Learning Temporal Relations for Evaluating Instruction-Guided Image Editing

Universität Regensburg [1] Canva [2]

Women in Data Science Worldwide Regensburg

Pia Donabauer[1,2], Udo Kruschwitz[1], Alexander Tack[2]

1. Introduction

Instruction-Guided Image Editing enables automated image modification based on natural language instructions, using trained AI models.

(a) Example of an image edit
The child's right hand made a scissors gesture

Problem: Lack of an established evaluation framework tailored to text-guided image editing.[1,2,3]
Existing metrics are often inadequate, risking misrepresentation of results and slowing research progress.

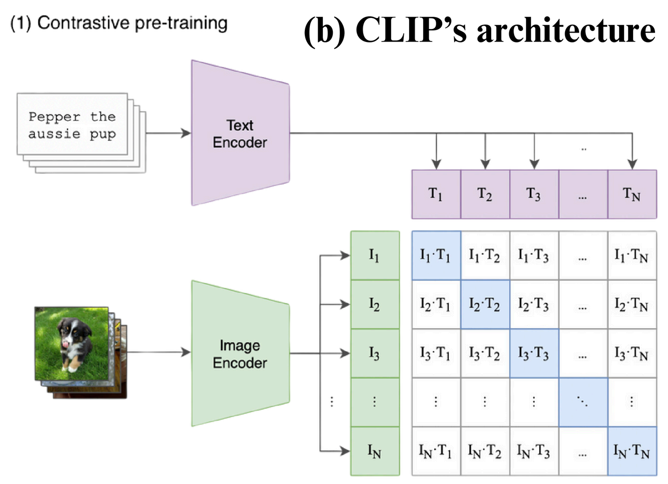
Limitations:

- **Automated metrics** are efficient,[2] but measure image quality rather than instruction fidelity.[3] Weak correlation with human judgement.[4]
- **Human evaluation** is more reliable, but costly, time-consuming and difficult to scale.[5,6]
- **Visual Language Models** (VLMs)-as-a-judge show promise,[7] but may hallucinate, be biased or misunderstand visuals.[8, 9]

2. Approach

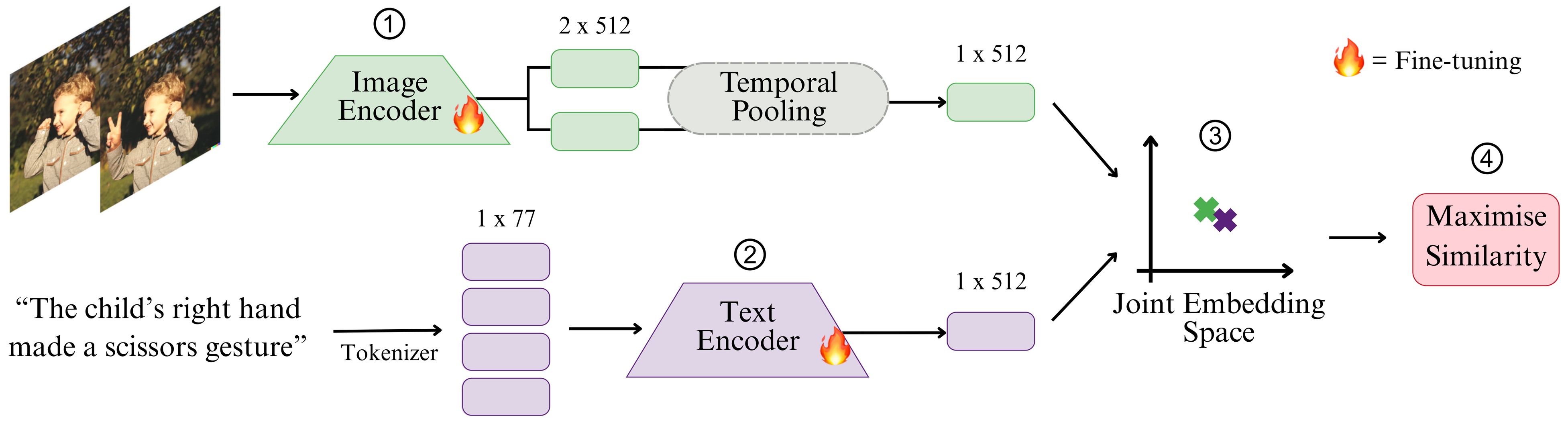
- Advances in **video understanding** using *Contrastive Language-Image Pre-Training*[10] ***(CLIP)*** suggest its potential to comprehend **temporal** dependencies.[11,12]

(b) CLIP's architecture



- *An image edit can be interpreted as a transformation from an original to an edited state.*
  → Use similar modelling techniques to evaluate image edits.
  → Learn **spatio-temporal relationships** of video sequences to better capture changes between image edit pairs and their instructions.

3. Methodology



① Image edit pairs are encoded as **2-frame videos:** [original image → edited image].
② Instruction is encoded.
③ Both the embeddings are projected into a **joint embedding space**.
④ **Training**: *Video-Finetuned CLIP*[13] (**ViFi-CLIP**) is fine-tuned on a **video-to-text retrieval task**, ranking edit-caption pairs by maximising similarity.
  - Dataset: **HumanEdit**[14] (5,751 samples, human-curated)

⑤ **Inference**: Assign each image pair and instruction a **similarity score**.

⑤

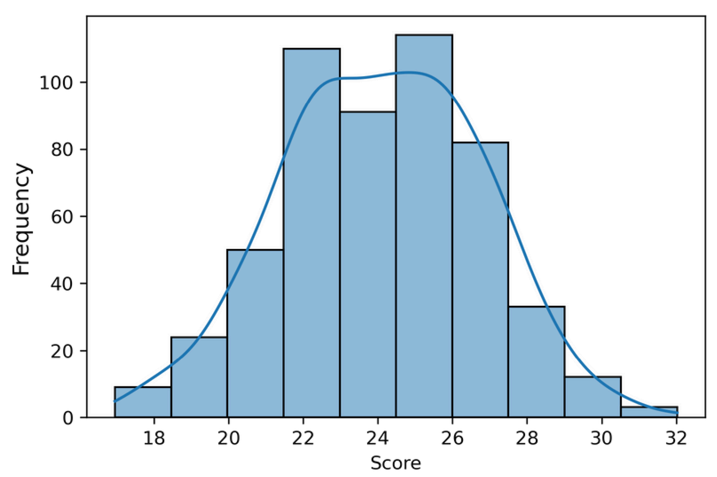
Instruction	Score
put a glass of soda on the table	**23.86**
let's add a man in the kitchen	21.52
spill milk onto the floor	21.05
let there be a cup of yogurt	19.25

⑥ **Evaluation**: Conduct **human annotation study** to assess human alignment. gradio
  - Dataset: **MagicBrush**[15] (10,388 samples, varying quality)

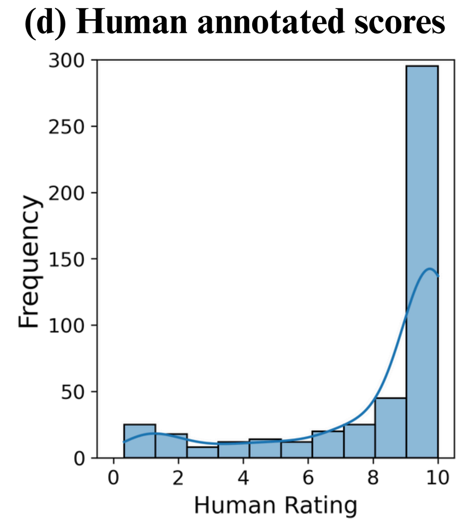
⑦ Classify edits using **GPT-4o** for a more detailed analysis.

4. Quantitative Results

(c) Scores predicted by ViFi-CLIP



(d) Human annotated scores



(e) Interrater agreement: Intraclass Correlation Coefficient (ICC)

	alignment	quality	consistency	overall
ICC (2,1)	0.73	0.51	0.38	0.66

- Mean rank of 161.22, Median rank of 29.60 (out of 1,128)
- Recall@k scores for k ∈ {1, 5, 10, 50}: 9.19, 24.13, 33.42, 59.49

(f) Spearman correlation coefficient between human annotations and ViFi-CLIP's predictions

Aspect	Spearman's ρ	*p*-value
alignment	0.12	0.01
quality	0.03	0.59
consistency	0.05	0.29
overall	0.09	0.04

(g) Spearman correlation on a subset of primarily extremely good and bad edits

Aspect	Spearman's ρ	*p*-value
alignment	0.42	< 0.001
quality	0.28	0.02
consistency	0.28	0.02
overall	0.40	< 0.001

5. Qualitative Results

(h) High human ratings (0.90* to 1.00) and low ViFi-CLIP scores (0.07 to 0.13)

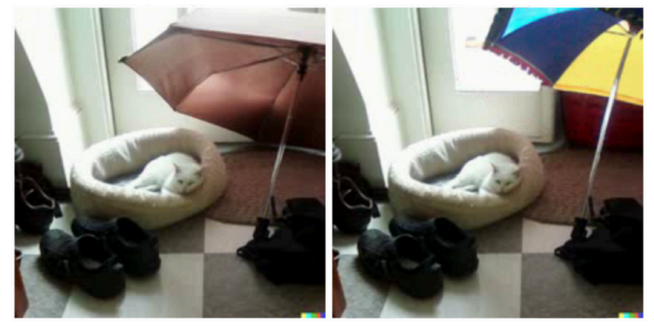
Can we have mountains on the background?

Let there be potted plant

(i) Low human ratings (0.04 and 0.14) and medium ViFi-CLIP scores (0.40 and 0.59)

Add a cotton candy machine

Add a shark next to the surfboard



(j) High disagreement among human raters (9-point difference on a 10-point Likert scale).

Let the umbrella be striped

*all values min-max normalised

6. Discussion

- **Highly subjective** nature of evaluating visual content.
- ViFi-CLIP shows **moderate retrieval performance**, struggling to capture fine-grained distinctions.
- Performance **improves** significantly for:
  - Extremely well or poorly executed image edits.
  - Specific edit types (e.g. object removals, food, drastic changes).
  - Larger high-quality training data available.
- **Lower** performance observed for attribute changes, subtle changes, edits involving people, backgrounds and objects.

Why is performance **limited**? (1) **Average pooling** may oversimplifiy complex instructions. (2) **CLIP's limitations** in fine-grained perception (pre-trained on 224x224 and prominent objects). (3) **Discrepancy** between training and eval data distribution.

[1] Xiang et al., Deep learning for image inpainting, in *Pattern Recogn.* (2023). [2] Xu et al., A Review of Image Inpainting Methods Based on Deep Learning, in *Appl. Sci.* (2023). [3] Kim et al., Augmentation-Driven Metric for Balancing Preservation and Modification in Text-Guided Image Editing (2024). [4] Li et al., Dreamedit, in *TMLR* (2023). [5] Peng et al., DreamBench++, in *ICLR* (2025). [6] Tan et al., Judgebench, in *ICLR* (2025). [7] Zheng et al., Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, in *NeurIPS* (2023). [8] Chen et al., MLLM-as-a-Judge, in *ICML* (2024). [9] Ku et al., VIEScore, in *Annual Meeting of the ACL* (2023). [10] Radford et al., Learning Transferable Visual Models From Natural Language Supervision, in *ICML* (2021). [11] Huang et al., FROSTER, in *ICLR* (2024). [12] Fang et al., CLIP2Video (2021). [13] Rasheed et al., Fine-tuned CLIP Models are Efficient Video Learners, in *CVPR* (2023). [14] Hui et al., HQ-Edit, in *ICLR* (2025). [15] Zhang et al., MagicBrush, in *NeurIPS* (2024).