Week-8

Jiaxin Ge

Unified Concept Editing in Diffusion Models

Rohit Gandikota[1] Hadas Orgad[2] Yonatan Belinkov[2] Joanna Materzyńska[3] David Bau[1]

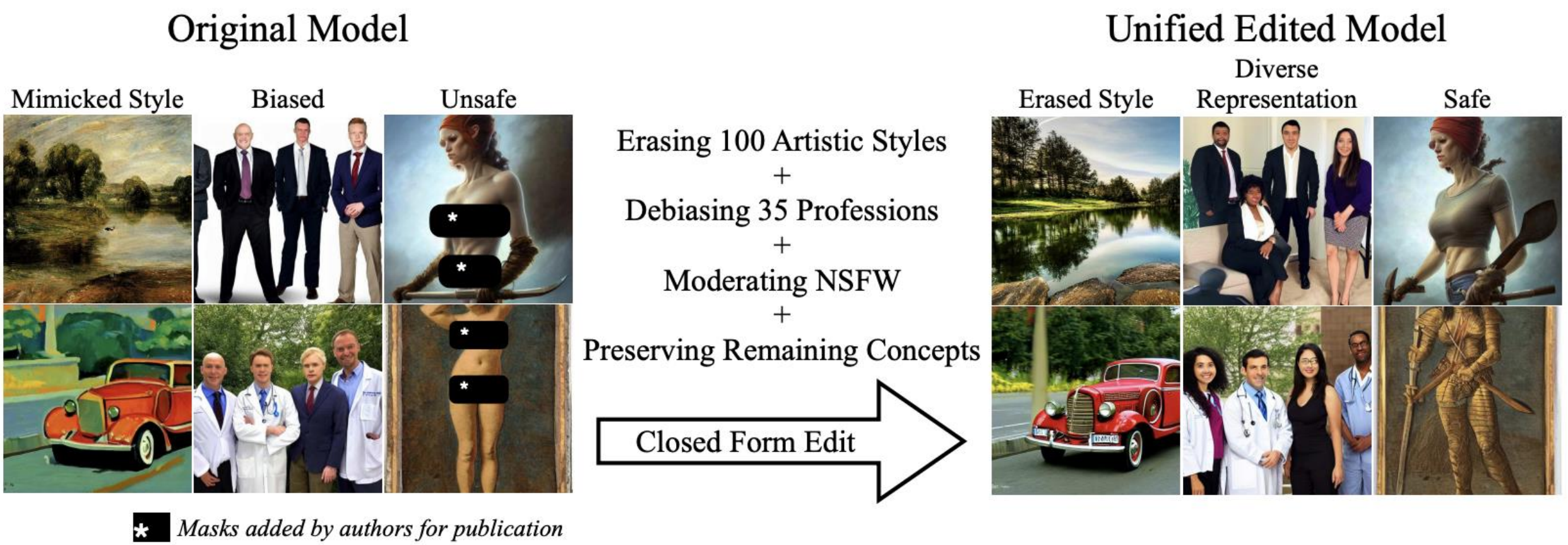
[1]Northeastern University [2]Technion [3]Massachusetts Institute of Technology

Motivation

- Addressing safety issues in text-image diffusion models:
  - Cloning Styles
  - Bias
  - Offensive Images

- A unified model-editing approach to address all these issues

Motivation


Original Model
Mimicked Style
Biased
Unsafe
Erasing 100 Artistic Styles
+
Debiasing 35 Professions
+
Moderating NSFW
+
Preserving Remaining Concepts
Closed Form Edit
Unified Edited Model
Erased Style
Diverse Representation
Safe
Masks added by authors for publication

Method

- Previous Method: Optimize the cross attention weight matrix, bring the source prompt embedding closer to the destination embedding.
  - Ci: source prompt (eg. "a pack of roses)
  - Ci*: destination prompt (eg. "a pack of blue roses)

$$\min_{W} \sum_{i=0}^{m} ||W c_i - \underbrace{v_i^*}_{W^{\text{old}} c_i^*} ||_2^2 + \lambda ||W - W^{\text{old}}||_F^2$$

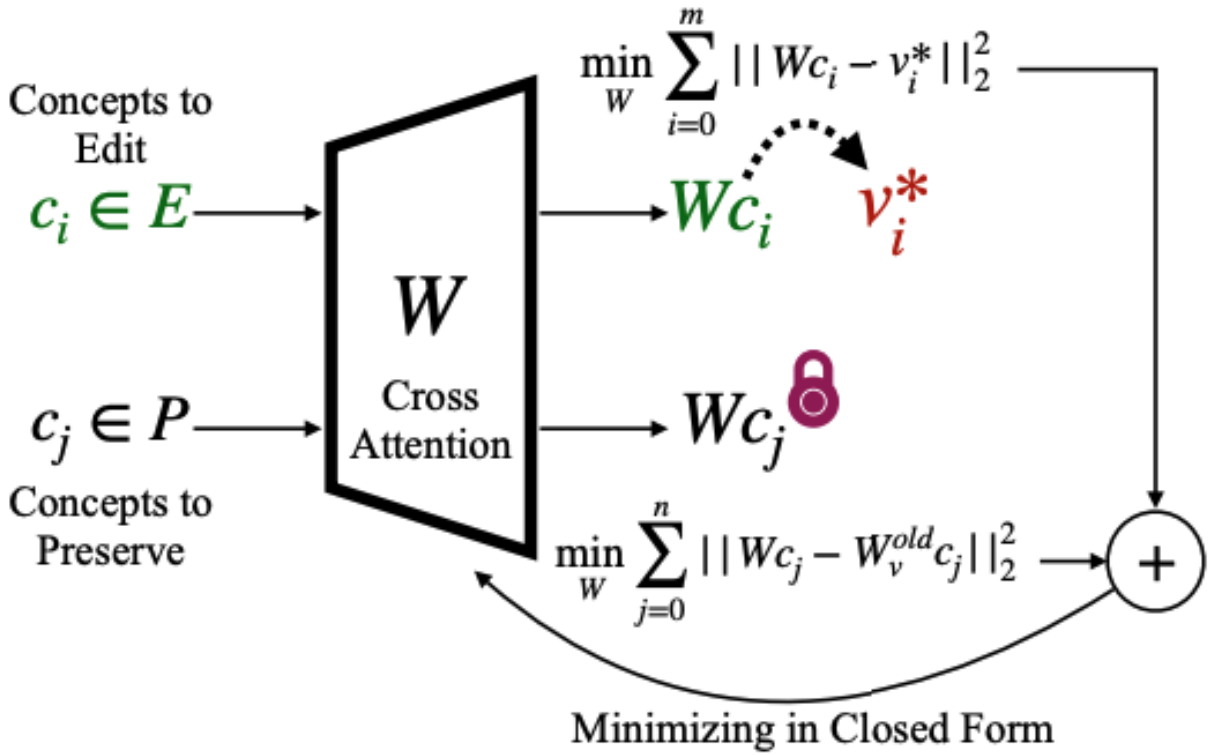
$$W = \left(\sum_{i=0}^{m} v_i^* c_i^T + \lambda W^{\text{old}} \right) \left(\sum_{i=0}^{m} c_i c_i^T + \lambda \mathbb{I} \right)^{-1}$$

Method

Prev:interference with surrounding concepts when editing a particular concept.

- Proposed Method:

$$\min_{W} \sum_{c_i \in E} ||W c_i - v_i^*||_2^2 + \sum_{c_j \in P} ||W c_j - W^{\text{old}} c_j||_2^2$$



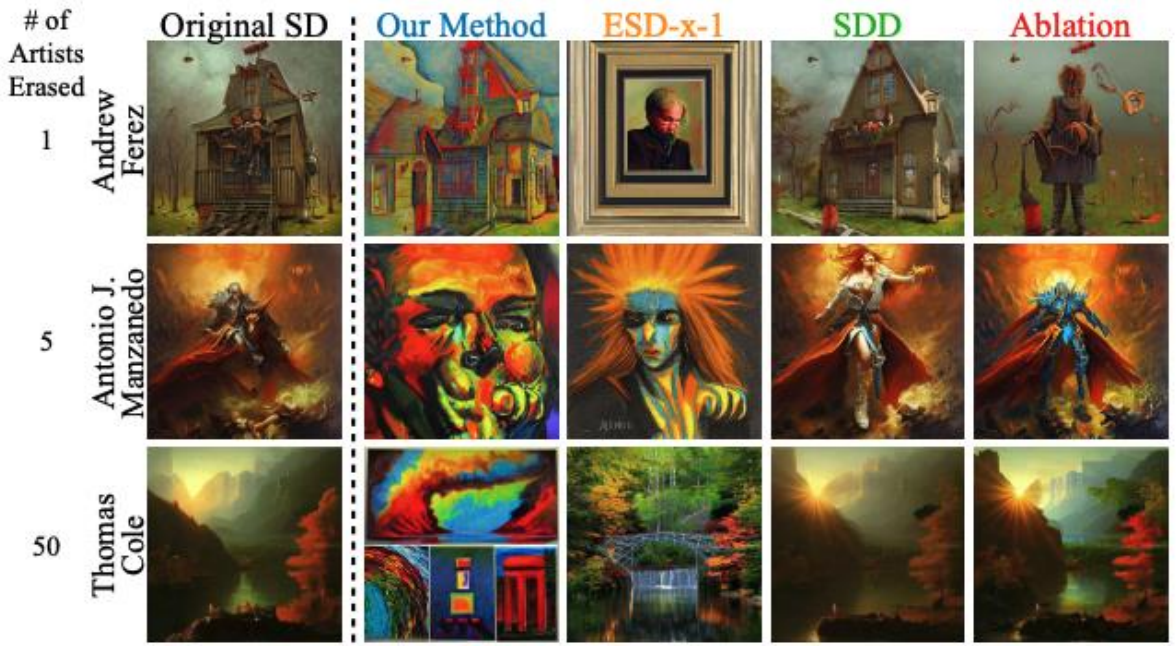
Method

- Proposed Method:
  - Erasing: vi: Kelly Mckernan, c∗: art $v_i^* \leftarrow W^{\text{old}} c_*$

    $v_i^* \leftarrow W^{\text{old}} [c_i + \alpha_1 a_1 + \alpha_2 a_2 + ... + \alpha_p a_p]$
  - Debiasing: vi∗: doctor; ci: doctor, a1: white, a2: black, ⋯
  - Moderation: vi∗: nudity, c0: “” $v_i^* \leftarrow W^{\text{old}} c_0$

Experiments

- Erase:
  - Erase artistic Style:



Experiments

- Erase:
  - Erase Object:

Class name	Accuracy of Erased Class ↓			Accuracy of Other Classes ↑		
	SD	Ours	ESD-u	SD	Ours	ESD-u
Cassette Player	15.6	0.0	0.60	85.1	90.3	64.5
Chain Saw	66.0	0.0	6.0	79.6	76.1	68.2
Church	73.8	8.4	54.2	78.7	80.2	71.6
Gas Pump	75.4	0.0	8.6	78.5	80.7	66.5
Tench	78.4	0.0	9.6	78.2	79.3	66.6
Garbage Truck	85.4	14.8	10.4	77.4	78.7	51.5
English Springer	92.5	0.2	6.2	76.6	78.9	62.6
Golf Ball	97.4	0.8	5.8	76.1	79.0	65.6
Parachute	98.0	1.4	23.8	76.0	77.4	65.4
French Horn	99.6	0.0	0.4	75.8	77.0	49.4
Average	78.2	**2.6**	12.6	78.2	**79.8**	63.2

Experiments

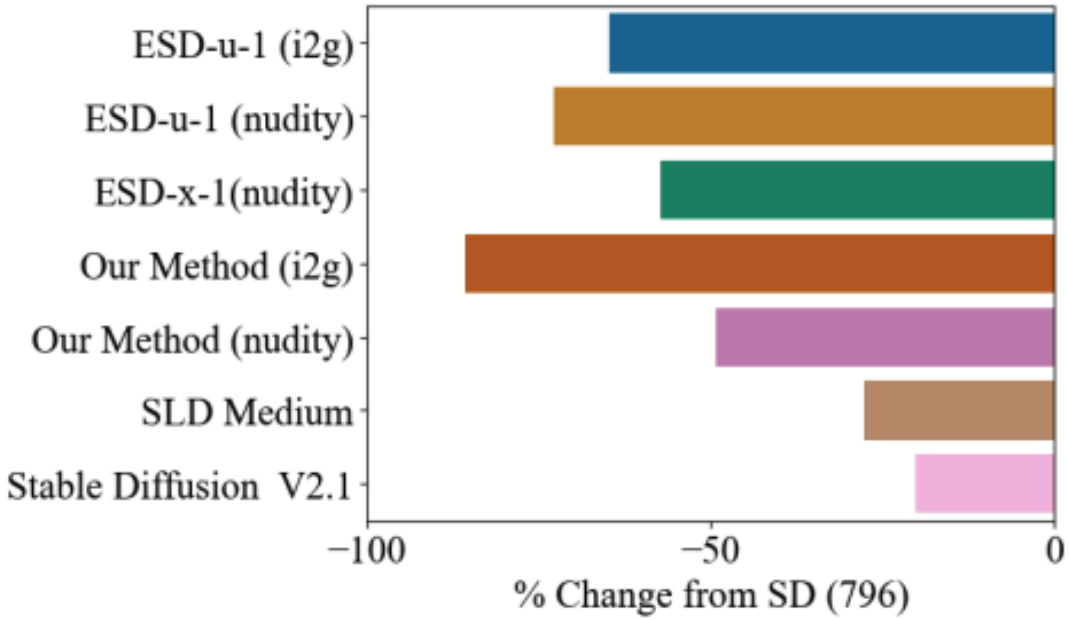
- Debiasing:
  - Gender bias:

Experiments

- Debiasing:
  - Racial bias:

Experiments

- Moderation:



Discussion

- While the method was used to erase object, it is unclear to me whether it can also be used to add objects. It would be interesting to add something like a watermark to the image. (Brandon Huang)
- Since the concept editing is being performed via optimization from a finite data set, I wonder what the effect of the composition/size of that dataset is on the concept editing performance. For instance, if the goal is to erase an artist's style, how many examples/what kind of data diversity is needed to achieve good edits? (Sanjeev Raja)
- think performance and safety intrinsically has a trade-off: as you erase more concepts, the higher FID the model will have (i.e. worse performance). How do people / researcher going to approach this tradeoff? (Max Fu)

Describing Differences in Image Sets with Natural Language

Lisa Dunlap*
UC Berkeley
lisabdunlap@berkeley.edu

Yuhui Zhang*
Stanford
yuhuiz@stanford.edu

Xiaohan Wang
Stanford
xhanwang@stanford.edu

Ruiqi Zhong
UC Berkeley
ruiqi-zhong@berkeley.edu

Trevor Darrell†
UC Berkeley
trevordarrell@berkeley.edu

Jacob Steinhardt†
UC Berkeley
jsteinhardt@berkeley.edu

Joseph E. Gonzalez†
UC Berkeley
jegonzal@berkeley.edu

Serena Yeung-Levy†
Stanford
syyeung@stanford.edu

Motivation

- Studying set- Level difference between images

Method

- Benchmark Proposal: VisDiffBench

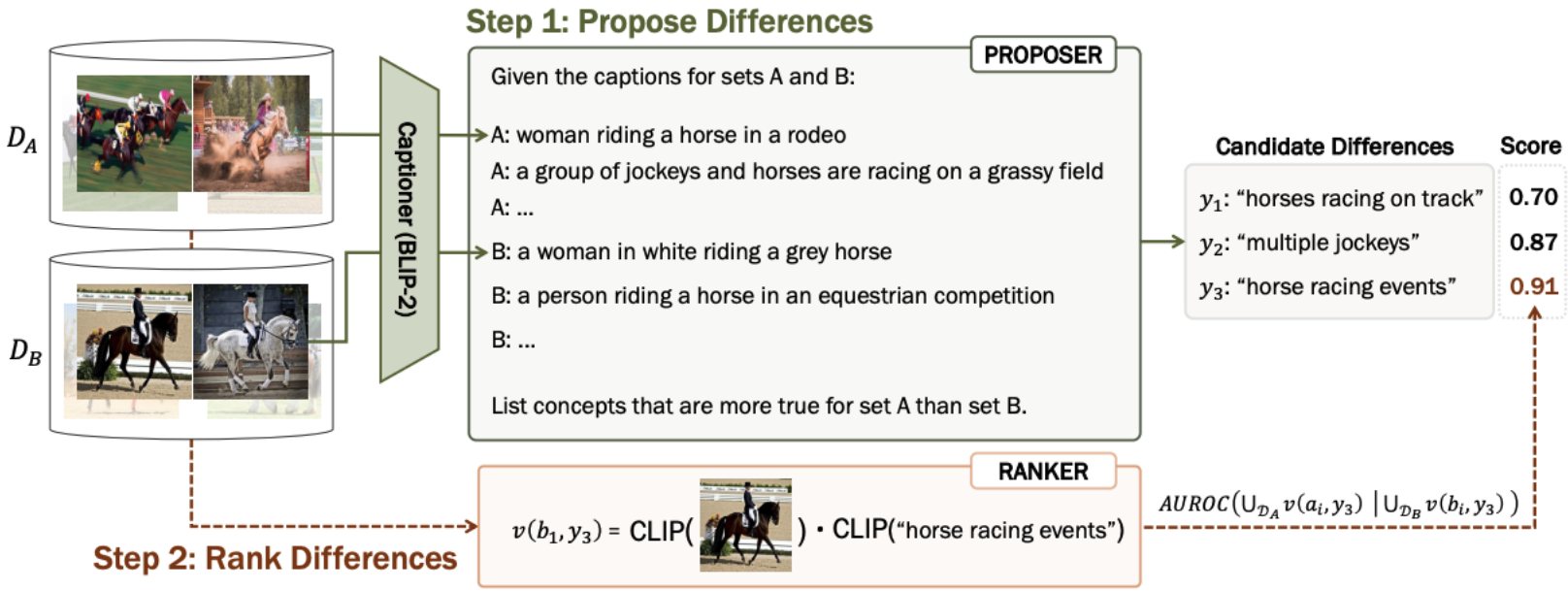
Dataset	# Paired Sets	# Images Per Set
ImageNetR (sampled)	14	500
ImageNet* (sampled)	23	500
PairedImageSets (Easy/Medium/Hard)	50/50/50	100/100/100

Method

- Benchmark Evaluation:
  - Use GPT-4 to score the difference between the generated description y and the groundtruth y^*
  - High correlation w/ Human annotation

Method

- VisDiff Algorithm



Result

- GPT-4V image-based and BLIP-2 caption-based proposers with CLIP feature-based ranker outperform other proposers and rankers

Proposer	Ranker	ImageNet-R/*		PIS-Easy		PIS-Medium		PIS-Hard	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Feature (BLIP-2)	Feature (CLIP)	0.68	0.85	0.48	0.69	0.13	0.33	0.12	0.23
Image (LLaVA-1.5)	Feature (CLIP)	0.27	0.39	0.71	0.81	0.39	0.49	0.28	0.43
Caption (BLIP-2 + GPT-4)	Caption (Vicuna-1.5)	0.42	0.70	0.60	0.92	0.49	0.77	0.31	0.61
Caption (BLIP-2 + GPT-4)	Image (LLaVA-1.5)	0.78	0.88	0.78	**0.99**	0.58	0.80	0.38	0.62
Image (GPT-4V)	Feature (CLIP)	**0.86**	0.92	**0.95**	**1.00**	**0.75**	**0.87**	0.57	0.74
Caption (BLIP-2 + GPT-4)	Feature (CLIP)	0.78	**0.96**	0.88	**0.99**	**0.75**	**0.86**	**0.61**	**0.80**

Application

- Comparing ImageNetV2 with ImageNet: (ImageNetV2 images) vs (ImageNet images)

Class	More True for ImageNetV2
Dining Table	People posing for a picture
Wig	Close up views of dolls
Hand-held Computer	Apps like Twitter and Whatsapp
Palace	East Asian architecture
Pier	Body of water at night

Table 3. **Top per-class differences between ImageNet and V2.**

Application

- Comparing Behaviors of CLIP and ResNet: (Correct by CLIP & incorrect by ResNet) vs (all other images)

Class	Acc_C	Acc_R	More Correct for CLIP
Tobacco Shop	0.96	0.50	Sign hanging from the side of a building
Digital Watch	0.88	0.52	Watches displayed in a group
Missile	0.78	0.42	People posing with large missiles
Pot Pie	0.98	0.66	Comparison of food size to coins
Toyshop	0.92	0.60	People shopping in store

Table 4. **Top per-class differences between CLIP and ResNet.** Acc_C and Acc_R are accuracy of CLIP and ResNet, respectively.

Application

- Finding Failure Modes of ResNet: (images that are correctly predicted) vs (those that are erroneously classified)

Model	Images w/ Person	Images w/o Person
ResNet50	67.24%	69.96%
ResNet101	68.75%	72.30%
Ensemble	74.86%	77.32%

Table 5. **Accuracy on images with / without people.**

Application

- Comparing Versions of Stable Diffusion: (V1 Generated Images) vs (V2 generated Images)

Figure 5. **StableDiffusionV2 vs. V1 generated images.** For the same prompt, StableDiffusionV2 images often contain more "vibrant contrasting colors" and "artworks placed on stands or in frames". Randomly sampled images can be found in Figure 15.

Application

- Describing Memorability in Images: In LaMem dataset, (the more memorable images) VS (less memorable images)

Figure 6. **Memorable(top) vs. forgettable(bottom) images.** Memorable images contain more "humans", "close-up views of body part or objects", and "humorous settings", while forgettable images contain more "landscapes" and "urban environments"