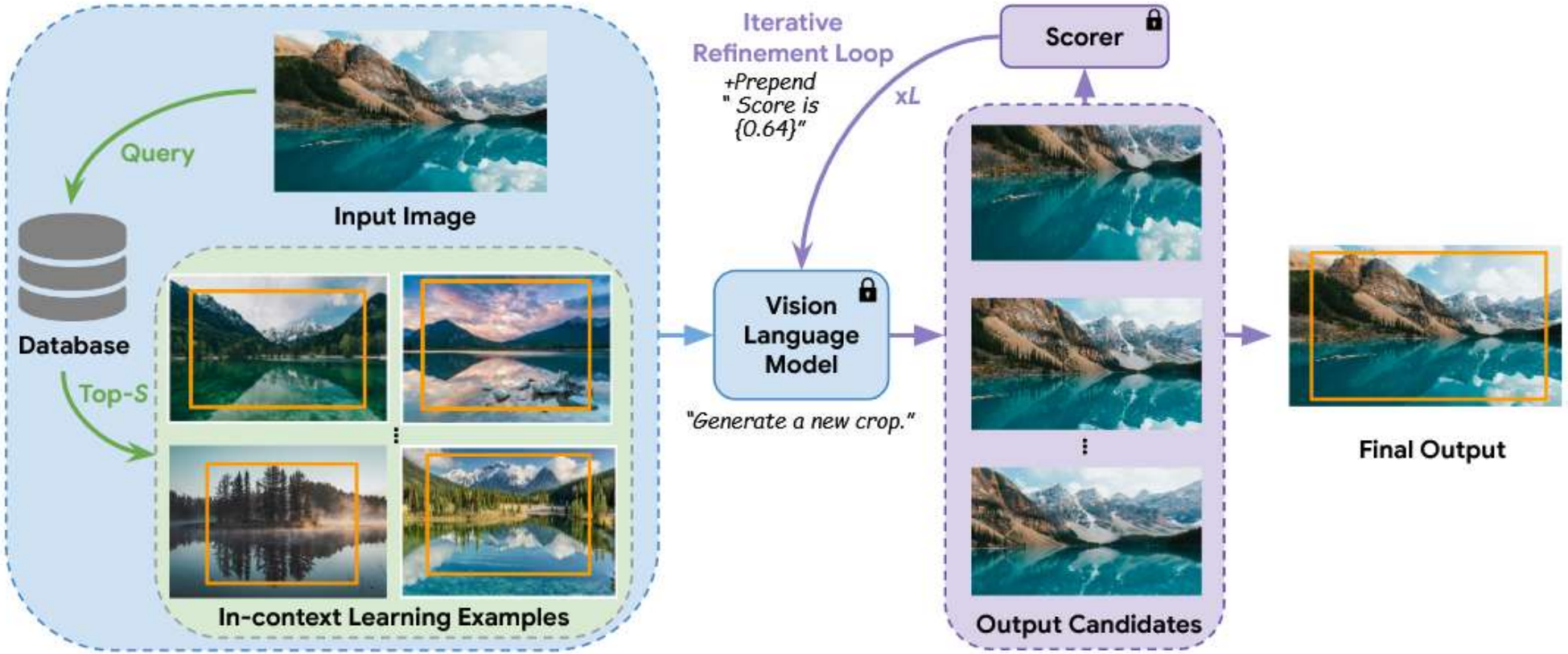


Figure 2. **Cropper Overview**. Cropper consists of two main steps: visual prompt retrieval and iterative crop refinement. Through visual prompt retrieval, top-S ICL examples are retrieved using an image similarity metric. In the iterative crop refinement stage, the VLM generates candidate crops based on these ICL examples and then these crops are subsequently scored by a scorer which measures aesthetics, content similarity, and area size. The VLM iteratively refines the crop candidates using the feedback from the scorer L times. All images are from Unsplash [30].

Visual Prompt Retrieval for Cropping

- Similar images are more likely to be cropped similarly. Thus, we aim to retrieve the **top-S** images and their most relevant ground-truth crops based on some similarity metric.

- Given an image query z_q and a dataset $\mathcal{D} = (z_i, C_i)_{i=1}^{M}$ containing M pairs of image z_i and crop ground-truth C_i, where C_i contains multiple crops $c_i, \ldots, c_s$ for some datasets.

- $\mathcal{Z}$ represents the set of top-S relevant images selected from the dataset $\mathcal{D}$ based on the similarity metric $Q(z_q, z_i)$.

- $\mathcal{H} = (z_j, c_j)_{j=1}^{S}$ represents the selected incontext images z_j along with their most relevant T crop ground-truths based on metric $G(z_q, c_j)$.

given an image query z_q

$$\mathcal{D} = (z_i, C_i)_{i=1}^{M}$$

$$\mathcal{Z} = \underset{z_i \in \mathcal{D}}{\arg\max}\, Q(z_q, z_i), \quad |\mathcal{Z}| = S,$$

$$\mathcal{H} = \underset{c_j \in C_j}{\arg\max}\, G(z_q, c_j), \quad z_j \in \mathcal{Z}, \quad |\mathcal{H}| = T,$$

Visual Prompt Retrieval for Cropping

- **Free-form cropping** aims to identify the best crop without additional constraints regarding aspect ratio or target subject. Q corresponds to the cosine similarity between the input image z_q and each training example $z_i \in \mathcal{D}$. We use the MOS score as G for selecting the groundtruth crops.

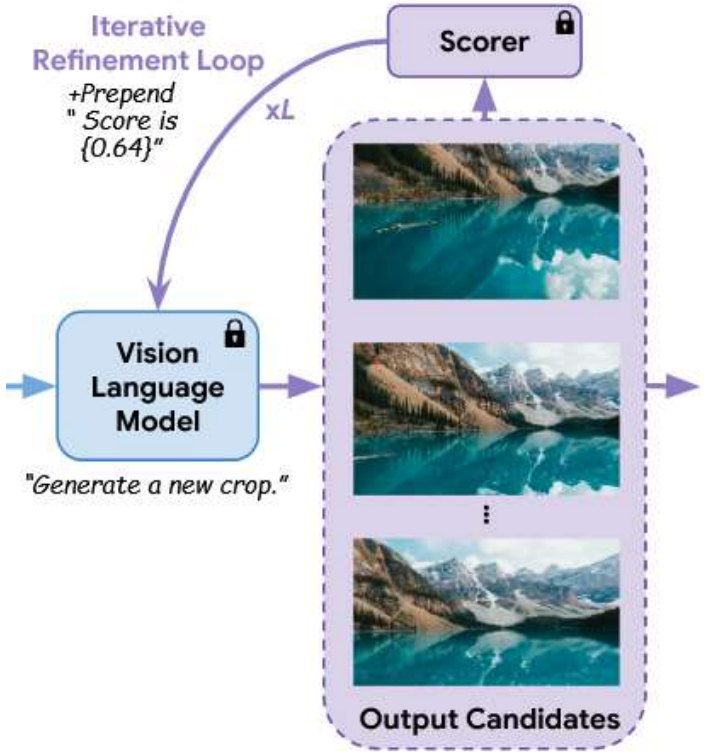
- **Subject-aware cropping** intends to identify an aesthetic crop containing the subject of interest, which is represented by binary masks provided by users. We first use CLIP image embedding similarity as Q for retrieving the top-S relevant images. G is defined as -L2 distance between the center points of the target mask m_q and mask from image $z \in Z$ to select the crop with closest masks.

- **Aspect ratio-aware cropping** requires the generated crop to conform to a specified aspect ratio r_q given the query image z_q. CLIP-based image similarity is adopted as Q. G is defined as the similarity between the crop c_i's aspect ratio and the target aspect ratio r_q.

Prompt & Output	Instruction
Initial Prompt	"Localize the aesthetic part of the image. (s, 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. s is MOS score. We provide several images here. {image 1} $(s_1^1, x_1^{1,1}, y_1^{1,1}, x_2^{1,1}, y_2^{1,1})$, $(s_1^2, x_1^{1,2}, y_1^{1,2}, x_2^{1,2}, y_2^{1,2})$, ..., $(s_1^T, x_1^{1,T}, y_1^{1,T}, x_2^{1,T}, y_2^{1,T})$, {image 2}, $(s_2^1, x_1^{2,1}, y_1^{2,1}, x_2^{2,1}, y_2^{2,1})$, $(s_2^2, x_1^{2,2}, y_1^{2,2}, x_2^{2,2}, y_2^{2,2})$, ..., $(s_2^T, x_1^{2,T}, y_1^{2,T}, x_2^{2,T}, y_2^{2,T})$, ... {image S}, $(s_S^1, x_1^{S,1}, y_1^{S,1}, x_2^{S,1}, y_2^{S,1})$, $(s_S^2, x_1^{S,2}, y_1^{S,2}, x_2^{S,2}, y_2^{S,2})$, ..., $(s_S^T, x_1^{S,T}, y_1^{S,T}, x_2^{S,T}, y_2^{S,T})$, {Query image},
Output	$(\hat{s}_1, \hat{x}_1^1, \hat{y}_1^1, \hat{x}_2^1, y_2^1)$,$(\hat{s}_2, \hat{x}_1^2, \hat{y}_1^2, \hat{x}_2^2, \hat{y}_2^2)$, ..., $(\hat{s}_R, \hat{x}_1^R, \hat{y}_1^R, \hat{x}_2^R, \hat{y}_2^R)$

Iterative Crop Refinement

- Generate R crop candidates.
- Crop the image according to each cropping proposal and feed the cropped images into scorers, such as VILA covering aesthetics, CLIP measuring content preserving, and area size, to obtain corresponding scores.
- In the refinement phase, we iteratively provide such feedback to the VLM by scoring the crop candidates and prompting it to generate new candidates to improve the score.
- This iterative process is repeated L times to generate the final output.



Iterative Crop Refinement Prompt	Initial Prompt + {Cropped image 1} $(\hat{s}_1, \hat{x}_1^1, \hat{y}_1^1, \hat{x}_2^1, \hat{y}_2^1)$, Score is {score 1} {Cropped image 2} $(\hat{s}_2, \hat{x}_1^2, \hat{y}_1^2, \hat{x}_2^2, \hat{y}_2^2)$, Score is {score 2} ... {Cropped image R} $(\hat{s}_R, \hat{x}_1^R, \hat{y}_1^R, \hat{x}_2^R, \hat{y}_2^R)$, Score is {score R} Propose similar crop that has high score. The region should be represented by (s, x_1, y_1, x_2, y_2).
Output	$(\hat{s}, \hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$

Experiment

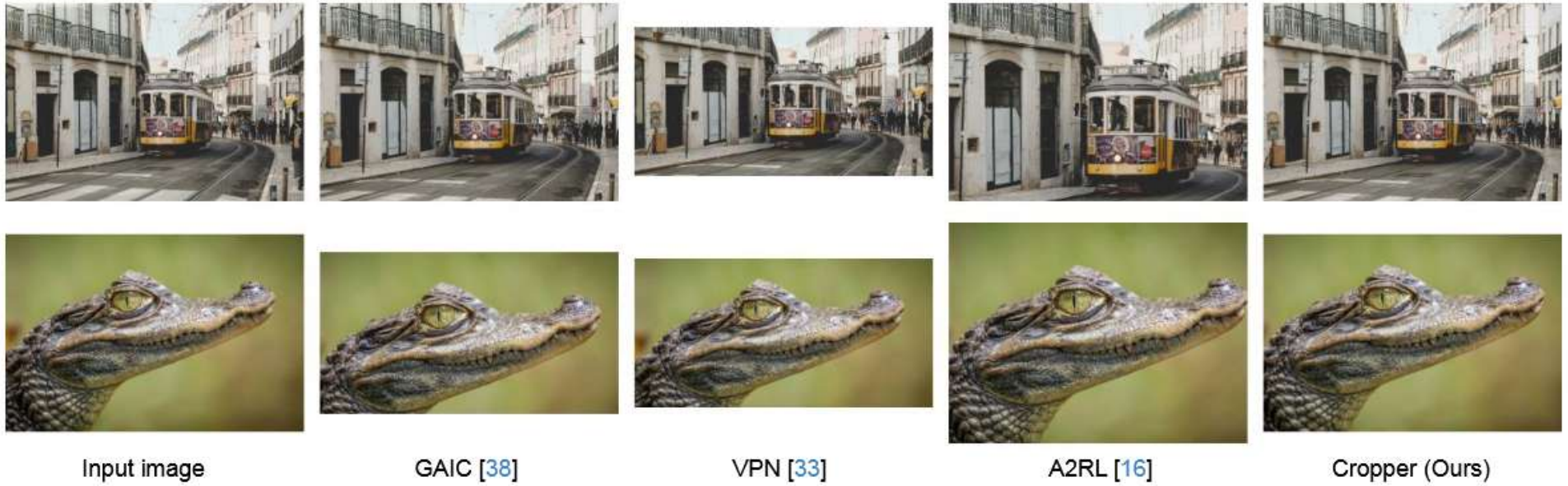


Figure 6. Qualitative comparing Cropper with GAIC [38], VPN [33], A2RL [16] on images from Unsplash [30] for free-form cropping.

Model	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$\overline{Acc}_5$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$\overline{Acc}_{10}$	$\overline{SRCC}$	$\overline{PCC}$
A2RL [16]	23.2	-	-	-	-	39.5	-	-	-	-	-	-
VPN [33]	36.0	-	-	-	-	48.5	-	-	-	-	-	-
VFN [5]	26.6	26.5	26.7	25.7	26.4	40.6	40.2	40.3	39.3	40.1	0.485	0.503
VEN [33]	37.5	35.0	35.3	34.2	35.5	50.5	49.2	48.4	46.4	48.6	0.616	0.662
GAIC [38]	68.2	64.3	61.3	58.5	63.1	84.4	82.7	80.7	78.7	81.6	0.849	0.874
CGS [19]	63.0	62.3	58.8	54.9	59.7	81.5	79.5	77.0	73.3	77.8	0.795	-
TransView [24]	69.0	66.9	61.9	57.8	63.9	85.4	84.1	81.3	78.6	82.4	0.857	0.880
Chao et al. [31]	70.0	66.9	62.5	59.8	64.8	86.8	84.5	82.9	79.8	83.3	0.872	**0.893**
Cropper (Ours)	**88.9**	**85.9**	**83.1**	**79.4**	**84.3**	**98.2**	**97.2**	**96.4**	**94.3**	**96.5**	**0.904**	0.860

Table 2. Quantitative comparison with existing free-form cropping methods on the GAICD [38] dataset. Cropper demonstrates significant superiority over other baselines despite using only a few in-context learning examples and no explicit training.

Experiment

Model	Training-Free	IoU ↑	Disp ↓
A2RL [16]	✗	0.667	0.0887
VFN [5]	✗	0.669	0.0887
VPN [33]	✗	0.704	0.0699
VEN [33]	✗	0.691	0.0765
LVRN [20]	✗	0.696	0.0765
GAIC [38]	✗	0.712	0.0696
SAC-Net [36]	✗	0.767	0.0491
Cropper (Ours)	✓	**0.769**	**0.0372**

Table 3. Quantitative comparison on the SACD [36] dataset in subject-aware cropping task.

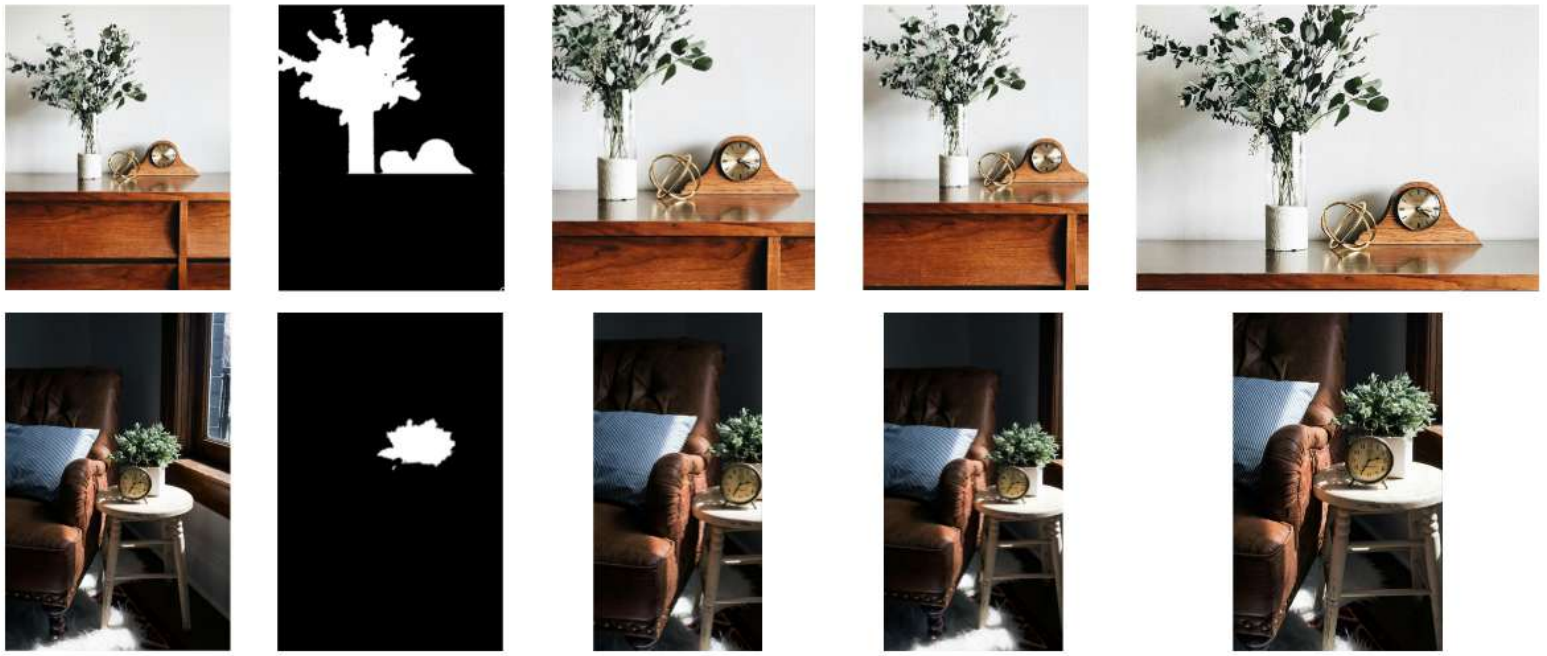


Figure 7. Visual comparisons on the subject-aware image cropping. Cropper preserves the important contents better than directly using VLMs, such as GPT-4o [1] and Gemini 1.5 Pro [26]. All input images are from Unsplash [30].

Experiment

Model	Training-free	IoU ↑	Disp ↓
GAIC [38]	✗	0.673	0.064
A2RL [16]	✗	0.695	0.073
VPN [33]	✗	0.716	0.068
Mars [18]	✗	0.735	0.062
Cropper (Ours)	✓	**0.756**	**0.053**

Table 4. Quantitative comparison on the FCDB dataset [4] for aspect ratio-aware cropping task. For other methods, we follow [18] to report the modified aspect ratio specified results, which are better than the ones in the original papers.

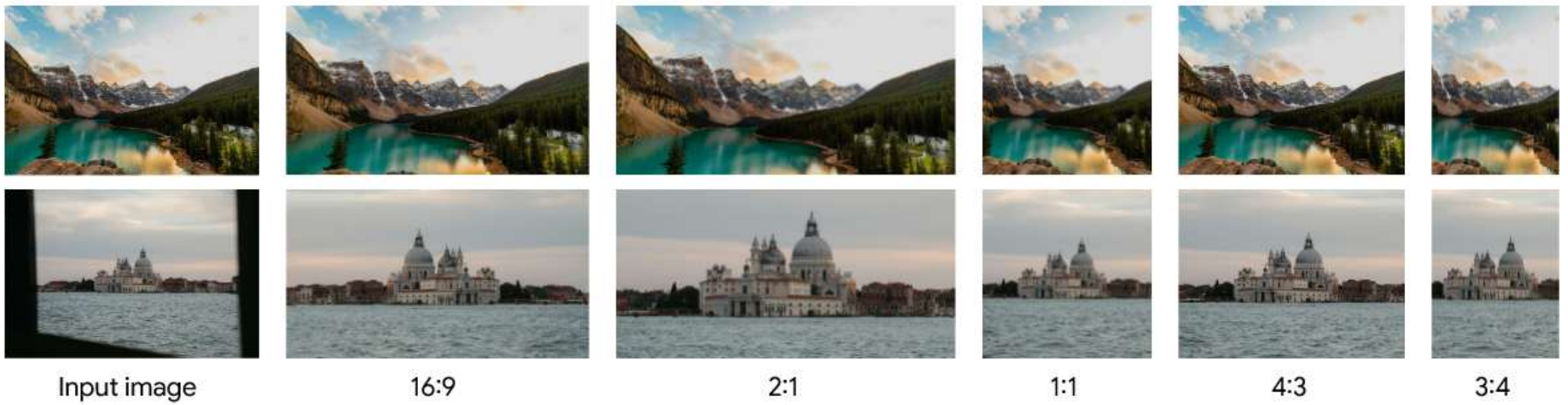


Figure 8. Example crops from Cropper for aspect ratio-aware cropping. This shows that our method can generate crops with the desired aspect ratios. All input images are from Unsplash [30].

Ablation Study

Scorer			Metrics					
VILA [15]	Area	CLIP [25]	IoU ↑	$\overline{Acc_5}$ ↑	$\overline{Acc_{10}}$ ↑	$\overline{SRCC}$ ↑	$\overline{PCC}$ ↑	Avg↑
✓	✗	✗	0.748	83.6	95.7	0.901	0.860	0.852
✗	✓	✗	0.752	83.9	96.0	0.882	0.838	0.854
✗	✗	✓	0.751	81.2	95.1	0.884	0.833	0.846
✓	✓	✗	0.748	**84.3**	**96.5**	0.904	0.860	**0.864**
✓	✗	✓	**0.754**	82.3	95.8	0.902	0.850	0.858
✗	✓	✓	0.752	83.9	96.1	0.902	0.858	0.862
✓	✓	✓	0.753	83.2	96.0	**0.907**	**0.869**	**0.864**

Table 5. Ablation study for different scorers on the GAICD [38] test dataset for free-form cropping.

Method	IoU ↑	Disp ↓
Zero-shot Gemini 1.5 Pro [26]	0.509	0.1385
Cropper with random retrieval	0.740	0.0660
Cropper with CLIP top-S	**0.748**	**0.0635**

Table 6. Ablation study for in-context learning. Comparing our methods with zero-shot Gemini 1.5 Pro [26], and Cropper with random retrieving in-context learning examples on the GAICD test set [38] for free-form cropping.

Dataset	Model	IoU ↑	Disp ↓
GAICD [38]	Cropper w/o Iter Refine.	0.722	0.0679
	Cropper (ours)	**0.748**	**0.0635**
FCDB [4]	Cropper w/o Iter Refine.	0.642	0.0925
	Cropper (ours)	**0.667**	**0.0865**

Table 7. Ablation study on iterative refinement.

	Free-form		Subject		Aspect-ratio	
Method	IoU ↑	Disp ↓	IoU ↑	Disp ↓	IoU ↑	Disp ↓
Highest score across all iters.	0.714	0.0843	0.760	0.0381	**0.756**	**0.0529**
From final iter.	**0.748**	**0.0635**	**0.769**	**0.0372**	0.714	0.0632

Table 8. Comparison of selection strategies for free-form cropping on GAICD test set [38], subject-aware cropping on the SACD [36] dataset, and aspect-ratio aware cropping on the FCDB dataset [4].

Model	$\overline{Acc_5}$↑	$\overline{Acc_{10}}$↑	$\overline{SRCC}$↑	$\overline{PCC}$↑	IoU↑	Avg↑
Cropper with Mantis-8B-Idefics2 [13]	80.2	88.6	0.874	0.797	0.672	0.806
Cropper with Gemini-1.5-flash [26]	**87.2**	**96.7**	0.805	0.758	**0.781**	0.837
Cropper with Gemini-1.5-pro [26]	84.3	96.5	**0.904**	**0.860**	0.748	**0.864**

Table 9. Comparing different vision-language models.

Thanks