Where's the Edit? Visual Models vs. Local Semantics

Audrey Zheng[1, 2], Xavier Thomas[2], Deepti Ghadiyaram[2]

Canyon Crest Academy, 5951 Village Center Loop Rd, San Diego, CA 92130[1], Boston University, 25 Bay State Road, Boston, MA 02215[2]

Introduction

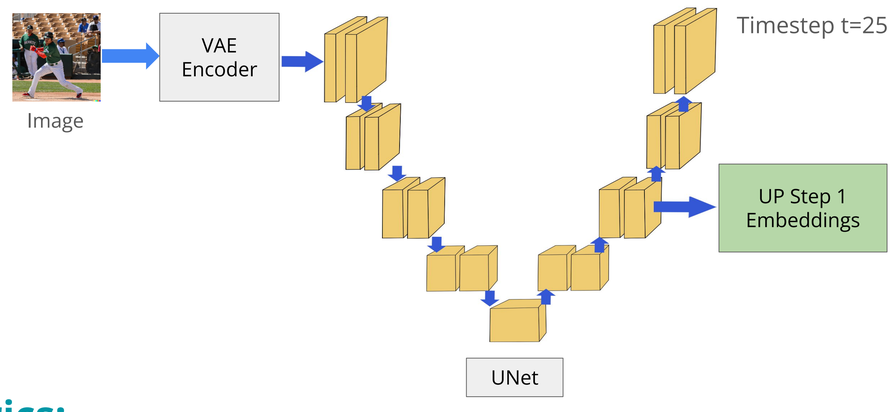
- How sensitive are feature spaces of vision models (e.g., CLIP, DINOv2) to **localized semantic edits?**
- How much does an **identical global context** influence the similarity of where an image lies in the feature space?

Methods

Dataset:

- Magicbrush [1]: 8807 image pairs with generated, object level edits

Stable Diffusion [2] Feature Extraction:



Metrics:

- **Similarity Score:** Cosine similarity between source and edited images
- **Attention-IoU (Intersection over Union):** Overlap between attention maps [3]

Key Finding

- CLIP and DINOv2 CLS features are **blind** to a significant amount of images with **identical global-contexts** but semantically meaningful, **object-level differences**

Example Failure Pair Similarities:

Observable Difference: The dog has its **head turned upwards** in the second image

CLIP CLS: 0.997 ✗	DINOv2 CLS: 0.995 ✗
Stable Diffusion: 0.899 ✓	

Observable Difference: The player is **not holding it's bat** in the second image

CLIP CLS: 0.998 ✗	DINOv2 CLS: 0.996 ✗
Stable Diffusion: 0.917 ✓	

Stable Diffusion features have a significantly lower similarity, indicating that it can **differentiate between the object-level change better than CLIP and DINOv2**

References

[1] Zhang et al. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. arXiv preprint arXiv:2306.10012 (2023)
[2] Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Model. arXiv preprint arXiv:2112.10752 (2021)
[3] Serianni et al. Attention IoU: Examining Biases in CelebA using Attention Maps. arXiv preprint arXiv:2503.19846 (2025)

Acknowledgements

I would like to thank my mentors, Xavier Thomas and Professor Deepti Ghadiyaram for their unwavering guidance and mentorship throughout this program. I would also like to thank the Boston University RISE program for this incredible opportunity.

Results

Failure on LLaVA:

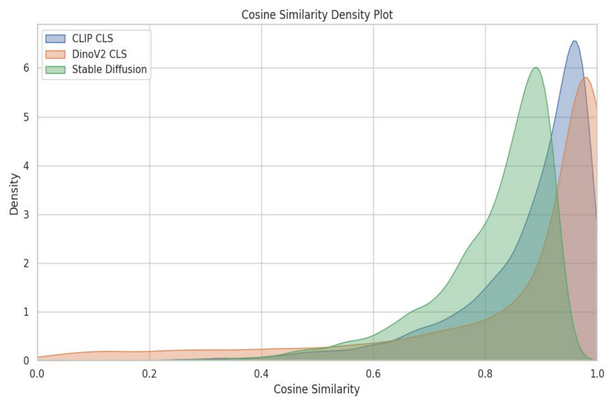
Question: What is the man holding?

A: The man is holding a baseball bat.

B: The man is holding a baseball bat.

Problem: CLIP based MLLMs **fail** to differentiate between these image pairs with high CLIP and DINOv2 similarities

Comparison with Stable Diffusion:



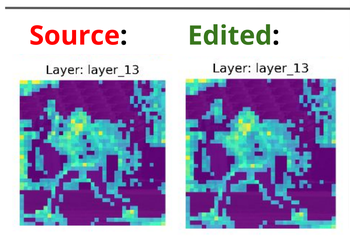
	% High CLIP & DINOv2	% High SD
Patch Tokens	**48%**	**0.3%**
CLS Tokens	**27%**	**0.3%**

- Stable Diffusion has on average **lower similarities** and significantly **less cases** with high (>0.95) similarities

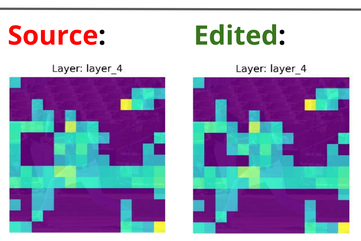
Key Finding: Stable Diffusion features are able to **better capture the differences in object-level edits**

Feature Attentions:

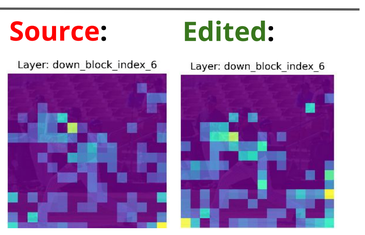
DINOv2: **Source:** **Edited:**



CLIP: **Source:** **Edited:**

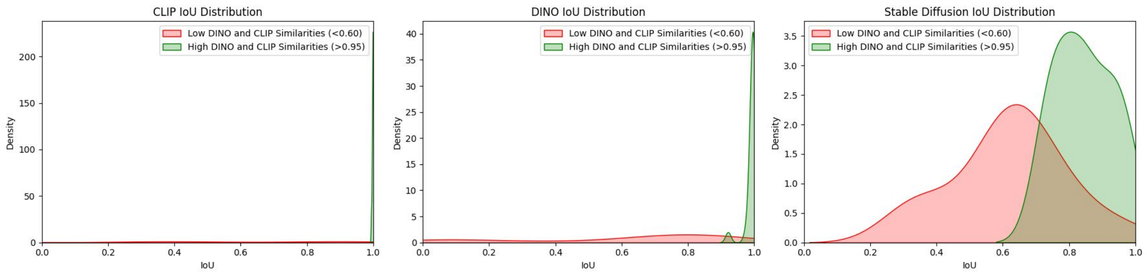


Stable Diffusion: **Source:** **Edited:**



- CLIP and DINOv2 attention maps have **little global difference** between original and edited images.
- Stable Diffusion attention maps are **significantly different**

Attention IoU Between Layers:



- Stable Diffusion attentions are **less highly concentrated** for images with **high CLIP and DINOv2 similarity**.
  - They are able to pick up on both the **object-level differences** (Low IoU) and the **identical global context** (High IoU) across multiple layers

Conclusion

Stable Diffusion features offer **a more semantically meaningful alternative** to CLIP and DINOv2 features in capturing object-level differences in identical global contexts