Enhanced Product Information Retrieval via Image-to-Searchable Layout Reconstruction

Jihoon Jang[1,2*] **Donghyuk Choi**[2*] **Janghoon Choi**[1†]

Major in Data Science Convergence, Kyungpook National University, Daegu, Korea [1] Cutib Inc., Seoul, Korea [2]

1 Motivation

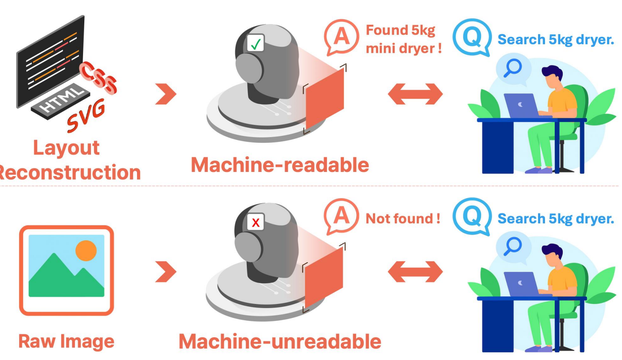

Fig. 1. A raw image is machine-unreadable, whereas a reconstructed layout is machine-readable.

- Online shops put key product details (**specs, prices, events**) **inside images**.
- Search engines (**SEO**) and AI engines (**GEO**) [1] **cannot read text in images**.
- So a user searches “5kg dryer”, but the answer is **“Not found!”**
- **Our idea:** Reconstruct the image as a **Searchable Layout**.

2 Objectives

Goal: Turn a product-page image into a **searchable and editable layout** that keeps its **text, style, and alignment**.

1. **New task:** Inverse layout reconstruction for dense e-commerce images, with word-level filtering, style, and alignment.
2. **ECTL dataset:** 3,055 real product pages from 20 categories, with word-level labels.
3. **Modular, OCR-agnostic pipeline:** Each module (OCR, VLM grouping, text removal, style) can be swapped — EasyOCR and PaddleOCR work **without retraining**.
4. **Better retrieval:** Our layouts beat image-only and OCR-only inputs on 3 models, and use **11–14% fewer tokens** than OCR-only.

3 Method

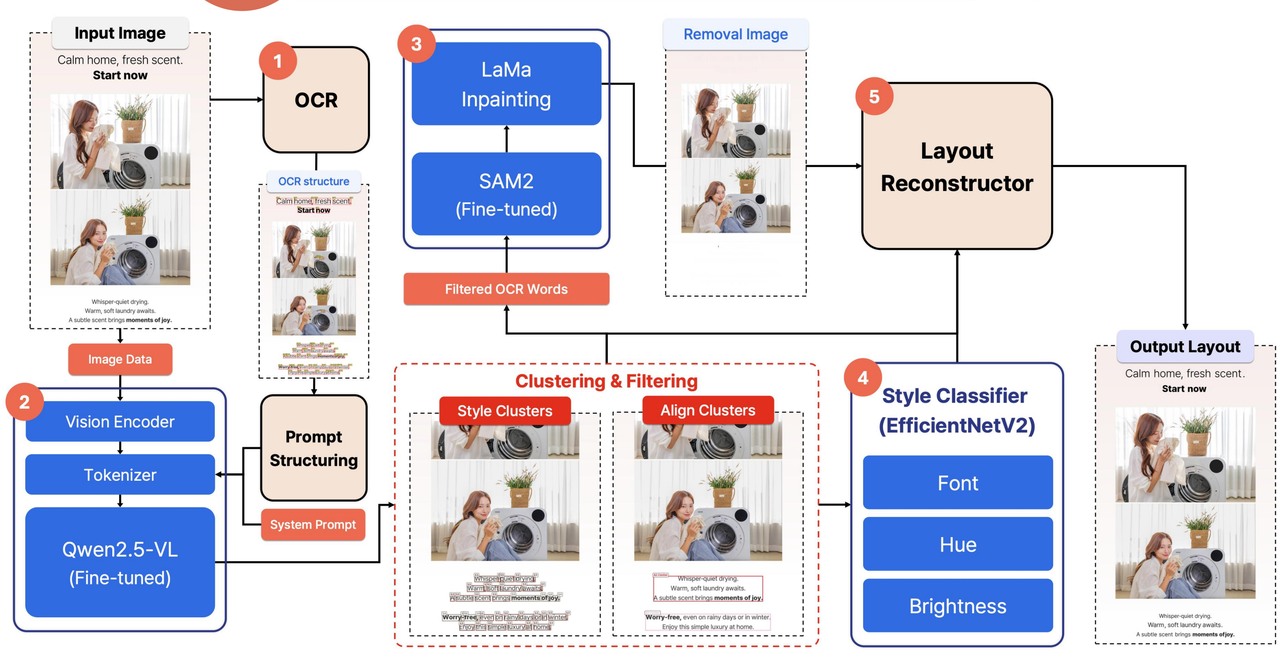

Fig. 2. Full pipeline of the proposed methodology. The numbers match the five steps below.

1. **OCR:** Google Cloud Vision reads the text as sentences → words → characters, and gives each word a **unique ID**.
2. **VLM clustering & filtering:** A fine-tuned Qwen2.5-VL (7B) [2] reads the image and the word IDs. In one output, it groups words by **style** and by **alignment** (left / center / right), and **drops decorative words**.
3. **Text removal:** A fine-tuned SAM2 [3] finds the exact text pixels, and LaMa [4] fills in a **clean background** behind them.
4. **Style classifier:** EfficientNetV2-S [5] predicts **font, hue, and brightness** for each word (Top-1 on synthetic data: 99.3% font, 99.7% color).
5. **Layout reconstructor:** Writes a structured layout — line blocks and word spans with inherited font and color (HTML in our implementation). The text is **real, editable, and searchable**. For example, in HTML (ours), for align groups and sentences, for words, with @font-face and inherited CSS.

E-Commerce Text Layout Dataset

3,055	20	47 → 35	500
Labeling data from 36,424 product images	**Product categories**	**Filter redundant words**	**Factual QA** pairs (100 images)

4 Experiments

QA-based Accessibility

Setup: 100 text-dense images × 5 factual QA pairs = **500 questions**. Every answer was checked by hand on the original image. Three foundation models answer from each input type.

Table 1. ROUGE-L/BERTScore between model answers and ground-truth answers on the QA validation set.

Input	Gemma-3 27B	Qwen3-VL 32B	Llama-3.2 11B	Tokens (input)*
Image-only	0.3538/0.8834	0.4396/0.9078	0.1338/0.8575	316/1281/**59**†
OCR-only	0.5574/0.9163	0.4794/0.9158	0.4782/0.9123	284/302/307
GPT-5.3 (E2E)	0.5752/0.9219	0.4752/0.9151	0.4719/0.9091	253/271/278
Layout | GCP OCR (Ours)	**0.5818/0.9277**	**0.4830/0.9214**	**0.5046/0.9225**	247/264/272
Layout | PaddleOCR	0.5449/0.9211	0.4412/0.9160	0.4464/0.9198	**242**/**257**/270
Layout | EasyOCR	0.5261/0.9159	0.4495/0.9145	0.4338/0.9152	251/266/276

* *Token values are listed in model order (left to right).* † *Llama 3.2 are encoded into fixed-length visual tokens.*

→ **Layout (Ours) is the best on all models** — in ROUGE-L, BERTScore and BLEU, too.

Reconstruction Quality

Table 2. Image-to-layout fidelity on the ECTL validation set.

Method	LPIPS ↓	SSIM ↑	FID ↓	PSNR ↑
Layout | GCP OCR	**0.0399**	**0.9394**	**13.58**	**34.86**
Layout | PaddleOCR	0.0535	0.9298	17.88	34.17
Layout | EasyOCR	0.0708	0.9149	39.49	33.29
GPT-5.3 (E2E)	0.1803	0.8786	68.76	21.72

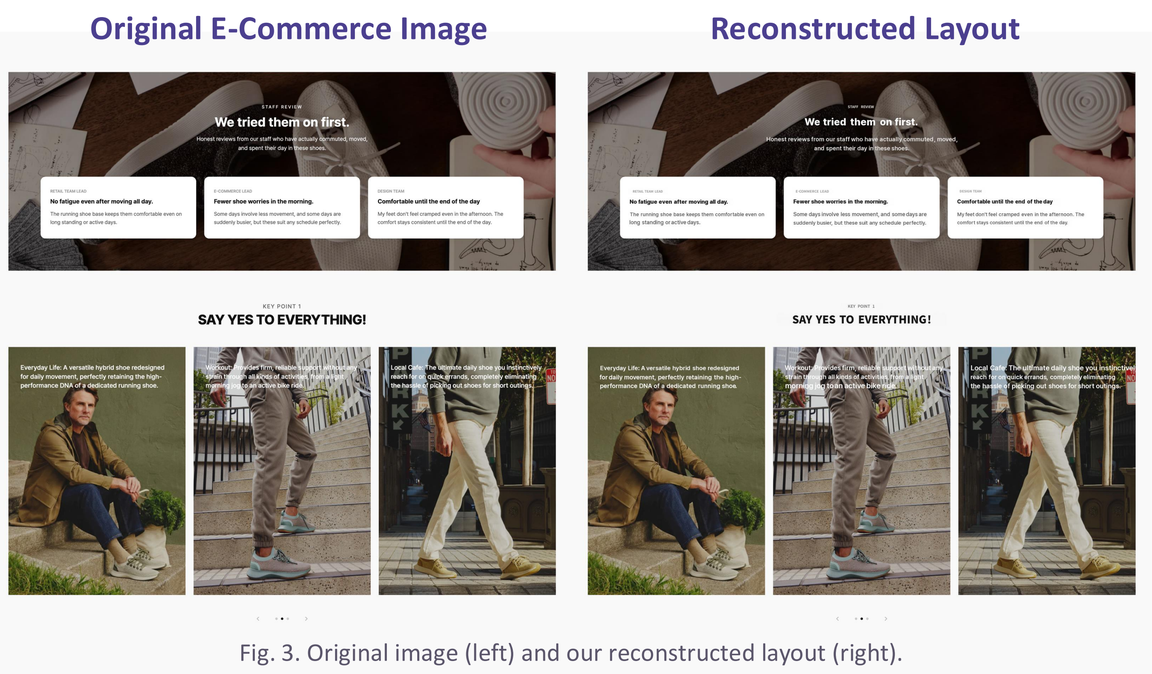

Fig. 3. Original image (left) and our reconstructed layout (right).

Table 3. QA answers from Llama 3.2 (shortened; ** = raw markdown in the model output).

Question	Image-only	Layout (Ours)	Ground truth
Q1. What do the feet not feel?	The feet do not feel *crampy* until the end of the day.	**Cramped, even in the afternoon.**	cramped
Q2. What is ‘Local Cafe’ called?	The shoe is called a Local Cafe.	**The ultimate daily shoe.**	The ultimate daily shoe
Q3. Heading under ‘STAFF REVIEW’?	**STAFF REVIEW** — … is: “**We tried them on first.**”	**We tried them on first.**	We tried them on first.
Q4. One activity in ‘Workout’?	One activity explicitly mentioned is a bike ride. …	**A morning jog.**	a light morning jog
Q5. What is ‘Everyday Life’ called?	The shoe is called “Everyday Life: A versatile hybrid shoe …”	**A versatile hybrid shoe.**	A versatile hybrid shoe

→ Image-only: long and wrong. | **Layout (Ours): short and correct.**

5 Conclusion

- Structured Layout, **searchable and editable**.
- **Better QA answers** with fewer tokens.
- **High-fidelity layout**, better than GPT-5.3 end-to-end.

Detailed Results HERE !!

References

[1] P. Aggarwal et al., “GEO: Generative Engine Optimization,” KDD 2024.
[2] S. Bai et al., “Qwen2.5-VL Technical Report,” arXiv:2502.13923, 2025.
[3] N. Ravi et al., “SAM 2: Segment Anything in Images and Videos,” arXiv:2408.00714, 2024.
[4] R. Suvorov et al., “Resolution-robust Large Mask Inpainting with Fourier Convolutions,” WACV 2022.
[5] M. Tan and Q. Le, “EfficientNetV2: Smaller Models and Faster Training,” ICML 2021.

ACM MM '26 · The 34th ACM International Conference on Multimedia · Rio de Janeiro, Brazil

* Equal contribution † Corresponding author