Model Diagnosis and Correction via Linguistic and Implicit Attribute Editing

Supplementary Material

Contents

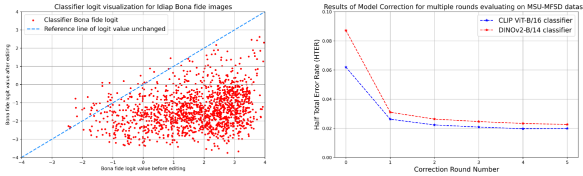


Figure 1. Left part. **Causal Relationship Verification** through classifier logits changes after editing with "overall orange color" attribute. From the figure, we can observe that after only editing the "overall orange color" attribute for the input images without changing others, the majority of output, *i.e.* the classifier logits, changes. This verifies there exists a causal relationship between this attribute and the performance of classifier. Right part. **Correction rounds vs Performance** relationship through analyzing the model from last round and select Top-1 attribute for correction for each round. As round number increases, performance firstly improves and then becomes saturated.

Dataset	Method	HTER(↓) / AUC (↑) / TPR@FPR=1% (↑)	
		CLIP ViT-B/16 [18]	DINOv2-B/14 [13]
MSU-MFSD	Train (O and C)	6.19% / 97.71% / 50.48%	8.71% / 96.92% / 44.54%
	Train + Validation (I)	3.96% / 98.98% / 80.95%	3.73% / 99.15% / 78.09%
	Train + Top 1 Edited Copy	2.41% / 99.33% / 78.56%	3.02% / 98.89% / 75.62%
	Train + Top 2 Edited Copies	2.21% / 99.48% / 94.01%	2.72% / 99.18% / 89.59%
	Train + Top 3 Edited Copies	2.17% / 99.46% / 94.78%	2.44% / 99.30% / 91.11%
	Train + Top 4 Edited Copies	**2.11% / 99.52% / 95.32%**	**2.40% / 99.44% / 92.02%**

Table 1. **Model Correction Performance Comparisons** between adding MDC edited counterfactual training sets *vs.* adding real validation set when testing on MSU dataset in face anti-spoofing domain. We select attributes from Figure 4 of the Main Text and then generate counterfactual training copies of OULU and CASIA datasets with the pattern of selected attributes. Finally, we add them to OULU and CASIA. "Top K Edited Copies" means we select K attributes and generate K edited training set copies. In terms of adding real validation set, we directly add Idiap to OULU and CASIA. Here, it shows leveraging the distilled causal information from validation set and further generating more training images is superior than using the data solely from the validation set.

1. Causal Relationship

In the main text, we leverage experimental analysis to disclose the "causal" relationship between the editing visual attributes and model performance, measuring solely on a decline in the visual model's recognition accuracy, which is accurate but not adequate. Actually, to reveal such relationship, we strictly follow the definition in [14], *i.e.* an output event E happens after many input events occur. Formally, the causal relationship is defined as "to determine whether one input event e_i is the cause for E, we generate counterfactual, which *only* change one event e_i from occurring to non-occurring, and observe whether the probability of E happens changes" [9]. Strictly applying this to our task, the attributes of images in validation set are considered input events, while the distribution of the classifier predicted logits is considered the output event (in Section 1 of main text). In left part of Figure 1, the values for majority Bona fide logits decrease after editing, verifying the causal relationship between added attribute and classifier output. This aligns with results in main paper (relying on accuracy decline, *i.e.* a consistent reflection for distribution changes of logits value). Such cross-validaitons by strictly following the definition of "causality" verify the attributes MDC found indeed own the causal effects to the model performance.

2. Correction rounds vs Performance

In the main text, we only correct model *once* with multiple attributes at the same time. Here, we run experiments over multiple runs to bring additional insights and show the visualization in right part of Figure 1. For each round, we analyze the model from last round and select Top-1 attribute for correction. As round number increases, performance firstly improves and then becomes saturated. This verifies that conducting one round correction is the best choice in

Dataset	**Method**	**Acc@1 / Acc@5** (↑)			**Dataset**	**Method**	**Acc@1 / Acc@5** (↑)		
		ViT-L/14	CLIP ViT-B/16 [18]	DINOv2-B/14 [13]			ViT-L/14	CLIP ViT-B/16 [18]	DINOv2-B/14 [13]
Stanford Dogs	Original Training Set	90.54% / 97.40%	82.90% / 90.52%	83.81% / 91.64%	Tsinghua Dogs	Original Training Set	88.20% / 94.98%	82.17% / 90.95%	84.41% / 91.76%
	LANCE [16] - Top 1	90.87% / 97.62%	83.38% / 90.87%	84.77% / 92.03%		LANCE [16] - Top 1	89.53% / 95.77%	82.64% / 90.95%	84.42% / 92.01%
	LANCE [16] - Top 2	90.98% / 97.70%	83.69% / 91.27%	84.69% / 92.16%		LANCE [16] - Top 2	89.60% / 96.01%	83.43% / 91.17%	84.10% / 91.87%
	LANCE [16] - Top 3	91.03% / 97.78%	84.11% / 91.57%	84.60% / 92.36%		LANCE [16] - Top 3	89.72% / 96.14%	83.38% / 91.24%	84.33% / 92.00%
	LANCE [16] - Top 4	91.00% / 98.05%	84.02% / 91.88%	84.77% / 92.11%		LANCE [16] - Top 4	89.99% / 96.32%	83.56% / 91.28%	84.31% / 92.05%
	MDC- Top 1	91.13% / 98.10%	84.58% / 91.75%	85.07% / 92.79%		MDC- Top 1	90.03% / 96.35%	83.77% / 92.04%	85.98% / 92.65%
	MDC- Top 2	91.56% / 98.37%	84.56% / 92.33%	86.02% / 93.26%		MDC- Top 2	90.78% / 96.99%	84.52% / 92.70%	86.50% / 93.21%
	MDC- Top 3	91.66% / **98.41%**	84.66% / **92.43%**	**86.72%** / 93.28%		MDC- Top 3	90.90% / 97.04%	84.60% / 92.75%	**86.61%** / 93.40%
	MDC- Top 4	**91.68%** / 98.40%	**84.69%** / 92.40%	86.66% / **93.31%**		MDC- Top 4	**90.91% / 97.09%**	**84.72% / 92.79%**	86.58% / **93.49%**
CUB	Original Training Set	90.80% / 96.67%	76.79% / 88.07%	82.19% / 90.25%	NABirds	Original Training Set	91.99% / 97.03%	85.77% / 92.18%	84.38% / 91.86%
	LANCE [16] - Top 1	91.43% / 97.00%	78.81% / 90.01%	83.20% / 90.88%		LANCE [16] - Top 1	91.77% / 96.67%	86.07% / 92.25%	84.22% / 91.76%
	LANCE [16] - Top 2	91.67% / 97.05%	78.51% / 90.10%	83.98% / 91.07%		LANCE [16] - Top 2	91.89% / 97.01%	86.48% / 92.73%	84.67% / 92.12%
	LANCE [16] - Top 3	91.63% / 97.08%	78.53% / 90.06%	84.00% / 91.43%		LANCE [16] - Top 3	91.78% / 97.05%	86.50% / 92.41%	84.70% / 92.33%
	LANCE [16] - Top 4	91.60% / 97.14%	78.56% / 90.15%	84.21% / 91.27%		LANCE [16] - Top 4	91.68% / 97.08%	86.55% / 92.60%	84.74% / 92.36%
	MDC- Top 1	91.58% / 97.27%	79.63% / 90.25%	84.44% / 91.56%		MDC- Top 1	92.44% / 97.51%	**86.63%** / 92.89%	85.05% / 92.07%
	MDC- Top 2	**91.88%** / 97.45%	81.62% / 91.22%	85.98% / 92.42%		MDC- Top 2	92.98% / **97.87%**	86.31% / 93.11%	**85.64%** / 92.45%
	MDC- Top 3	91.83% / **97.48%**	**81.66% / 91.25%**	86.10% / 92.56%		MDC- Top 3	**93.04%** / 97.61%	86.35% / 93.18%	85.56% / **92.67%**
	MDC- Top 4	91.80% / 97.43%	81.57% / 91.11%	**86.30% / 92.77%**		MDC- Top 4	93.00% / 97.65%	86.41% / **93.21%**	85.49% / 92.61%

Table 2. **Model Correction Complete Results** in dog and bird species domains. We select Top K attributes from Figure 3 of the Main Text based on the results of diagnosis only using ViT-L/14 classifiers on *validation sets*, *i.e*. test parts of Stanford Dogs and CUB. For classifier retraining and correction, we augment the training part of each dataset and test on corresponding test part. Here, it shows that MDC is superior than LANCE in leveraging causal attributes to boost the performance of the model. This Table is a complete version for Table 1 of the Main Text.

Dataset	**Method**	**HTER**(↓) / **AUC**(↑) / **TPR@FPR=1%**(↑)		**Dataset**	**Method**	**HTER**(↓) / **AUC** (↑) / **TPR@FPR=1%** (↑)	
		CLIP ViT-B/16 [18]	DINOv2-B/14 [13]			CLIP ViT-B/16 [18]	DINOv2-B/14 [13]
Idiap Replay Attack	Original Training Set	4.14% / 98.63% / 94.10%	6.06% / 98.18% / 80.13%	MSU-MFSD	Original Training Set	6.19% / 97.71% / 50.48%	8.71% / 96.92% / 44.54%
	LANCE [16] - Top 1	3.89% / 98.88% / 95.30%	5.87% / 98.44% / 82.05%		LANCE [16] - Top 1	5.87% / 98.02% / 52.77%	8.77% / 96.79% / 48.55%
	LANCE [16] - Top 2	3.43% / 99.03% / 96.02%	5.37% / 98.73% / 85.59%		LANCE [16] - Top 2	5.74% / 98.31% / 55.49%	7.99% / 97.02% / 53.14%
	LANCE [16] - Top 3	3.38% / 99.01% / 96.10%	5.38% / 98.77% / 85.87%		LANCE [16] - Top 3	5.88% / 98.10% / 57.14%	8.05% / 97.30% / 56.41%
	LANCE [16] - Top 4	3.44% / 99.04% / 96.14%	5.29% / 98.67% / 85.74%		LANCE [16] - Top 4	5.67% / 98.66% / 59.04%	8.00% / 97.11% / 58.50%
	MDC- Top 1	1.66% / 99.60% / 98.64%	2.78% / 99.03% / 90.77%		MDC- Top 1	2.62% / 99.24% / 75.24%	3.10% / 98.59% / 70.11%
	MDC- Top 2	**1.23% / 99.66%** / 99.09%	1.99% / **99.12%** / 96.34%		MDC- Top 2	2.46% / **99.44%** / 93.33%	2.88% / **99.03%** / 88.20%
	MDC- Top 3	1.37% / 99.63% / **99.15%**	1.78% / 99.10% / 96.58%		MDC- Top 3	**2.43%** / 99.40% / 94.01%	2.66% / 99.00% / 89.03%
	MDC- Top 4	1.40% / 99.60% / 99.13%	**1.71%** / 99.01% / **96.62%**		MDC- Top 4	2.49% / 99.36% / **94.71%**	**2.63%** / 98.99% / **89.28%**

Table 3. **Model Correction Complete Results** in face anti-spoofing domain. We select attributes from Figure 4 of the Main Text based on the results of diagnosis on Idiap and with CLIP ViT-B/16 classifier. For classifier retraining and correction, we augment the generated data to *training sets*, *i.e*. OULU and CASIA. This Table is a complete version for Table 3 of the Main Text.

balancing the resource costs and performance gains.

3. Additional Model Correction Results

We provide more results for model correction. In Section 3.1, we aim to verify that that adding edited counterfactual training data with the pattern of causal attributes is superior than simply adding the real validation set, which is leveraged for discovering causal attributes, to the training set. In Section 3.2, we present the results via selecting more attributes (Top 3 and 4) under the same model correction strategy with main text.

3.1. Add Edited Training Set *vs.* Real Validation Set

In the main text, we claim one benefit of MDC is that “once causality is confirmed, we can generate unlimited training samples depicting the identified error pattern, which can be further used for model correction without requiring additional data sourcing, saving time and cost while improving accuracy, robustness.” This raises a straight forward question: *Now that we can generate more or even unlimited counterfactual training samples with causal patterns to augment the training set and boost the model, how is it compared with simply adding the real validation set, which is leveraged for discovering causal attributes, to the original training set?* To answer this question, we keep the strategy of selecting causal attributes unchanged but modify our model correction strategy. Instead of replacing part of the original training set with counterfactual samples, we directly edit the whole training set to contain the causal pattern of selected attributes, and then directly add this counterfactual training copy to original training data. Specifically, if we select Top 2 attributes, the final training data volume is 3 times larger compared with the original training data volume. For comparison, we directly combine the *training set* and *validation set* for training and leverage *test set* for evaluation as another baseline.

We conduct this experiment on face anti-spoofing domain since the training (OULU and CASIA), validation (Idiap), and test (MSU) sets have the same classes. We select attributes from Figure 4 of the Main Text. From Table 1, we can observe that leveraging the distilled causal information from validation set and further generating more training im-

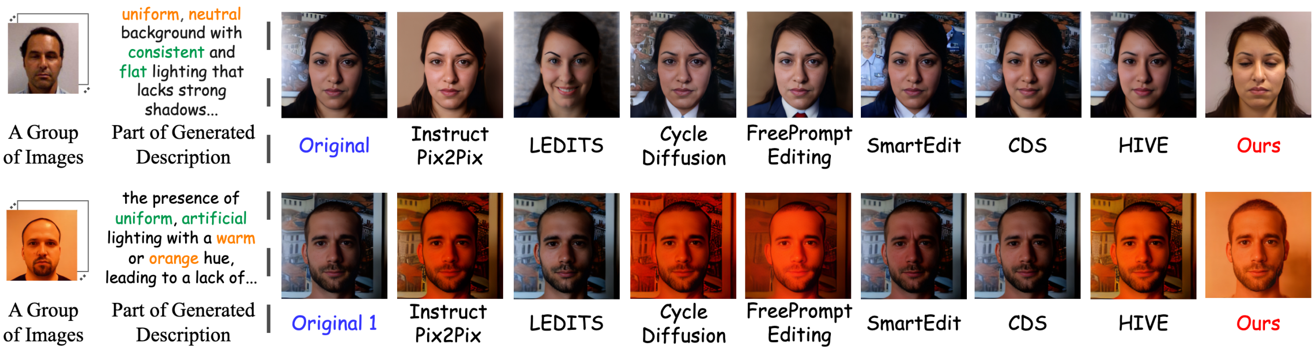

uniform, neutral background with consistent and flat lighting that lacks strong shadows...
A Group of Images
Part of Generated Description
Original
Instruct Pix2Pix
LEDITS
Cycle Diffusion
FreePrompt Editing
SmartEdit
CDS
HIVE
Ours
the presence of uniform, artificial lighting with a warm or orange hue, leading to a lack of...
A Group of Images
Part of Generated Description
Original 1
Instruct Pix2Pix
LEDITS
Cycle Diffusion
FreePrompt Editing
SmartEdit
CDS
HIVE
Ours

Figure 2. **Comparison of Implicit Attribute Editing** of ours against more baselines using language-guided editing. Like what we did in Section 4.3 of the Main Text, we firstly generate a long description of the image group using a MLLM and then use the generated description to serve as the guidance for the language-guided editing baselines. Here, it shows that ours is better at capturing the patterns resembled in the group of images representing implicit attributes. This Figure is a complete version for Figure 6 of the Main Text.

Method	Stanford Dogs			CUB200			Idiap Replay Attack		
	CLIP Score↑	FID↓	Unchanged Ratio↑	CLIP Score↑	FID↓	Unchanged Ratio↑	CLIP Score↑	FID↓	Unchanged Ratio↑
Original Validation Set	16.66	97.23	100%	21.83	68.07	100%	18.41	77.05	100%
Edited Validation Set - Pix2Pix [3]	**20.03**	92.57	85.74%	22.87	64.04	82.31%	20.11	81.34	84.66%
Edited Validation Set - LEDITS [20]	18.62	100.56	79.45%	21.99	77.21	73.09%	18.87	87.04	73.40%
Edited Validation Set - Cycle [24]	18.77	95.66	87.46%	22.05	66.04	**85.98%**	19.55	78.15	87.36%
Edited Validation Set - FreePrompt [10]	19.74	97.90	83.66%	22.51	67.90	83.88%	20.08	76.00	81.39%
Edited Validation Set - SmartEdit [6]	18.23	93.99	85.81%	20.03	71.48	84.17%	18.99	72.72	84.27%
Edited Validation Set - CDS [12]	19.02	87.01	85.04%	21.06	64.82	82.05%	19.64	71.46	86.72%
Edited Validation Set - Hive [27]	19.92	95.03	84.44%	20.39	67.00	80.74%	20.11	69.71	82.12%
Edited Validation Set - MDC	19.94	**78.24**	**88.16%**	**23.58**	**55.13**	85.75%	**20.37**	**40.81**	**90.12%**

Table 4. **Comparisons of Editing Capability with Quantitative Evaluation** between MDC and language-guided editing baselines. Like what we did in Section 4.3 of the Main Text, we compute the CLIP Score between each linguistic attribute's description and every image of corresponding edited validation set (or original validation set). We compute the FID Score between the edited sets (or original set) and corresponding classified group of validation error images containing the implicit patterns to measure the distribution distance, which represents the editing capability for implicit attributes. For unchanged ratio, we compute the ratio of CLIP prediction, between every image in original and edited validation sets, for appearance-related attributes in Table 7 to 9, keeping unchanged during editing. Here, it shows MDC is superior in implicit attribute editing, competitive in linguistic attribute editing and keeping unrelated attributes unchanged. This Table is a complete version for Table 3 of the Main Text.

ages show superior performance than using the data solely from the validation set. We can even find that the HTER performance of MDC surpasses the "Train + Validation" baseline by absolute 1.55% and 0.71% on CLIP ViT-B/16 and DINOv2-B/14 architectures respectively even with only one attribute is selected.

3.2. Augment with More (Top 3, 4) Attributes

To further evaluate the effectiveness of our discovered causal attributes, we select more attributes (Top 3 and Top 4) from Figure 3 and 4 of the Main Text to conduct targeted editing for model correction. The model correction strategies are kept the same as main text (replacing part of the training data). From Table 2 and 3, we can observe MDC still achieves consistent better results than both the "Original Training Set" and LANCE baselines. However, the rate of increase by adding third and fourth attributes drops significantly compared with adding the first two attributes, and sometimes the performance even regresses, *e.g*. adding Top 4 attributes (91.80%) getting worse performance than only adding top 2 attributes (91.88%) for ViT-L/14 classifier on CUB200 dataset for MDC. This is possibly because as the number of selected attributes increases, the intensity of their causality decreases and making the augmented training set lack of adequate causal information of error patterns.

4. Additional Editing Capability Evaluation

4.1. Additional Comparisons with More Baselines

Like what we did in Section 4.3 of the Main Text, we compare with more language-guided editing baselines from both qualitative perspective in Figure 2 and quantitative perspective in Table 4. Our method is more effective in simulating implicit attributes, while achieving competitive performance with the baselines in linguistic attribute editing and keeping unrelated attributes unchanged.

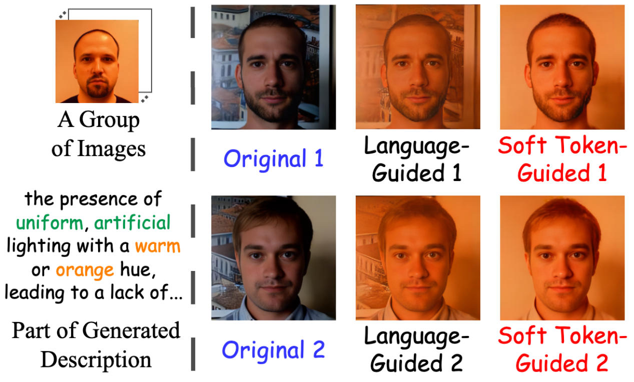

A Group
of Images
Original 1
Language-Guided 1
Soft Token-Guided 1
the presence of uniform, artificial lighting with a warm or orange hue, leading to a lack of...
Part of Generated Description
Original 2
Language-Guided 2
Soft Token-Guided 2

Figure 3. **Ablation Study for Language *vs.* Soft Token Guided Editing** for implicit attribute simulation, *i.e*. "orange hue style". For language-guided editing, we only change the guidance from soft tokens to the generated description. Here, it shows that the soft token guided editing method is superior in simulating implicit attribute represented by a group of images. The background of soft token guided images is closer to images in the classified group (top left corner).

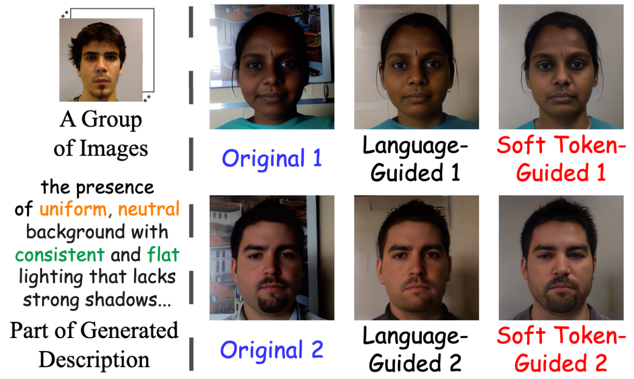

A Group
of Images
Original 1
Language-Guided 1
Soft Token-Guided 1
the presence of uniform, neutral background with consistent and flat lighting that lacks strong shadows...
Part of Generated Description
Original 2
Language-Guided 2
Soft Token-Guided 2

Figure 4. **Ablation Study for Language *vs.* Soft Token Guided Editing** for implicit attribute simulation, *i.e*. "neutral background".

Dataset	"orange hue style" in Figure 3		"neutral background" in Figure 4	
	FID↓	Unchanged Ratio ↑	FID↓	Unchanged Ratio ↑
Original Validation Set (Idiap)	94.61	100%	71.10	100%
Edited Idiap - MDC, Language	68.28	91.15%	50.19	87.41%
Edited Idiap - MDC, Soft Token	**37.99**	**91.23%**	**29.83**	**88.19%**

Table 5. **Ablation Study for Language *vs.* Soft Token Guided Editing with Quantitative Evaluation** for implicit attribute simulation, *i.e*. "orange hue style" in Figure 3 and "neutral background" in Figure 4. We compute the FID Score between the Idiap edited validation sets and corresponding classified group of Idiap error images containing the implicit "orange hue style" and "neutral background" patterns to measure the distribution distance between them. For unchanged ratio, we compute the ratio of CLIP prediction, between every image in original and edited validation sets, for appearance-related attributes in Table 9, keeping unchanged. Here, it shows that the soft token guided editing method is superior in simulating implicit attributes.

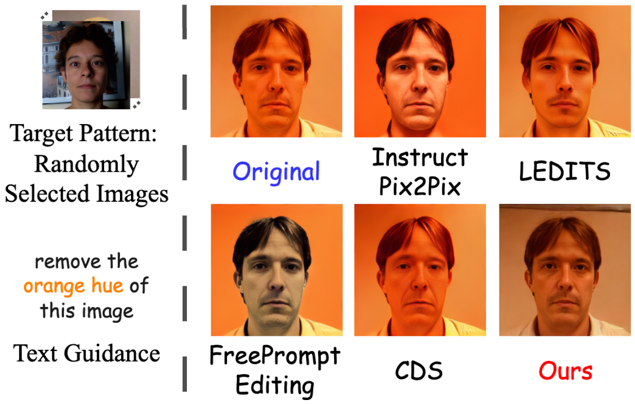

Target Pattern: Randomly Selected Images
Original
Instruct Pix2Pix
LEDITS
remove the orange hue of this image
Text Guidance
FreePrompt Editing
CDS
Ours

Figure 5. **Failure Cases**, *i.e*. removing the "orange hue" pattern from the original image. The randomly selected group of images are leveraged for the guidance of our method, while the text guidance is used for guiding baselines. Neither the baselines nor our method remove the pattern successfully. However, our method indeed reduces the "orange hue" pattern to some extent.

4.2. Ablation Study for Soft Token Guided Editing

In previous sections, we have verified that MDC is superior than language-guided editing baselines in simulating implicit attributes. However, due to the differences in architecture, training strategy, and other components, comparing with other language-guided editing methods can not fully validate soft token guided editing is better than the language guided one. To make an accurate comparison, we conduct ablation study for soft token guided editing *vs.* language guided editing. Specifically, in order to simulate implicit patterns, we only change the guidance from the optimized soft tokens to the generated text descriptions (presented in column 4 of Figure 8 to 10).

From several examples in face anti-spoofing domain in Figure 3 and 4, we can observe the soft token guided edited images are visually closer to the patterns resembled in the group of images than the edited ones via language. Specifically, in Figure 3, the language guided images failed to remove the background while the soft token guided images successfully did. To quantitatively evaluate whether the pattern "orange hue style" and "neutral background" are simulated correctly, we compute the FID Score and "Unchanged Ratio" like what we did in previous sections. We leverage the classified groups of images obtained in causal attribute discovery stage and compute their distribution distance using FID with original or counterfactual edited datasets. From the results in Table 5, the edited sets guided by soft tokens are distribution-wised closer to group of images with desired implicit patterns. For the unchanged ratio, we compute the ratio of CLIP prediction for appearance-related attributes in Table 9 keeping unchanged during editing.

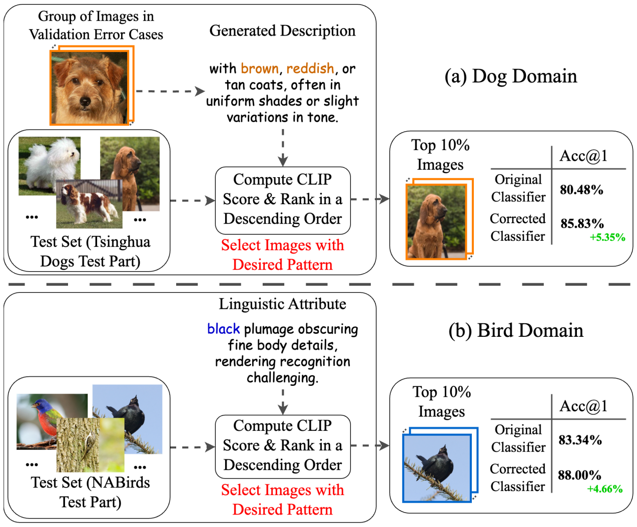

Group of Images in Validation Error Cases
Generated Description
with brown, reddish, or tan coats, often in uniform shades or slight variations in tone.
(a) Dog Domain
Top 10% Images
Acc@1
Original Classifier
80.48%
Corrected Classifier
85.83%
+5.35%
Compute CLIP Score & Rank in a Descending Order
Select Images with Desired Pattern
Test Set (Tsinghua Dogs Test Part)
Linguistic Attribute
black plumage obscuring fine body details, rendering recognition challenging.
(b) Bird Domain
Top 10% Images
Acc@1
Original Classifier
83.34%
Corrected Classifier
88.00%
+4.66%
Compute CLIP Score & Rank in a Descending Order
Select Images with Desired Pattern
Test Set (NABirds Test Part)

Figure 6. **Test Set-Level Robustness Evaluation for Corrected Classifier** to the "overall orange color" implicit image pattern in (a) dog domain and the "black plumage" linguistic attribute in (b) bird domain selected from Figure 3 of the Main Text. Original classifier is trained under the training parts of Tsinghua Dogs and NABirds, and the training set of corrected classifier is augmented with "overall orange color" and "black plumage" pattern respectively. Specifically, to select images from *test sets* with desired pattern, we compute the CLIP Score between the text description and the whole test sets, rank in a descending order, and choose top 10% images to evaluate two classifiers. Here, it shows after retraining classifiers with augmented training set containing desired pattern, the robustness to such patterns improves significantly.

4.3. Failure Cases

As shown in Figure 5, MDC as well as the baselines fails in simulating some patterns, *i.e*. removing the "orange hue" from the image. Specifically, we randomly select a group of images for MDC to simulate. After editing, we expect to remove such "orange hue" effect from the original image. However, MDC fails though such pattern is reduced to some extent.

5. Additional Set Level Robustness Evaluation

In Figure 5 of the Main Text, we evaluate the robustness of corrected classifier from different perspectives. In this section, we conduct further set-level robustness evaluation over real *test images* of animal species domains, *i.e*. the test parts of Tsinghua Dogs and NABirds. As shown in Figure 6, we select Top 1 attributes from Figure 3 of the Main Text for model correction and classifier retraining, which the strategies are kept the same as main text. To evaluate whether the corrected classifier become more robust to the pattern of selected attributes, we attempt to grab images with such pattern from the test set. Specifically, we compute the CLIP Score between the text description and the whole test sets, rank in a descending order, and choose top 10% images to evaluate. Results in Figure 6 demonstrate the corrected classifier indeed become more robust to the images with desired attribute patterns, verifying the usefulness and effectiveness of our selected causal attributes.

6. More Successful Editing Cases

Our editing model can successfully respond to more editing requests beyond the causal attributes discovered by our MDC system. In Figure 7, we display some qualitative editing results in bird domain. We can observe our edited images successfully follow the text guidance, demonstrating the capability of our editing model.

7. Experimental Setting Details

Additional Baselines. We select additional state-of-the-art language-guided editing baselines, *i.e*. SmartEdit [6], CDS [12], and Hive [27] for comparison.

Datasets Recapture. (1) Animal species classification: The training parts of Stanford Dogs [8] and CUB200 [22] are set as our *training sets*, corresponding testing parts as our *validation sets*. The training parts of Tsinghua Dogs [29] and NABirds [21] datasets are leveraged for targeted editing and the corresponding test parts are served as our *test sets*. (2) Face anti-spoofing: Combining OULU-NPU [2] and CASIA-MFSD [28] as *training set*, setting Idiap Replay Attack [4] as *validation set*, using MSU-MFSD [23] as *test set*.

Model	Stanford Dogs	Tsinghua Dogs	CUB200	NABirds
Our classifier	90.80%	88.20%	90.54%	91.99%
He et al. [5]	-	88.03%	91.70%	-
Xu et al. [25]	91.80%	-	91.80%	90.80%

Table 6. **Comparison of our Fine-Grained Classifiers with SOTAs** on four animal species datasets. The results are reported in Top-1 accuracy. Here, it shows our classifier achieves competitive performance with SOTAs, setting a good baseline for diagnosis.

Implementation Details. *1) Classifier Training:* To obtain the classifier having competitive performance with state-of-the-art works, in dog and bird domains, we leverage the top-performing ImageNet-22K pre-trained ViT-L/14 model [1] and fine-tune it with *training sets* for 100 epochs. We compare our trained classifier with some state-of-the-art works [5, 25] in Table 6. For the face anti-spoofing domain, we fine-tune a pretrained ViT-B/16 model from CLIP [18] for 500 steps with the batch size as 32. *2) Candidate Attribute Discovery.* We set N_{max} as 20 and unsupervised clustering method as K-Means and K equals 3. *3) Soft Token Learning.* We set q as 3 and λ_{reg} as 0.85 [26] and the optimizing process only takes approximately 5 minutes with 1 A100 GPU. *4) Editing Model Training.* Following [17], in fine-grained

[1] timm/eva02_large_patch14_448.mim_m38m_ft_in22k_in1k

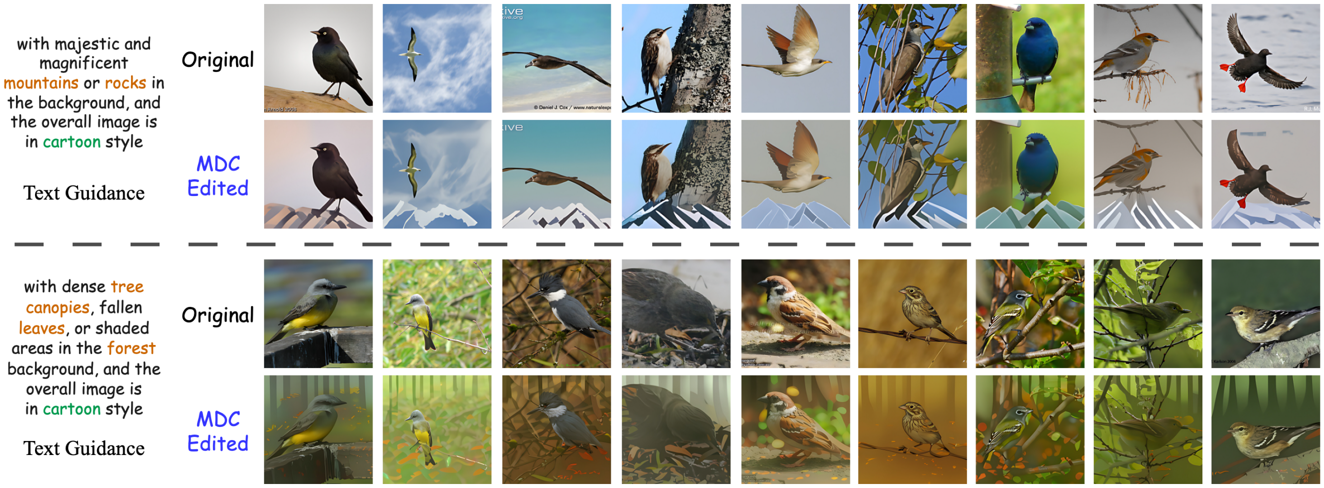

with majestic and magnificent mountains or rocks in the background, and the overall image is in cartoon style
Text Guidance
Original
MDC Edited
with dense tree canopies, fallen leaves, or shaded areas in the forest background, and the overall image is in cartoon style
Text Guidance
Original
MDC Edited

Figure 7. **Some Qualitative Editing Examples** beyond the causal attributes discovered by MDC in bird domain. Here, it shows our editing model owns the capability in response to other text guidance request.

classification domain, we firstly conduct a pre-training step on ImageNet-22K for 400M steps for better initialization, while for face anti-spoofing domain, we directly pick up a checkpoint pre-trained on FFHQ [7]. During the model training stage, we freeze the U-Net, update all mapping networks simultaneously, and set the semantic encoder as trainable to increase representation power and adaptability to various attribute semantics. We train the model for 300K steps with batch size as 1, which takes around 12 hours with 32 A100 GPU. We set s as 1, T_{edit} as 10 for training, 200 for inference.

8. Candidate Attributes Discovery Details

In this section, we provide more details of our candidate attribute discovery stage. After categorizing the error cases into small sub-groups using our attribute discovery model (ADM) [1] and further merging these sub-groups into larger K groups of images, we prompt ADM with two more requests as mentioned in Section 3.1 of the Main Text to acquire the linguistic attributes. From Figure 8 to 10, we can observe the details of linguistic attributes discovery from implicit group of images in dog, bird, and face anti-spoofing domains. Specifically, from classified image groups, we initially leverage ADM to summarize the image pattern into "one phrase" (column 2). Then, we prompt ADM to list some other types of reasons (column 3) that could cause misclassification based on that phrase. In the last column of each figure, we also list the generated descriptions for every image patterns.

9. Appearance-Related Attributes Details

To evaluate whether other unrelated attributes remain unchanged during editing, we compute the ratio of CLIP's prediction keeping unchanged between the original and counterfactual edited validation sets for these attributes. Specifically, for dog and bird domain, we prompt MLLMs to generate dozens of appearance-related attributes listed in Table 7 and 8. For face anti-spoofing domain, we directly leverage 40 CelebA [11] attributes shown in Table 9.

For example, if we want to compute the ratio of attribute "eyeglasses" unchanged between the original and one counterfactual edited set, we pick one image from the original set and construct the input text of CLIP as ["a photo of a

Category	Attributes
Coat Color	Black, Brown, White, Grey, Golden, Spotted, Brindle
Coat Length	Short, Medium, Long, Curly, Wiry
Coat Texture	Smooth, Rough, Silky, Dense
Size	Small, Medium, Large, Giant
Body Shape	Stocky, Slender, Muscular, Lean
Ear Type	Floppy, Erect, Semi-erect, Cropped
Tail Type	Long, Short, Curled, Straight, Docked
Eye Color	Brown, Blue, Hazel, Amber, Mismatched
Snout Shape	Short, Medium, Long
Leg Length	Short, Long, Proportional
Facial Expression	Alert, Friendly, Serious, Droopy
Markings	Spots, Stripes, Patches, Masks, Saddle-like markings

Table 7. **Appearance-Related Attributes for Dog Domain.**

Category	Attributes
Body Size	Small, Medium, Large
Plumage Color	Black, White, Brown
Plumage Pattern	Stripes, Spots, Streaks
Plumage Texture	Glossy, Dull, Fluffy
Beak	Small, Medium, Large
Tail	Forked tail, Rounded tail, Pointed tail
Wings Size	Small, Medium, Large
Wings Shape	Rounded and short, Long and narrow, Broad
Leg shape	Perching, Grasping, Wading
Eye color	Dark brown or black, Bright yellow, Red or reddish-orange
Posture	Upright, Horizontal, Crouched

Table 8. **Appearance-Related Attributes for Bird Domain.**

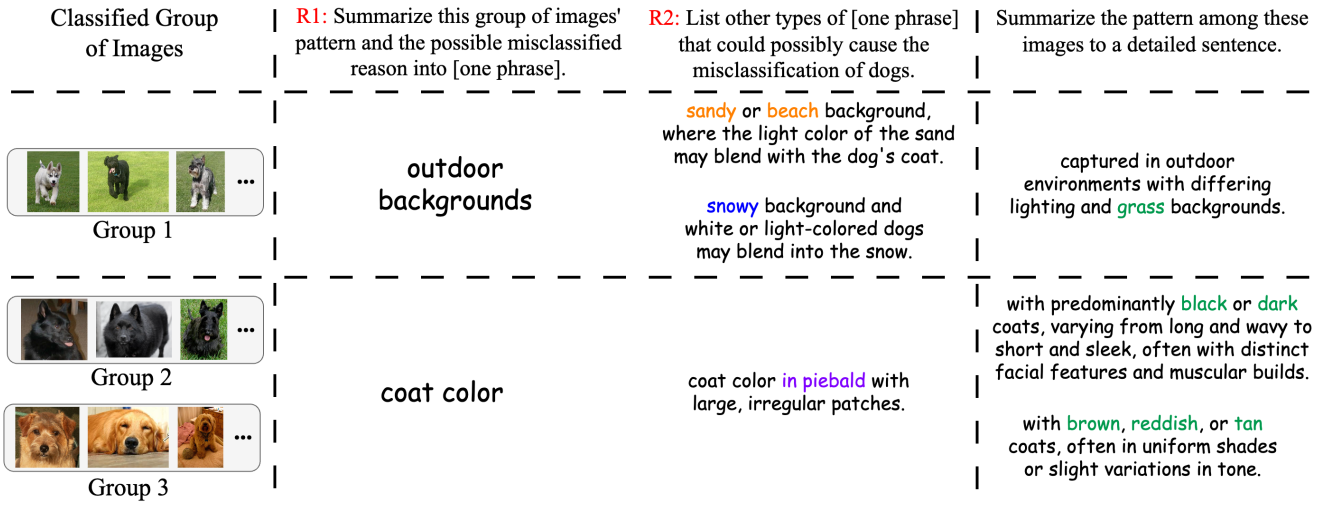

Classified Group of Images
R1: Summarize this group of images' pattern and the possible misclassified reason into [one phrase].
R2: List other types of [one phrase] that could possibly cause the misclassification of dogs.
Summarize the pattern among these images to a detailed sentence.
Group 1
outdoor backgrounds
sandy or beach background, where the light color of the sand may blend with the dog's coat.
snowy background and white or light-colored dogs may blend into the snow.
captured in outdoor environments with differing lighting and grass backgrounds.
Group 2
Group 3
coat color
coat color in piebald with large, irregular patches.
with predominantly black or dark coats, varying from long and wavy to short and sleek, often with distinct facial features and muscular builds.
with brown, reddish, or tan coats, often in uniform shades or slight variations in tone.

Figure 8. **Details of Linguistic Attributes Discovery from Implicit Group of Images** in the *validation set*, *i.e*. Stanford Dogs test set, of dog domain and corresponding generated descriptions (column 4), which are further used as the guidance for the language-guided editing baselines. From classified image groups, we initially leverage ADM to summarize the image pattern into "one phrase" (column 2). Then, we prompt ADM to list some other types of reasons (column 3) that could cause misclassification based on that phrase.

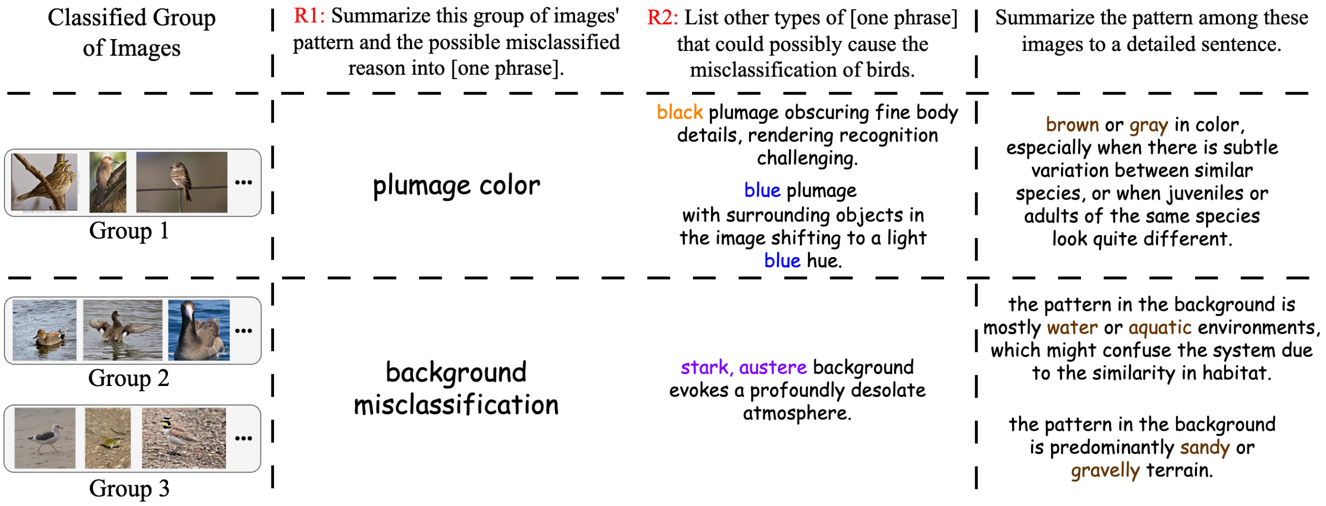
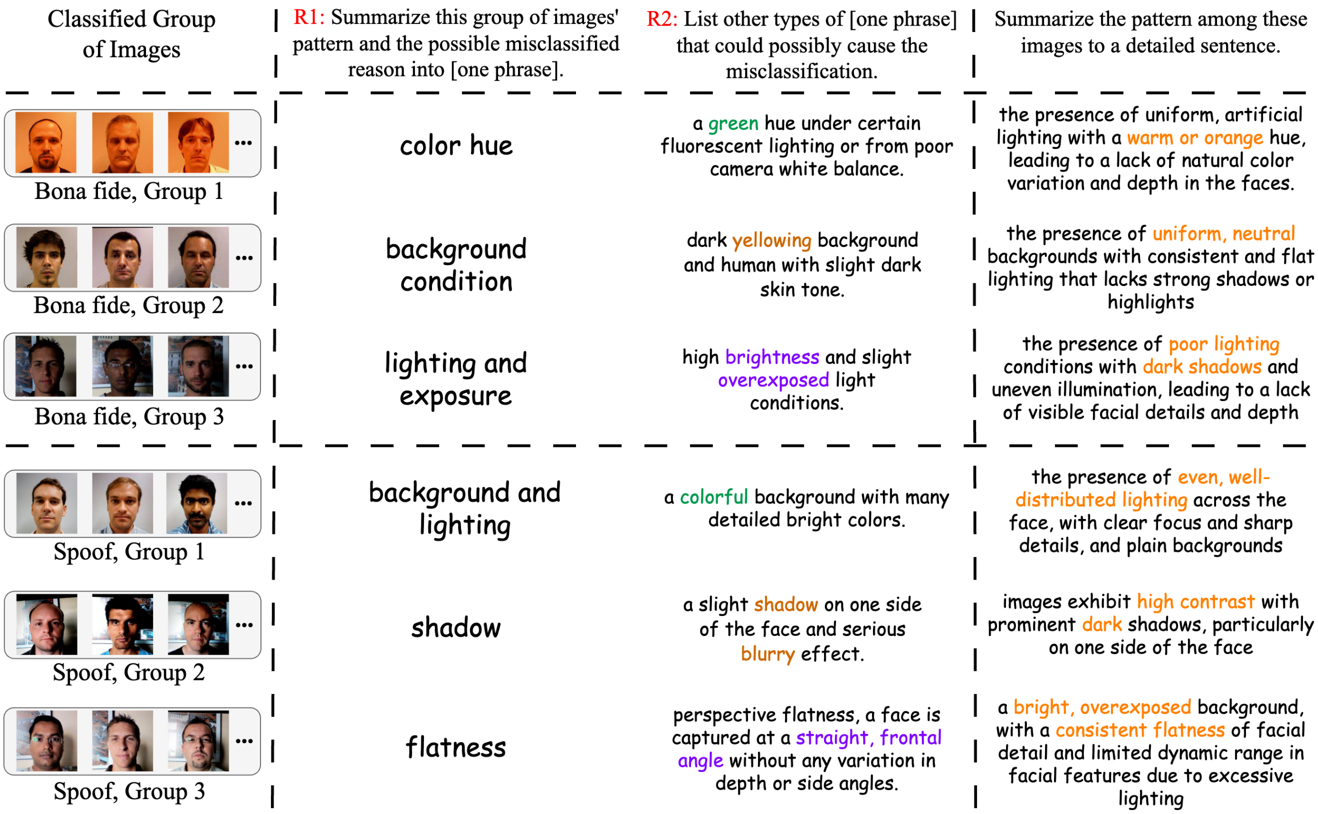

Classified Group of Images
R1: Summarize this group of images' pattern and the possible misclassified reason into [one phrase].
R2: List other types of [one phrase] that could possibly cause the misclassification of birds.
Summarize the pattern among these images to a detailed sentence.
Group 1
plumage color
black plumage obscuring fine body details, rendering recognition challenging.
blue plumage with surrounding objects in the image shifting to a light blue hue.
brown or gray in color, especially when there is subtle variation between similar species, or when juveniles or adults of the same species look quite different.
Group 2
Group 3
background misclassification
stark, austere background evokes a profoundly desolate atmosphere.
the pattern in the background is mostly water or aquatic environments, which might confuse the system due to the similarity in habitat.
the pattern in the background is predominantly sandy or gravelly terrain.

Figure 9. **Details of Linguistic Attributes Discovery from Implicit Group of Images** in the *validation set*, *i.e*. CUB200 test set, of bird domain and corresponding generated descriptions.

person", "a photo of a person with eyeglasses"]. We then compare the CLIP *argmax* predictions for this original image and corresponding paired image from counterfactual set using the same CLIP text input. If the predictions are the same, we consider this "eyeglasses" appearance-related attribute as unchanged for this image. We iterate all appearance-related attributes, all images in the dataset and average the ratio as the final score for these paired datasets. Corresponding results are shown in Table 3 of the Main Text, Table 4 and 5.

CelebA [11] Attributes
5'o Clock Shadow, Arched Eyebrows, Attractive, Bags Under Eyes, Bald, Bangs, Big Lips, Big Nose, Black Hair, Blond Hair, Blurry, Brown Hair, Bushy Eyebrows, Chubby, Double Chin, Eyeglasses, Goatee, Gray Hair, Heavy Makeup, High Cheekbones, Male, Mouth Slightly Open, Mustache, Narrow Eyes, No Beard, Oval Face, Pale Skin, Pointy Nose, Receding Hairline, Rosy Cheeks, Sideburns, Smiling, Straight Hair, Wavy Hair, Wearing Earrings, Wearing Hat, Wearing Lipstick, Wearing Necklace, Wearing Necktie, Young

Table 9. **Appearance-Related Attributes for Face Domain.**



Classified Group of Images	R1: Summarize this group of images' pattern and the possible misclassified reason into [one phrase].	R2: List other types of [one phrase] that could possibly cause the misclassification.	Summarize the pattern among these images to a detailed sentence.
Bona fide, Group 1	color hue	a green hue under certain fluorescent lighting or from poor camera white balance.	the presence of uniform, artificial lighting with a warm or orange hue, leading to a lack of natural color variation and depth in the faces.
Bona fide, Group 2	background condition	dark yellowing background and human with slight dark skin tone.	the presence of uniform, neutral backgrounds with consistent and flat lighting that lacks strong shadows or highlights
Bona fide, Group 3	lighting and exposure	high brightness and slight overexposed light conditions.	the presence of poor lighting conditions with dark shadows and uneven illumination, leading to a lack of visible facial details and depth
Spoof, Group 1	background and lighting	a colorful background with many detailed bright colors.	the presence of even, well-distributed lighting across the face, with clear focus and sharp details, and plain backgrounds
Spoof, Group 2	shadow	a slight shadow on one side of the face and serious blurry effect.	images exhibit high contrast with prominent dark shadows, particularly on one side of the face
Spoof, Group 3	flatness	perspective flatness, a face is captured at a straight, frontal angle without any variation in depth or side angles.	a bright, overexposed background, with a consistent flatness of facial detail and limited dynamic range in facial features due to excessive lighting

Figure 10. **Details of Linguistic Attributes Discovery from Implicit Group of Images** in the *validation set*, *i.e*. Idiap, of face anti-spoofing domain and corresponding generated descriptions.

10. Limitations & Future Work

Despite the superior performance, MDC has several limitations. Firstly, our editing model lacks of zero-shot inference capability. Currently, for each specific task, MDC requires to retrain the editing model and learn multiple mapping networks, each taking charge of one specific editing request. In the future, we may construct a universal editing model, trained using a much larger dataset, *e.g*. LAION-5B [19] or built upon a more powerful foundation generation model, *e.g*. SDXL [15]. With such editing model, we eliminate the need of retraining with different training data for every specific task. Another limitation is though MDC is orthogonal to any computer vision task, we only verify its capability in image classification. Therefore, another future work is to employ MDC to other computer vision tasks, *e.g*. object detection, segmentation. In this way, we can adopt our system to more computer vision tasks and enhance the performance of relevant models.

References

[1] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024. 6

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017. 5

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[4] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012. 5

[5] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *AAAI*, 2022. 5

[6] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 3, 5

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019. 6

[8] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop*, 2011. 5

[9] David Lewis. Causation. *The journal of philosophy*, 70(17): 556–567, 1973. 1

[10] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7817–7826, 2024. 3

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6, 7

[12] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9192–9201, 2024. 3, 5

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2

[14] Judea Pearl. *Causality*. Cambridge university press, 2009. 1

[15] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 8

[16] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. *Advances in Neural Information Processing Systems*, 36:25165–25184, 2023. 2

[17] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022. 5

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 8

[20] Linoy Tsaban and Apolinário Passos. Ledits: Real image editing with ddpm inversion and semantic guidance. *arXiv preprint arXiv:2307.00522*, 2023. 3

[21] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015. 5

[22] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[23] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 5

[24] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023. 3

[25] Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, 25:9015–9028, 2023. 5

[26] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 5

[27] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024. 3, 5

[28] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012. 5

[29] Ding-Nan Zou, Song-Hai Zhang, Tai-Jiang Mu, and Min Zhang. A new dataset of dog breed images and a benchmark for finegrained classification. *Computational Visual Media*, 6: 477–487, 2020. 5