When a Zero-Shooter Cheats: Improving Age Estimation via Activation Steering

Erik Imgrund
BIFOLD & TU Berlin

Pia Hanfeld
BIFOLD & TU Berlin

Klim Kireev
BIFOLD & TU Berlin

Konrad Rieck
BIFOLD & TU Berlin

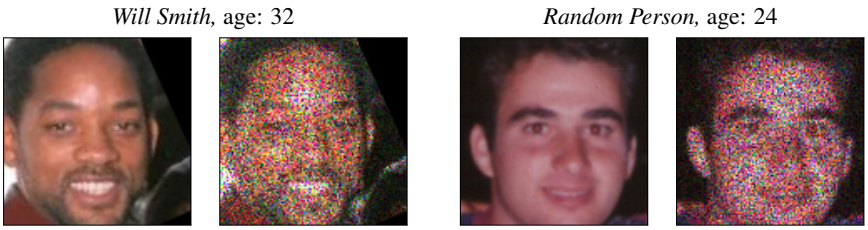


Figure 1: Example of the identity shortcut. The age of Will Smith is consistently mispredicted by Gemma 3 to be close to his age at the knowledge cutoff. The effect persists even in the presence of severe noise, which causes deviations for random people that the model does not recognize. The random person is sampled from the FG-Net dataset, the image of Will Smith from AgeDB.

Abstract

Different age-related regulations have been proposed to protect minors from harmful content and interactions online. Automated age estimation is central to enforcing such regulations, and vision-language models (VLMs) achieve state-of-the-art performance on this task. However, we find that the zero-shot nature of VLM-based age estimation produces an unexpected side effect we call the *identity shortcut*: Instead of estimating age from visual features, VLMs tend to identify the depicted person and infer their age from memorized knowledge. This phenomenon leads to substantially incorrect predictions when non-celebrities are misidentified as celebrities. It also produces deceptively high robustness to noise and adversarial perturbations on celebrity images, which dominate popular benchmarks. To mitigate this, we propose an *activation steering* method that suppresses the shortcut by intervening on the hidden states of the VLM. This method improves age estimation accuracy for both memorized and unseen identities, reducing mean absolute error by up to 25% across popular benchmarks.

1 Introduction

Age estimation from facial images has quietly become safety-critical infrastructure. New legislation in the EU and Australia [Commonwealth of Australia, 2024, European Commission, 2025] require platforms to gate access for minors, and at the scale of deployment, such gating needs to be performed by automated age estimators [Hanaoka et al., 2024, Age Check, 2021][1]. Vision-language models

[1] Non-exhaustive list of companies providing certified age verification: Yoti, ROC AI, Scytáles, Sumsub.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

(VLMs) have recently surpassed domain-specific image classifiers for this task, demonstrating vastly improved capabilities over the traditional fully-supervised models. However, due to the lack of internal transparency, this improved performance may come at the cost of unexpected behavior.

An example of such counterintuitive behavior is illustrated in Figure 1. Rather than estimating the age of the depicted person from visual features, the VLM first identifies the subject in the photo and reports their current age from memory, producing a strikingly incorrect prediction. Moreover, unlike any traditional detection method, the prediction does not change even in the presence of high-amplitude noise, implying a persistent underlying bias.

We call this phenomenon the *identity shortcut*: instead of inferring the queried attribute from visual evidence, the VLM first attempts to recognize the subject and reports an attribute associated with it from memory. This shortcut has severe consequences: First, it causes prediction errors on celebrity images, where the model returns an age tied to the memorized identity. Second, it degrades performance on non-celebrities who visually resemble known figures, as the model anchors its prediction on the closest identity. Third, and most insidiously, it manufactures *deceptive robustness*. Models appear exceptionally resilient to perturbations on popular benchmarks, but only because those benchmarks are dominated by celebrity images.

In this paper, we show that this phenomenon is not limited to a specific model or dataset, but instead is a general issue across VLMs. To mitigate it, we introduce an inference-time activation steering framework that identifies the hidden-state directions responsible for identity-based retrieval and suppresses them, forcing the model to condition its prediction on visual evidence rather than on a recalled identity. Extensive evaluations across state-of-the-art VLMs demonstrate that our intervention not only eliminates the asymmetric robustness between memorized and unseen identities but also reduces mean absolute error by 10.8% on average (up to 25.3%) across existing benchmarks.

In summary, we make the following contributions:

- **Identity shortcut:** We introduce the phenomenon of the *identity shortcut*, where VLMs prioritize identity retrieval over visual feature analysis. This effect is responsible for systematic performance degradation and can inflate benchmark results.
- **Deceptive Robustness:** We show that zero-shot VLMs exhibit fundamentally different behavior under perturbations than traditional methods, as their inference can be guided by retrieval rather than visual analysis, leading to misleading conclusions about their robustness.
- **Inference-Time Mitigation:** We propose an intervention that steers the model's activations to suppress the identity shortcut. Our method improves accuracy and restores expected robustness, ensuring that age estimates are grounded in visual evidence rather than memorized associations.

2 Preliminaries

To contextualize our discovery of the identity shortcut, we first review prior work on age estimation and the robustness of VLMs. We then describe the datasets and models that constitute the foundation of our empirical evaluation.

2.1 Related Work

Our work sits at the intersection of three active directions in research: methods for automated age estimation from facial images, the robustness of VLMs to input perturbations, and the use of VLMs as zero-shot predictors for image understanding tasks.

Age Estimation. Age estimation is an important image understanding task with many applications including enforcing age-based restrictions [Hanaoka et al., 2024], data mining [Ni et al., 2009] or removing underage pictures from image generation datasets [Cretu et al., 2025, Gaul et al., 2026]. Pan et al. [2018] use convolutional neural networks (CNNs) to perform age estimation based on a cropped image of a person's face. Kuprashevich and Tolstykh [2023] achieve better performance using a vision transformer and combine facial and body images. This approach is further extended by Kuprashevich et al. [2024], achieving the best performance of any specialized model for age estimation. More recently, Ren et al. [2026] show that VLMs achieve even lower error despite lacking any task-specific training. In a related line of research, Szu-Tu et al. [2025] show that VLMs achieve

high performance in estimating the age of buildings through memorization of famous landmarks. We extend this work by quantifying how identity-based biases degrade VLM age estimation and propose a mitigation strategy that restores general performance.

Robustness of VLMs. VLMs have been utilized widely for multimodal input analysis across multiple domains [Li et al., 2025]. While most research uses benchmarks that evaluate performance under near-ideal scenarios [Yin et al., 2023, Li et al., 2024, Liu et al., 2024b, Fu et al., 2026], real-world deployments demand accuracy even on degraded inputs, such as blurry or noisy uploads [Ye et al., 2024]. Following this work, Usama et al. [2025] show that VLMs are not robust towards common noise corruptions. Regarding adversarial robustness, Wang et al. [2025] demonstrate that adversarial training can restore robust performance to almost clean performance, while Cui et al. [2024] show that appending LLM-generated descriptions to prompts enhances resistance to white-box attacks. Meanwhile, Latif et al. [2026] improve robustness against visual and textual corruption functions by preprocessing the input to the VLM through expert generative denoising models.

However, existing work focuses primarily on input corruptions, implicitly assuming a homogeneous data distribution. We extend this line of work by first investigating the robustness in a regression task, namely age estimation, and comparing robustness properties across different data distributions. We also demonstrate that the robustness of VLMs is dependent on the data distribution, revealing that current state-of-the-art benchmarks provide a false sense of security.

Zero-shot Reasoning Capabilities. Nagar et al. [2024] raise the question of whether state-of-the-art visual question answering (VQA) benchmarks, developed with CNNs and ViTs in mind, allow meaningful judgement of the actual visual reasoning capabilities of VLMs. They design a benchmark on new synthetic data to asses if the zero-shot performance of the models relies on reasoning instead of memorized knowledge. Overall, all evaluated models perform significantly worse compared to a human test group in this study.

Pombal et al. [2025] introduce an automatic benchmark for VLMs on a validation set of a public VQA dataset. One VLM is tasked to generate a new question-answer pair from an input image which is then judged by a second model. Their approach automates the generation of VQA tasks for benchmarking from existing, real-world datasets. However, their approach does not explicitly analyze the visual reasoning capabilities of the models. Rizwan et al. [2025] analyze the capabilities of several VLMs to classify potentially offensive memes. They observe that most models they analyze are rigid about their initial classification even under partial occlusion.

Overall, extensive work analyzing VLM zero-shot performance remains scarce, with only a few publications explicitly distinguishing between world knowledge and visual reasoning capabilities. Building on these concerns, we specifically investigate how identity-based memorization compromises zero-shot age estimation.

2.2 Experimental Setup.

In this work, we follow the standard evaluation procedure established in the age estimation literature, relying on publicly available datasets and models.

Datasets. In particular, we use the following three public datasets and one self-collected dataset throughout our study.

- **CelebA** [Liu et al., 2015], a widely-used dataset containing over 200,000 images of over 10,000 celebrities annotated with 40 binary attributes,
- **AgeDB** [Moschoglou et al., 2017], consisting of over 16,000 images captured under uncontrolled, real-world conditions of 570 celebrities labeled with age and gender,
- **FG-Net** [Lanitis et al., 2002], composed of 1,002 images of 82 non-celebrities captured at different stages of their lives,
- **VideoScrape**, a dataset consisting of 200 images that we manually scrape from a video platform after the knowledge cutoff of the considered VLMs. We describe it in Appendix A.4.

In order to evaluate our method on roughly equal distributions regarding age and gender, we first gather demographics from our VideoScrape dataset and randomly sample as many images from AgeDB and FG-Net. The demographics for our evaluation sets are reported in Table 8 in Appendix A.4.

Models. We study different age estimators: six zero-shot vision-language models prompted for age estimation [Liu et al., 2024a, Wang et al., 2024, Team, 2025, Qwen Team, 2026, Google DeepMind, 2026], and two fully-supervised methods: the state-of-the-art vision transformer MiVOLO [Kuprashevich et al., 2024] and a ResNet-18 trained for age estimation according to Pan et al. [2018]. We summarize the information about evaluated methods in Table 1.

Table 1: Characteristics of the age estimation methods used in our study. The type is either fully-supervised (FS) or zero-shot (ZS). We did not include Gemini 3 Flash in the table, as the details are mostly proprietary.

	Task-specific		**Vision-language models**				
	CNN	MiVOLO	LLaVA 1.5	Qwen-VL 2.5	Gemma 3	Qwen 3.5	Gemma 4
Parameters	12 M	27 M	7 B	7 B	4 B	9 B	31 B
Release	2018	2023	2023	01/25	03/25	02/26	04/26
Setting	FS	FS	ZS	ZS	ZS	ZS	ZS

3 Identity Shortcut

Age estimation systems are assumed to operate on visual features of the input image (face, posture, context, etc.). However, since VLMs are pretrained on vast corpora of data containing factual information about people, they can take a *shortcut* by identifying the depicted person and then performing factual inference from their memorized world knowledge. While one might assume that this mechanism aids performance by retrieving the correct age for identified individuals, it fundamentally deviates from the intended task of visual age estimation. Specifically, returning a person's current age based on the model's knowledge cutoff introduces systematic temporal errors.

Even worse, we find that VLMs frequently attempt to assign identities to non-public figures as well, leading to hallucinated identifications and consequently erroneous age predictions. In this section, we analyze the presence and impact of this *identity shortcut*.

3.1 Definitions.

Given an age estimator f, we define $f(x \mid \mathrm{k})$ as its output on an input image x drawn from an image distribution $\mathbb{X}$, with knowledge of the depicted person. Similarly, $f(x \mid \neg\mathrm{k})$ denotes that a person on an image x is unknown to the estimator. $\mathbb{K}_f$ in our notation is the distribution of images containing people known to the estimator. We then define the expected impact δ_k of the identity shortcut for an age estimator f as

$$\delta_k = \mathop{\mathbb{E}}_{x \sim \mathbb{X}} \left| f(x \mid k) - f(x \mid \neg k) \right|. \tag{1}$$

Unfortunately, the direct computation of this term is infeasible, as it would require constructing a model that memorizes exactly the same set of people except for the person shown in the image x. Consequently, we would need to either modify x so that it is no longer identified correctly, or modify the model itself. Neither option is practical. Instead, we introduce a surrogate estimator $g(\cdot \mid \neg k)$ that is guaranteed not to have specific knowledge of any individual and compute

$$\mathop{\mathbb{E}}_{x \sim \mathbb{K}_f} \left| f(x \mid k) - g(x \mid \neg k) \right| \approx \delta_k + E_{f,g}, \tag{2}$$

which is accurate up to a modeling error $E_{f,g}$ based on the mismatch between the surrogate and actual estimator. Then, we approximate this term on the distribution of images with unknown identities via

$$E_{f,g} = \mathop{\mathbb{E}}_{x \sim \mathbb{X} \setminus \mathbb{K}_f} \left| f(x \mid \neg k) - g(x \mid \neg k) \right|. \tag{3}$$

3.2 Impact on Known Identities

In order to quantify the effect of memorized identity knowledge on the prediction of VLMs, we split the CelebA [Liu et al., 2015] dataset into two parts, the first one consists of randomly sampled images, and the second one of images where a VLM correctly identifies a celebrity. We measure the difference between the VLMs' age predictions and a baseline age estimator (MiVOLO [Kuprashevich et al., 2024]) that was not trained with knowledge of any identities. The choice of MiVOLO as the baseline estimator is dictated by the following: MiVOLO was trained on manually annotated samples from the Open Images dataset and production data of their age estimation service and both of them are unlikely to have celebrity images. The Open Images dataset contains permissively licensed images from Flickr filtered for images appearing on any other site and therefore have a low probability of including any celebrities. The resulting modeling error and identity shortcut impact δ_k are reported in Table 2.

Table 2: Impact of the identity shortcut δ_k estimated using Equation (2) and the modeling error $E_{f,g}$. All shortcut impacts higher than 1-sigma are shown in bold.

Model	$\delta_k + E_{f,g}$	$E_{f,g}$	δ_k
LLaVa 1.5	3.71 ± 0.20	3.46 ± 0.17	**0.25**
QwenVL 2.5	4.57 ± 0.24	4.16 ± 0.21	**0.41**
Gemma 3	5.55 ± 0.25	4.39 ± 0.21	**1.16**
Qwen 3.5	7.82 ± 0.37	7.89 ± 0.32	-0.07
Gemma 4	3.32 ± 0.15	3.05 ± 0.14	**0.27**
Gemini 3 Flash	4.48 ± 0.37	3.03 ± 0.25	**1.45**

It is immediately apparent that the identity shortcut has a measurable impact on most VLMs. The effect does not show any trend along model size or recency and is strongest for *Gemma 3* and *Gemini 3 Flash*, where its expected impact on the model's prediction is greater than one year. While the outlying behavior of *Qwen 3.5* looks counter-intuitive, we explain it in the following way: *Qwen 3.5 always* takes the identity shortcut for images in CelebA, even when not correctly identifying them. Indeed, its disagreement with MiVOLO is the highest on both sets, and as we show in Section 4, it benefits from the identity shortcut removal the most.

3.3 Robustness across Distributions

Besides the performance impact, we also discover that the identity shortcut has an asymmetric effect on model robustness. We estimate the robustness by applying a set of common corruptions $\mathcal{C}$ introduced by [Hendrycks and Dietterich, 2019] with different noise levels and measure the change in the age estimate. More specifically, we compute the mean deviation across all corruptions. We normalize the deviations of each corruption by the maximum deviation across all data points of all datasets to weigh each corruption inversely proportional to its severity. More formally, we compute

$$\mathop{\mathbb{E}}_{c \in \mathcal{C}, x \in \mathbb{X}} \frac{|f(x) - f(c(x))|}{\max_{x' \in \mathbb{X}} |f(x') - f(c(x'))|}. \tag{4}$$

In order to distinguish natural asymmetry on different datasets from the deceptive robustness caused by the identity shortcut, we perform this experiment on multiple datasets (CelebA-K, AgeDB, FG-Net and VideoScrape).

In Figure 2, we report the mean deviation of all tested models across the different data distributions averaged over different corruptions and samples, normalized by the maximum deviation achieved for each corruption type. We compute the mean deviation across all corruption schemes, normalized by the maximum deviation achieved for each individual corruption scheme, so as to weigh them evenly. Both specialized models display nearly identical robustness across all datasets, confirming prior research on robustness under different data distributions. Most VLMs, in contrast, show non-uniform robustness with more fragility on AgeDB and FG-Net and far higher robustness on our dataset of known celebrities. *Qwen 3.5* is an outlier, showing lower robustness only for AgeDB and similar robustness across FG-Net, CelebA and VideoScrape. A reason for that could be that it recognizes more celebrities on AgeDB than the other datasets.

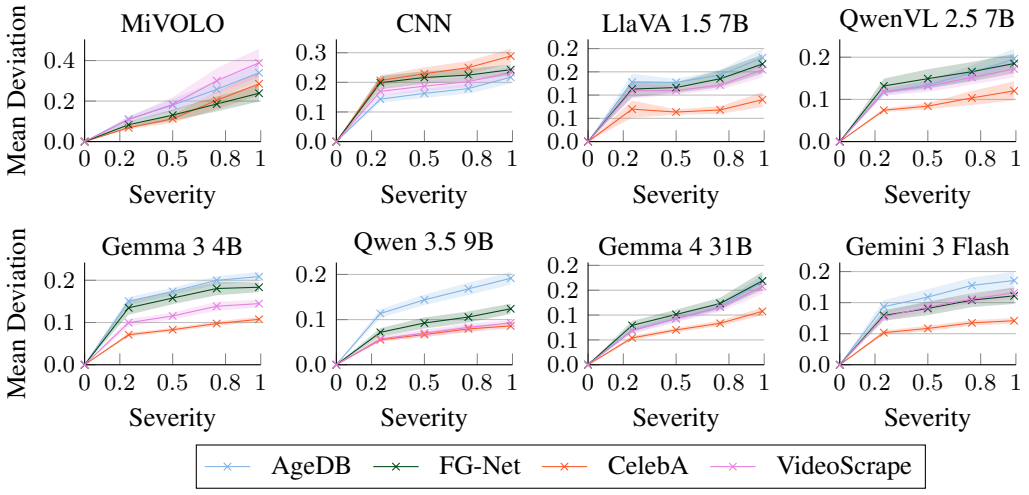


Figure 2: Robustness to common corruptions of different age estimation models. The deviation from the base prediction is averaged over all types of corruption, normalized by the maximum deviation for each corruption. The shaded area shows the standard error of the mean across all corruptions. Vision language models show asymmetrical robustness depending on the data distribution and demonstrate distinctively high robustness on the CelebA dataset.

Interestingly, robustness is similar on three different benchmark datasets (AgeDB, FG-Net, VideoScrape) for all VLMs except for *Gemma 3*. Since VideoScrape is guaranteed to be excluded from the training sets, we can deduce that the benchmark datasets have been excluded from the training corpora of the studied VLMs as well. Lastly, the increased robustness on CelebA indicates that the impact of the identity shortcut is less prone to change under image corruptions than normal predictions. This implies that the memorization effect persists even when noise of various means is applied. It is worth noting that this finding holds even under various modifications of the inference prompt that we explored in preliminary experiments across the studied vision language models. However, we did not observe a significant difference for the identity shortcut, so we conclude that prompt design is not a viable explanation for the performance discrepancy.

Besides the analysis presented above, we report the disaggregated results together with the adversarial robustness evaluation in Appendix C. In summary, the discovered trends present across the majority of corruptions for most of the models. Regarding adversarial robustness, we observe a similar discrepancy between CelebA and other datasets, though it is worth to mention how starkly non-robust all studied models are (all models are reliably attacked with $\varepsilon < 1/255$).

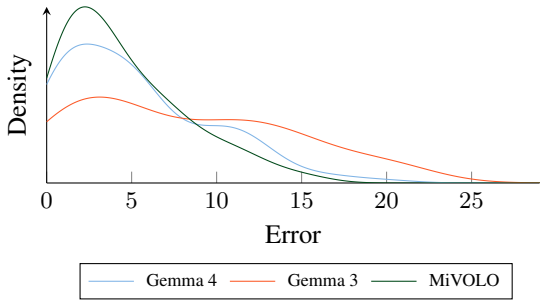


Figure 3: Distribution of mean absolute error of selected models on FG-Net. While MiVOLO's error distribution has a single pronounced peak, the error distribution of the VLMs is bimodal.

3.4 Impact on Unknown Identities

As we show in Sections 3.2 and 3.3, there is a significant difference in model behavior for images with known and unknown people for vision-language models. However, even within the unknown people distribution, there is a discrepancy between traditional methods such as MiVOLO and VLMs. In Figure 3, we show the model's error distribution on FG-Net, a dataset with a guaranteed absence of celebrities. The error distribution of MiVOLO behaves as expected, while the distribution for both *Gemma 3* and *Gemma 4* has a pronounced bimodal structure, which indicates that different behavior is at play even for unknown people in some cases. We hypothesize that models try to guess their identity and hence produce grossly incorrect predictions.

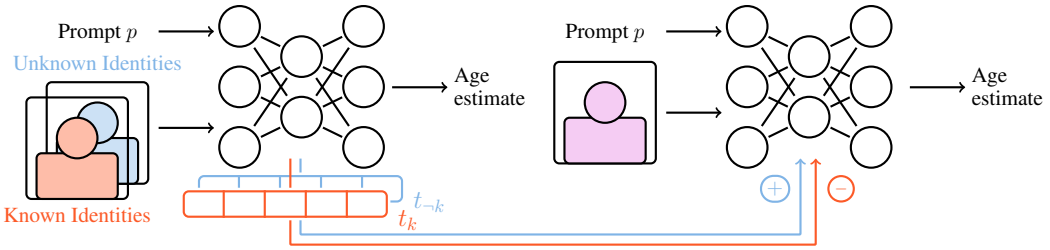


Figure 4: Overview of our task activation steering method. The VLM computes a task vector from the image and prompt, based on which the age is estimated. We modify this task vector by moving it towards the distribution of unknown identities, effectively disabling the identity shortcut.

As an additional experiment, we analyze the reasoning of *Qwen 3.5* and *Gemma 3* for evidence of identity usage. For Qwen 3.5, we use the native reasoning capabilities and for Gemma 3 we use chain-of-thought prompting. On the FG-Net dataset, which does not contain celebrity images by design, we find that 3.96% of the reasoning traces contain the keyword "identity" and 2.0% contain the keyword “identification,” providing a lower bound on reasoning that at least considers the identity and identification of the depicted person. Using an automatic analysis of the reasoning traces with Qwen 3.6 35B as a judge, we further find that 22.8% of the cases mention identifying the person as a strategy.

Interestingly, none of Gemma 3’s chains of thought explicitly mention identification as a strategy. Instead, we find that in 4% of the cases, the model skips chain-of-thought generation and immediately returns the result. In those specific cases, the prompt identification strategy described in Section A.2 also suggests that the model predicts a celebrity identity. For CelebA, we find that the pattern described above appears in 51.3% of the cases, providing further evidence for this hypothesis.

These results suggest that the identity shortcut is indeed being used and not an artifact of our methodology. For Gemma 3 in particular, we see the shortcut in action as a literal shortcut, resulting in skipped reasoning.

4 Task Activation Steering

To further analyze the mechanism behind the identity shortcut, we utilize task vectors, first introduced by Hendel et al. [2023]. The concept of task vectors is based on the idea of decomposing a language model f, which receives a task description p and an input sample x into two functions applied in sequence. The first function, $t(\cdot)$ is responsible for extracting a task vector from the task description which is then used to produce the final prediction, so that $f(x, p) = a(x, t(p))$. Based on our findings, we extend this definition to task vectors also dependent on the input, i.e. we decompose f by

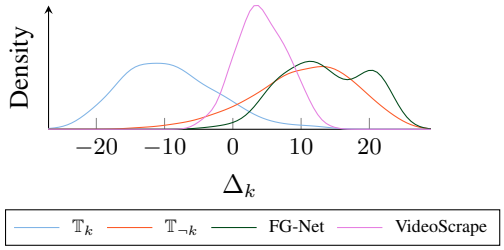
$$f(x) = a(x, t(x, p)). \tag{5}$$

This formulation allows us to describe the identity shortcut as a different task vector being produced depending on whether the identity is known or unknown. Thus, we define the task vector when the identity is known as t_k and when it is unknown as $t_{\neg k}$, resulting in

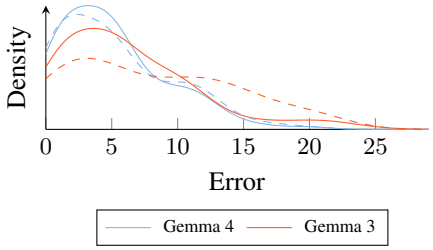
$$f(x \mid k) = a(x, t_k), \quad f(x \mid \neg k) = a(x, t_{\neg k}). \tag{6}$$

From an implementation point of view, we base our task vector computation on the findings by Hendel et al. [2023], which say that the shallow activations at the last token position of the prompt are responsible for encoding the task. While Hendel et al. [2023] use activations from a single layer, we instead take all activations of the first half of the model at this last token position as the task vector to get a robust representation and avoid the search over layer indices. Then, we can compute the task vector for known and unknown identities by

$$t_k = \mathop{\mathbb{E}}_{x \sim \mathbb{K}_f} t(x, p), \quad t_{\neg k} = \mathop{\mathbb{E}}_{x \sim \mathbb{X} \setminus \mathbb{K}_f} t(x, p). \tag{7}$$



(a) Density of different task vectors for Gemma 4 along the axis $t_k - t_{\neg k}$ of different datasets. FG-Net and VideoScrape are closer to unknown identities but still have some overlap with known identities.



(b) Error distribution of selected models on FG-Net with (solid line) and without (dashed) activation steering. The steered model error distribution becomes closer to normal.

Figure 5: Effects of the identity shortcut on photos containing unknown identities.

4.1 Task Vector Distribution

Utilizing the introduced task vectors, we can now investigate the cause of the unusual error distribution for non-celebrity images. We measure the number of samples for which the predicted task vector $t(x)$ is closer to the identity shortcut task vector t_k than the task vector for unknown identities $t_{\neg k}$. To express it as a single-dimensional value, we define the difference in distance to the task vectors

$$\Delta_k(t) = \|t - t_k\| - \|t - t_{\neg k}\|. \tag{8}$$

For task vectors closer to the identity shortcut we get $\Delta_k(t) < 0$ and for those further from the identity shortcut, we have $\Delta_k(t) > 0$. To account for the spread of task vectors, we additionally define the distributions $\mathbb{T}_k$ and $\mathbb{T}_{\neg k}$ for task vectors on known and unknown identities of AgeDB, respectively.

Figure 5a displays example distributions of task vectors computed from *Gemma 4* on known identities from AgeDB, unknown identities and two datasets consisting of unknown identities, FG-Net and VideoScrape. We see that there is a clear separation between known and unknown identities of AgeDB. The FG-Net distribution is close to the unknown identities distribution, while VideoScrape, is still closer to the unknown identities distribution, but situated further in between both. This indicates that *Gemma 4* does take the identity shortcut for some samples, even though the identity cannot be known by the model and must thus be misidentified. To show how many samples activate the identity shortcut, we test $\mathbb{P}(t \in \mathbb{T}_{\neg k}) < 0.1 \wedge \mathbb{P}(t \in \mathbb{T}_k) > 0.1$. The resulting ratios are shown in Table 3.

Table 3: Ratio of samples for which the task vector indicates usage of the identity shortcut.

Dataset	LLaVA 1.5	QwenVL 2.5	Gemma 3	Qwen 3.5	Gemma 4
FG-Net	6.98%	0.58%	0.00%	1.16%	1.16%
VideoScrape	6.82%	0.00%	0.91%	17.73%	10.91%

We see that all models have samples on which the identity shortcut is prominent, but *QwenVL 2.5* and *Gemma 3* have particularly small ratios. While there is no clear trend along the models, *Qwen 3.5* and *Gemma 4* have the highest ratios with *Qwen 3.5* activating the identity shortcut for nearly 18% of the samples in the VideoScrape dataset, additionally reinforcing our suggestion in Section 3.2, that *Qwen 3.5* is especially prone to the shortcut even on datasets without celebrities present.

4.2 Disabling the Identity Shortcut

As the identity shortcut has a detrimental impact on model performance, it is beneficial to remove it. One possible way to do so, would be to remove the identity knowledge from the model during the training process or via machine unlearning techniques. This approach, however, is either computationally heavy to implement (full retraining), or introduces additional loss of performance (machine unlearning). Among the inference-time techniques, prompt-engineering could be used. However as we discuss in Section 3.3, our attempts to change the prompt did not affect the model prediction, even

if we prompted the model to ignore its knowledge of the person or focus on the visual appearance in different ways. Instead, we propose to keep the model on the correct task, via an inference-time activation-level intervention. The high-level process is explained in Figure 4. We use the task vectors t_k and $t_{\neg k}$ to intervene in the hidden states and move the model from the identity shortcut distribution towards the distribution of task vectors with unknown identity. More concretely, we approximate age estimation with unknown identity via

$$f(x \mid \neg k) \approx a(x, t(x, p) + \alpha \cdot (t_{\neg k} - t_k)), \tag{9}$$

where α is a hyperparameter. We choose $\alpha = 3$ through manual tuning on a set of manually collected validation images of celebrities and non-celebrities that are not included in any of the benchmark datasets. We apply this activation steering to the VLMs during age estimation and compute the mean absolute error on AgeDB and FG-Net. The results are displayed in Table 4.

Table 4: Mean absolute error of vision language models on AgeDB and FG-Net in the default mode and with activation steering to fix shortcut mispredictions.

Dataset	Variant	LLaVA 1.5	QwenVL 2.5	Gemma 3	Qwen 3.5	Gemma 4
AgeDB	Default	9.21 ± 0.06	8.94 ± 0.06	8.48 ± 0.05	13.52 ± 0.09	**6.50** ± 0.04
	Steered	**8.93** ± 0.06	**7.78** ± 0.05	**8.01** ± 0.05	**10.72** ± 0.07	6.73 ± 0.04
FG-Net	Default	7.06 ± 0.50	9.03 ± 0.58	8.36 ± 0.57	8.80 ± 0.54	5.23 ± 0.38
	Steered	7.13 ± 0.49	**7.02** ± 0.50	**6.24** ± 0.45	**7.17** ± 0.48	**5.08** ± 0.35

The resulting errors show that steering the activations towards $t_{\neg k}$ and away from t_k does indeed help performance in almost all cases. The mean improvement across all setups is 10.8%, showing the measurable practical impact of the proposed method. *Gemma 3* sees the biggest improvement with a 25.36% reduction of error on FG-Net. This indicates that the identity shortcut causes mispredictions, both for celebrities on AgeDB and non-celebrities on FG-Net. Interestingly, we see no clear correlation between the ratio of samples for which the task vector indicates usage of the identity shortcut and the performance uplift. We suspect that older models, such as LLaVa 1.5 are not capable enough to perform well on age estimation and therefore removing the identity shortcut only reverts them back to their poor baseline performance. Lastly, Qwen 3.5 shows the nearly biggest improvement, providing additional evidence of our previous findings regarding this model.

Further analysis on AgeDB shows that improvements for correctly identified people are higher than for misidentified samples, confirming that the identity shortcut's impact on celebrities is higher than on non-celebrities. Nevertheless, improvements for all models except LLava 1.5 on FG-Net show that the effect on non-celebrities is prominent. It is worth mentioning that activation steering has little impact on the inference time and therefore has a negligible overhead for real-world deployment. In addition, we show the resulting error distribution when applying activation steering in Figure 5b. It is apparent that the bimodal nature of the error distribution is dampened. Therefore, we conclude that the identity shortcut is indeed the reason for its appearance in the first place.

5 Conclusion

It is hard to overlook the progress in zero-shot computer vision enabled by vision-language models. Age estimation is a particularly revealing case: associations acquired during large-scale pretraining align well with the task, and zero-shot VLMs now define the state of the art. However, as we show in this paper, some of these associations play an insidious role and lead to erroneous results. The identity shortcut we expose is one such phenomenon, and we expect it is only the first of many failure modes to surface within the internal mechanisms of VLMs. Our activation-steering method addresses the shortcut, offering a path toward correcting VLM failure modes more broadly.

From a broader perspective, our results point to an increasing dilemma in AI research. As learning models grow in size and capability, we are departing from task-specific learning, where the inference process is dictated by the employed architecture. VLMs and other large models become so flexible, that they are capable of solving the same task through different means. This is an unmatched advantage technically; at the same time, it can undermine our ability to measure genuine progress if models solve tasks in unexpected or even incorrect ways.

Societal Impact. Our robustness analysis further shows that current VLM-based age estimators fail under low-magnitude perturbations, with error rates unacceptable for security-critical systems. Given that automated age estimation is already being deployed under recent legislation governing minors' access to online services, the resulting gap between benchmark performance and real-world reliability creates a deployment risk that cannot be ignored. We therefore echo the community's concerns about employing age estimators in this setting.

Limitations and Future Work. One limitation of our method, is its reliance on open weight models. Even though the technique is applied during inference time, it still requires access to the activations, limiting its use for the best-performing closed weight VLMs. Additionally, each experiment relies on a surrogate to measure the effect of the identity shortcut, such as using MiVOLO to get an unbiased estimate even though it might have biases itself and testing identity knowledge through model generated answers. Another limitation of our work is that we study the identity shortcut only on one task—age estimation. Even though this setup is important for real-world deployments, we hypothesize that these identity shortcuts generalize to other attribute inference tasks like gender, height and weight. Rigorous validation across a wider array of models, datasets, and demographic distributions remains a critical direction for future research.

Acknowledgments and Disclosure of Funding

This work was supported by the European Research Council (ERC) under the Consolidator Grant "MALFOY" (101043410), by the German Federal Ministry of Research, Technology and Space (BMFTR) under the grant "AIgenCY" (16KIS2012), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2092 CASA – 390781972.

References

Age Check. ACCS 2:2021 Technical Requirements for Data Protection and Privacy. Technical report, Age Check Certification Scheme Ltd, 2021. Approved by the Information Commissioner's Office in accordance with the Commissioner's tasks and powers under Articles 57(1)(n) and 58(3)(f) pursuant to Article 42(5) of the UK General Data Protection Regulation.

Commonwealth of Australia. Online Safety Amendment (Social Media Minimum Age) Bill 2024. Act No. 127 of 2024, 2024. Assented 10 December 2024.

A.-M. Cretu, K. Kireev, A. Abdalla, W. Obinna, R. Meier, S. A. Bargal, E. M. Redmiles, and C. Troncoso. Evaluating Concept Filtering Defenses against Child Sexual Abuse Material Generation by Text-to-Image Models. *Computing Research Repository*, 2025.

X. Cui, A. Aparcedo, Y. K. Jang, and S.-N. Lim. On the Robustness of Large Multimodal Models Against Image Adversarial Attacks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

European Commission. Communication from the Commission: Guidelines on measures to ensure a high level of privacy, safety and security for minors online, pursuant to Article 28(4) of Regulation (EU) 2022/2065. Technical Report C(2025) 6826 final, European Commission, Brussels, 2025.

C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *Advances in Neural Information Processing Systems*, 38, 2026.

C. Gaul, E. Fidalgo, E. Alegre, Rocío Alaiz Rodríguez, and E. Pérez Corral. Underage detection through a multi-task and multiage approach for screening minors in unconstrained imagery. *Pattern Recognition*, 178, 2026. ISSN 0031-3203.

Google DeepMind. Gemma 4. Model Card, 2026.

K. Hanaoka, M. Ngan, J. Yang, G. W. Quinn, A. Hom, and P. Grother. Face Analysis Technology Evaluation: Age Estimation and Verification, institution = National Institute of Standards and Technology, 2024.

R. Hendel, M. Geva, and A. Globerson. In-Context Learning Creates Task Vectors. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics, 2023.

D. Hendrycks and T. G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proc. of the International Conference on Learning Representations*, 2019.

M. Kuprashevich and I. Tolstykh. MiVOLO: Multi-input Transformer for Age and Gender Estimation. In *Analysis of Images, Social Networks and Texts*. Springer Nature Switzerland, 2023.

M. Kuprashevich, G. Alekseenko, and I. Tolstykh. Beyond Specialization: Assessing the Capabilities of MLLMs in Age and Gender Estimation. *Computing Research Repository*, 2024.

A. Lanitis, C. Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4), 2002.

S. S. Latif, S. Shiper, K. M. R. Kiran, M. F. Ishmam, M. A. Hossain, A. R. M. Kamal, and M. H. Ashmafee. Enhancing Vision Language Corruption Robustness using Cross-Distribution & Prompted Denoisers. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2024.

Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi. A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved Baselines with Visual Instruction Tuning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. MMBench: Is Your Multi-modal Model an All-around Player? In *European Conference on Computer Vision*. Springer, 2024b.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *Proc. of International Conference on Computer Vision*, 2015.

S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. AgeDB: the first manually collected, in-the-wild age database. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2017.

A. Nagar, S. Jaiswal, and C. Tan. Zero-Shot Visual Reasoning by Vision-Language Models: Benchmarking and Analysis. In *Proc. of the International Joint Conference on Neural Networks*, 2024.

B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. *ACM Multimedia*, 2009.

H. Pan, H. Han, S. Shan, and X. Chen. Mean-Variance Loss for Deep Age Estimation From a Face. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

V. S. Patel, Z. Nie, T.-N. Le, and T. V. Nguyen. Masked Face Analysis via Multi-Task Deep Learning. *Journal of Imaging*, 7(10), 2021.

J. Pombal, N. M. Guerreiro, R. Rei, and A. F. T. Martins. Zero-shot Benchmarking: A Framework for Flexible and Scalable Automatic Evaluation of Language Models, 2025.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. Blog Post, February 2026.

S. Ren, X. Shen, A. Raj, A. Dai, Y. Xu, Z. Chen, S. Wu, C. Gong, Y. Zhang, and others. Out of the box age estimation through facial imagery: A Comprehensive Benchmark of Vision-Language Models vs. out-of-the-box Traditional Architectures. *Computing Research Repository*, 2026.

N. Rizwan, P. Bhaskar, M. Das, S. S. Majhi, P. Saha, and A. Mukherjee. Exploring the Limits of Zero Shot Vision Language Models for Hate Meme Detection: The Vulnerabilities and their Interpretations. *Proc. of the International AAAI Conference on Web and Social Media*, 2025.

L.-Z. Szu-Tu, T.-L. Wu, C.-J. Chang, H. Syu, and Y.-L. Liu. Exposing VLM Memorization of Famous Landmarks: A 55K Building Age Dataset Revealing Popularity Bias, 2025.

G. Team. Gemma 3 Technical Report. *Computing Research Repository*, 2025.

M. Usama, S. A. Asim, S. B. Ali, S. T. Wasim, and U. B. Mansoor. Analysing the Robustness of Vision-Language-Models to Common Corruptions. *Computing Research Repository*, 2025.

P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *Computing Research Repository*, 2024.

Z. Wang, C. Xie, B. Bartoldson, and B. Kailkhura. Double Visual Defense: Adversarial Pre-training and Instruction Tuning for Improving Vision-Language Model Robustness. *Computing Research Repository*, 2025.

J. Ye, Y. Wu, S. Gao, C. Huang, S. Li, G. Li, X. Fan, Q. Zhang, T. Gui, and X. Huang. RoTBench: A Multi-Level Benchmark for Evaluating the Robustness of Large Language Models in Tool Learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024.

Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, X. Huang, Z. Wang, L. Sheng, L. Bai, et al. LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. *Advances in Neural Information Processing Systems*, 36, 2023.

K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10), 2016.

A Experimental Details

To aid reproducibility and foster further research in this area, we aim to release as many details about our experiments as possible. Any details not found in this section may be inferred from the provided at github.com/mlsec-group/identity-shortcut.

A.1 Generation Settings

To query the VLMs and ensure reproducible, comparable results between them, we always utilized the same prompt: **"Estimate the age of the person in this photograph. Respond with ONLY a single integer representing their age in years. Do not include any other text, explanation, or units"**. Additionally, we set the temperature to 0 and the maximum number of generated tokens to 10. We then parse the estimated age as an integer from the response. We report the precisions of the models and their licenses in Table 5. For activation steering, we use $\alpha = 3$ across all models.

Table 5: Models, their precision, and licenses.

Model	Precision	License
ResNet-18	bfloat16	BSD-3-Clause
MiVOLO	bfloat16	Apache-2.0
LLaVa 1.5 7B	bfloat16	Apache-2.0
QwenVL 2.5 7B	bfloat16	Apache-2.0
Gemma 3 4B	bfloat16	Gemma License
Qwen 3.5 9B	quantized 4 bit	Apache-2.0
Gemma 4 31B	quantized 4 bit	Apache-2.0
Gemini 3 Flash	n/a	Gemini Terms of service

A.2 Identifying Identity Knowledge

We identify when a VLM knows the identity of a person in two different ways, depending on the availability of labels. For AgeDB, labels with the person's name exist, so we simply prompt the model to guess the identity of the depicted person and accept answers with an edit distance of less than 5. This allows for some leeway in the names such as accents on characters being included or omitted in the ground truth and model answer while keeping the number of false positives to a minimum. The prompt used for this task is **"Do you know this person? If yes, provide their name only. If not, say 'Unknown"**.

On CelebA, we do not have the correct labeled identities, so we instead check the answers by cross-checking the answer through a separate prompt in a separate context. For this, we use the following prompt: **"Is this person 'NAME'? Answer with yes or no only."**, where "NAME" is the answer provided by the previous prompt. Manual checks of 100 randomly chosen samples confirm that the person is indeed identified correctly.

A.3 Robustness Settings

In order to evaluate the performance of the VLMs under common noise corruptions, we manipulated the images from the utilized datasets with the corruptions listed in Table 6. We analyzed several severity levels, namely severity levels $\in \{0.25, 0.5, 0.75, 0.99\}$. Depending on these levels, we calculated parameters for the different corruptions. We report the values for the parameters or parameter pairs for each of the four severity levels. As displayed in Figure 2, we average over all these corruption functions and severity levels, as well as over the applied adversarial perturbations.

A.4 Datasets

We acquire AgeDB [Moschoglou et al., 2017]. FG-Net [Lanitis et al., 2002], and CelebA [Liu et al., 2015] from Kaggle. They are widely-used image datasets of celebrities and non-celebrities. AgeDB, in particular, consists of over 16 000 images of 570 subjects labeled with age and gender. FG-Net is composed of 1 002 images of 82 subjects. For most subjects, the dataset contains multiple photos

Table 6: Common noise corruptions and the hyperparameters depending on our four severity levels.

Corruption	Parameter	Values given severity
Gaussian Noise	σ	$\{0.13, 0.22, 0.31, 0.40\}$
Shot Noise	c	$\{17.25, 31.50, 45.75, 59.43\}$
Impulse Noise	c	$\{0.09, 0.15, 0.21, 0.27\}$
Speckle Noise	c	$\{0.26, 0.38, 0.49, 0.60\}$
Defocus Blur	(radius, alias blur)	$\left\{\begin{matrix} 4.75, & 0.20 \\ 6.50, & 0.30 \\ 8.25, & 0.40 \\ 9.93, & 0.50 \end{matrix}\right\}$
Glass Blur	(σ, max δ, iterations)	$\left\{\begin{matrix} 0.90, & 1, & 2 \\ 1.10, & 2, & 2 \\ 1.30, & 3, & 2 \\ 1.49, & 3, & 2 \end{matrix}\right\}$
Motion Blur	(radius, σ)	$\left\{\begin{matrix} 12.50, & 6.00 \\ 15.00, & 9.00 \\ 17.50, & 12.00 \\ 19.90, & 14.88 \end{matrix}\right\}$
Zoom Blur	zoom factor	$\left\{\begin{matrix} 1.00, & 1.01, & 1.03, & 1.04, & 1.06, & 1.07, & 1.09, & 1.10 & \\ 1.00, & 1.02, & 1.04, & 1.06, & 1.08, & 1.10, & 1.12, & 1.14, & 1.16 \\ 1.00, & 1.02, & 1.05, & 1.07, & 1.10, & ,1.12, & 1.15, & 1.17, & 1.20 \\ 1.00, & 1.03, & 1.06, & 1.09, & 1.12, & 1.15, & 1.18, & 1.21, & 1.24 \end{matrix}\right\}$
Gaussian Blur	σ	$\{2.25, 3.50, 4.75, 5.95\}$
Snow	$(c_1, \ldots, c_7)$	$\left\{\begin{matrix} 0.21, & 0.30, & 3.38, & 0.59, & 10.50, & 5.00, & 0.94 \\ 0.33, & 0.30, & 3.75, & 0.68, & 11.00, & 6.00, & 1.08 \\ 0.44, & 0.30, & 4.12, & 0.76, & 11.50, & 7.00, & 1.21 \\ 0.55, & 0.30, & 4.48, & 0.85, & 11.98, & 7.96, & 1.34 \end{matrix}\right\}$
Frost	(c_1, c_2)	$\left\{\begin{matrix} 0.90, & 0.49 \\ 0.80, & 0.57 \\ 0.70, & 0.66 \\ 0.60, & 0.75 \end{matrix}\right\}$
Fog	(c_1, c_2)	$\left\{\begin{matrix} 1.88, & 1.85 \\ 2.25, & 1.70 \\ 2.62, & 1.55, \\ 2.98, & 1.41 \end{matrix}\right\}$
Spatter	$(c_1, \ldots, c_6)$	$\left\{\begin{matrix} 0.66, & 0.33, & 1.75, & 0.68, & 0.82, & 0.00 \\ 0.66, & 0.35, & 2.50, & 0.67, & 1.05, & 0.00 \\ 0.67, & 0.38, & 3.25, & 0.67, & 1.27, & 0.00 \\ 0.67, & 0.40, & 3.97, & 0.66, & 1.49, & 0.00 \end{matrix}\right\}$
Brightness	c	$\{0.20, 0.30, 0.40, 0.50\}$
Contrast	c	$\{0.31, 0.23, 0.14, 0.05\}$
Saturate	(c_1, c_2)	$\left\{\begin{matrix} 5.22, & 0.05 \\ 10.15, & 0.10 \\ 15.07, & 0.15 \\ 19.80, & 0.20 \end{matrix}\right\}$
Elastic	(c_1, c_2, c_3)	$\left\{\begin{matrix} 0.80, & 4.80, & 2.16 \\ 1.60, & 3.20, & 1.76 \\ 2.40, & 1.60, & 1.36 \\ 3.17, & 0.06, & 0.98 \end{matrix}\right\}$
Pixelate	c	$\{0.46, 0.33, 0.19, 0.06\}$
JPEG	quality	$\{20, 16, 12, 7\}$
Adversarial	ϵ	$\{0.1/255, 0.3/255, 1/255, 8/255\}$

taken at different stages of their lives. CelebA contains over 200 000 images of a total of over 10 000 individuals. These images are not annotated with a name, age, or gender.

Table 7: Datasets and their licenses.

Dataset	License
AgeDB	non-commercial research purposes only
FG-Net	MIT
CelebA	non-commercial research purposes only
VideoScrape	Terms Of Use, Privacy Policy

VideoScrape To ensure that we evaluate our hypotheses on images of real human subjects which are not included in the training dataset of the analyzed VLMs, i.e. published after the knowledge cut-off date, we compose a dataset of 200 images acquired from publicly uploaded videos from

autonomous drones. Owners of the HoverAir models from ZeroZero Robotics can upload videos within the Hover app and share them with third parties via URLs. We use a Python script that semi-automatically browses the app for videos, copies, and saves the share link to a file. We then automatically go through all acquired links and check the metadata for the upload date and the actual video content URL.

If the upload date is after the latest known VLM cutoff date, we download the video. For models without a known knowledge cutoff date, we use the public date instead as the upper bound on the possible knowledge cutoff. We then center crop each frame to a 500 pixel wide square and detect faces within the cropped frame via a pretrained MTCNN [Zhang et al., 2016] (MIT License). According to 7 (m) of the Terms Of Use of ZeroZero Robotics, retrieving data of their websites or apps is prohibited only if it was "...not made generally available to visitors to the Sites or to Registrants...". If users actively decide to share their videos and photos captured with a HoverAir drone within the app, the content is publicly shared with all registered users to receive likes and comments. The code to retrieve the images is published within our code base. We are not storing any videos and we will delete the images of human subjects with the publication of this paper.

Subsampling To ensure a roughly equal distribution of ages and genders, we analyze our VideoScrape dataset using MiVOLOv2 and obtain the demographics as presented in Table 8. We then analyze AgeDB and FG-Net and sample images randomly to match the VideoScrape demographics given enough images for a certain age bin are present in the dataset. FG-Net does not officially label their images with the ground-truth gender but the extended FG-Net Mask dataset [Patel et al., 2021] annotates the images accordingly.

Table 8: Demographics for the utilized datasets. We first analyze our VideoScrape dataset and then subsample images from AgeDB and FG-Net to match the same distribution as best as possible.

	Age Bins							
	0-2	3-6	7-12	13-20	21-32	33-43	44-53	54-100
VideoScrape								
Male	0	1	5	1	25	55	38	41
Female	0	3	4	1	15	18	5	3
AgeDB								
Male	0	1	5	1	25	55	38	41
Female	0	3	4	1	15	18	5	3
FG-Net								
Male	0	1	5	1	25	38	19	10
Female	0	3	4	1	15	18	5	3

A.5 Compute Resources

For all experiments except those involving Gemini 3 Flash, we utilize a cluster of NVIDIA GPUs consisting of H100, A100, A40, H200 and RTX PRO 6000 Blackwell GPUs. We did not control which GPU was used for each experiment and therefore report aggregated GPU hours instead in Table 9. For Gemini 3 Flash, we use a total of 109.3 million input tokens and 199.9 thousand output tokens.

Table 9: Compute required for running the experiments.

Experiment	GPU Hours
Impact on Known Identities	4
Robustness across Distributions	1 499
Impact on Unknown Identities	2
Task Vector Distribution	2
Disabling the Identity Shortcut	8

B Ablation Experiment

To show the robustness of our method across different hyperparameter choices, we perform an ablation study of the intervention strength α and the layer until which the task vector is chosen using Gemma 3. Lastly, we also perform a short experiment on height estimation.

The results can be seen in Table 10 and Table 11. For the intervention strength, a range of values can be used, and further tuning to the task could even improve performance. As a sanity check for our method, we have also included negative values, which correspond to performing steering in the opposite direction. As expected, steering in the opposite direction worsens performance. For the cutoff layer, the results are indeed sensitive to the chosen layer, but there is a range of layers that all result in activation steering remaining beneficial. While an earlier cutoff provides marginal gains on AgeDB, the middle layer serves as a compromise across datasets to avoid hyperparameter search.

Table 10: Sensitivity analysis of the intervention strength α. Values represent the change in mean absolute error relative to our chosen baseline setting of $\alpha = 3$.

α	AgeDB	FG-Net
-3	$+3.27$	$+5.70$
-2	$+2.47$	$+4.80$
-1	$+1.21$	$+3.52$
0	$+0.47$	$+2.12$
$+1$	-0.53	$+0.99$
$+2$	-0.48	-0.17
$+3$	± 0.00	± 0.00
$+4$	$+1.11$	$+0.01$

Table 11: Ablation over the layer range used to define the task vector. Values denote the shift in the cutoff layer relative to our baseline (the exact middle of the model, denoted as 0). Results represent the change in mean absolute error relative to this baseline.

Cutoff Layer	AgeDB	FG-Net
-3	-0.21	$+1.82$
-2	-0.42	$+1.38$
-1	-0.57	$+0.61$
± 0	± 0.00	± 0.00
$+1$	$+4.44$	$+1.01$
$+2$	$+7.44$	$+4.40$

To check the robustness to different prompts, we try multiple variants. The first prompt is the one used for all other experiments. The next two prompts are variants of that and the last two prompts are engineered to reduced reliance on the identity shortcut:

1. "Estimate the age of the person in this photograph. Respond with ONLY a single integer representing their age in years. Do not include any other text, explanation, or units."
2. "Look at the photograph and provide your best guess of the person's age as a single whole number. Output nothing else."
3. "Based on this image, estimate how old the person is and respond with only an integer age in years, with no additional words or symbols."
4. "Examine the person in the photograph and estimate their age by considering visible facial features, skin texture, hair characteristics, body proportions, and any other age-related cues apparent in the image. Provide your best age estimate as a single integer in years only. Output nothing except the number: no words, no punctuation, no units, and no explanation."
5. "Examine the person in the photograph and estimate their age based only on visible, image-based cues such as facial features, skin texture, hair appearance, posture, and overall physical presentation. Do not use or consider identity, name, known background, public profile,

or any external knowledge about who the person is. Your output must be exactly one whole number representing the estimated age in years, with no extra text, punctuation, units, explanation, or qualifiers."

The results with and without activation steering can be seen in Table 12. Without steering, the models still return a mean absolute error of 9.06 ± 0.75 across the five different prompts for AgeDB. With steering using the above prompts, the average mean error is 8.22 ± 0.25, leading to a reduction in error of $9.31\,\%$ on average. Similarly, for FG-Net, the average mean error without steering is 8.93 ± 0.95 years and with steering 6.49 ± 0.47 years, reducing the error by $27.3\,\%$ on average. As this data shows, the prompt variations do not remove the bias; moreover, our method improves MAE for the engineered prompts, resulting in similar values as the normal prompts.

Table 12: Mean absolute error of Gemma 3 on AgeDB and FG-Net with different prompts.

Dataset	Variant	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5
AgeDB	Default	8.47	9.41	8.41	10.39	8.63
	Steered	8.01	8.18	8.13	8.71	8.07
FG-Net	Default	8.20	9.20	8.27	10.67	8.31
	Steered	5.95	6.56	6.15	7.31	6.47

We also perform a short experiment on whether a partial reasoning trace analysis transfers to a different task in the form of height prediction. We prompt Qwen 3.5 9B to estimate the height of a person based on an image of this person. We use FG-Net for those purposes again as it does not contain celebrities and Qwen 3.6 35B to judge whether identity was considered and rejected by the VLM. In 74.5% of the reasoning traces, identification is mentioned as a possible strategy, and in 56.4% of cases it is not rejected later on, which suggests that the identity shortcut extends beyond age.

C Additional Results

We provide additional detailed results across all models and corruption modes in Figures 6-13.

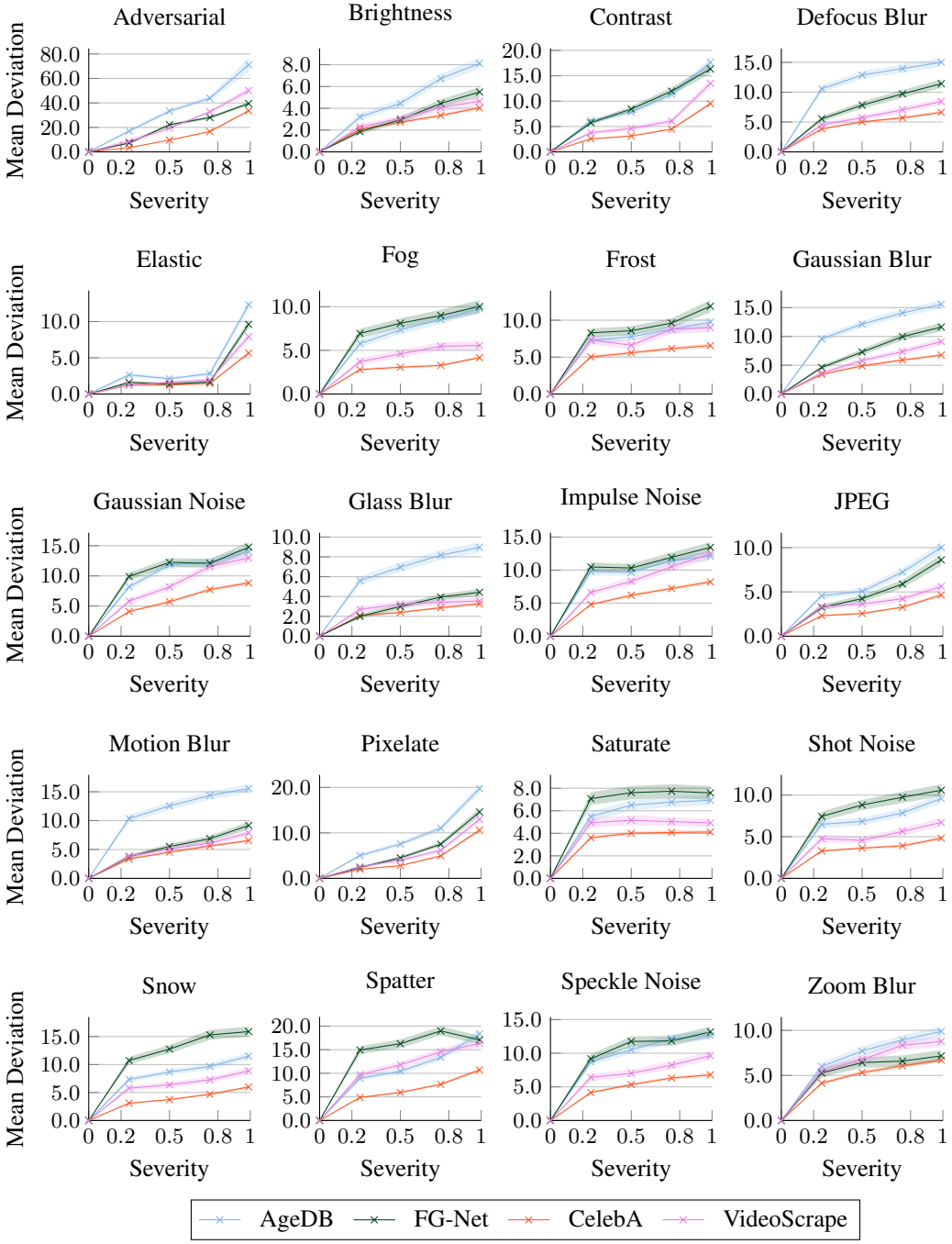


Figure 6: Detailed results for the deviations achieved by each corruption for Gemma 3.

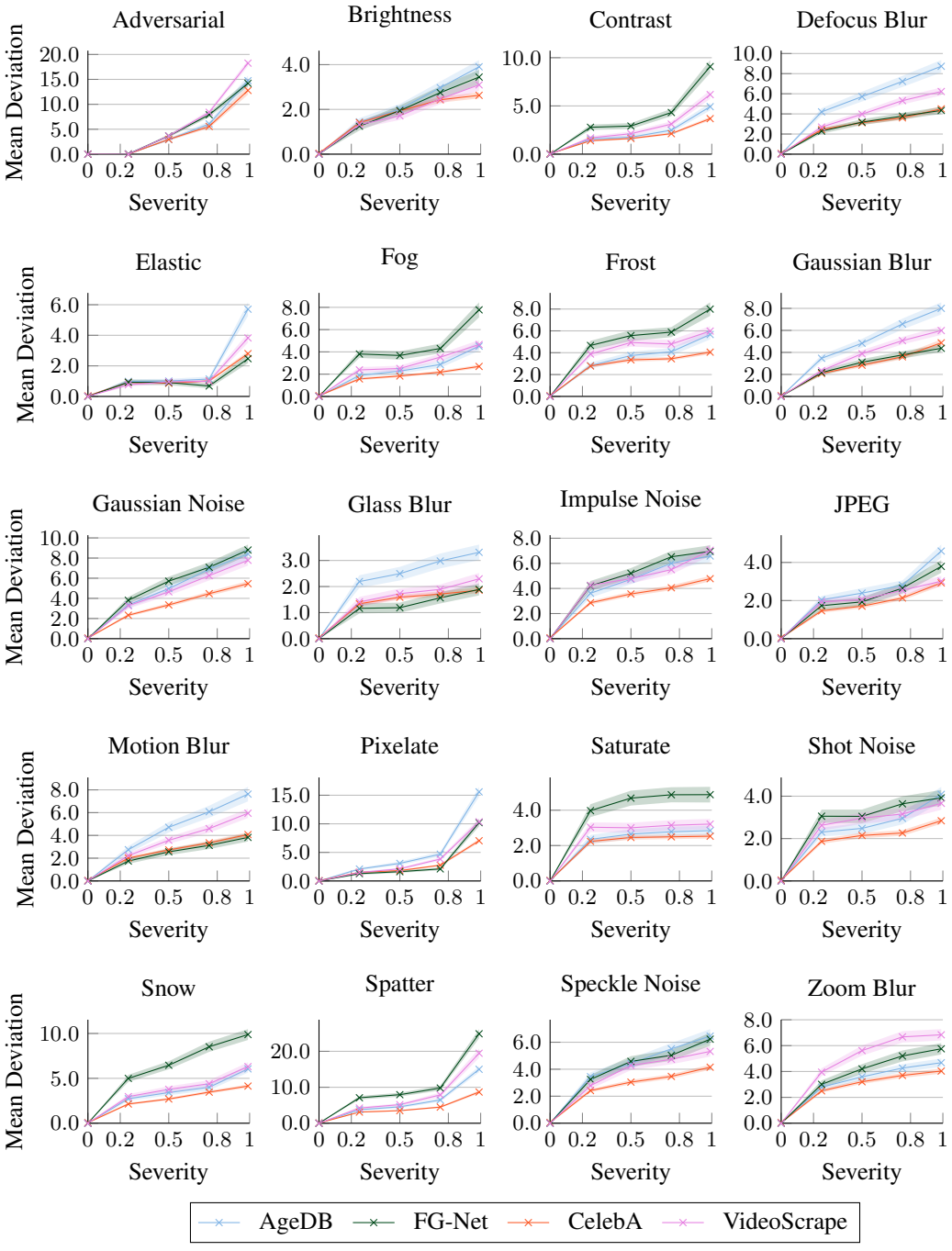


Figure 7: Detailed results for the deviations achieved by each corruption for Gemma 4.

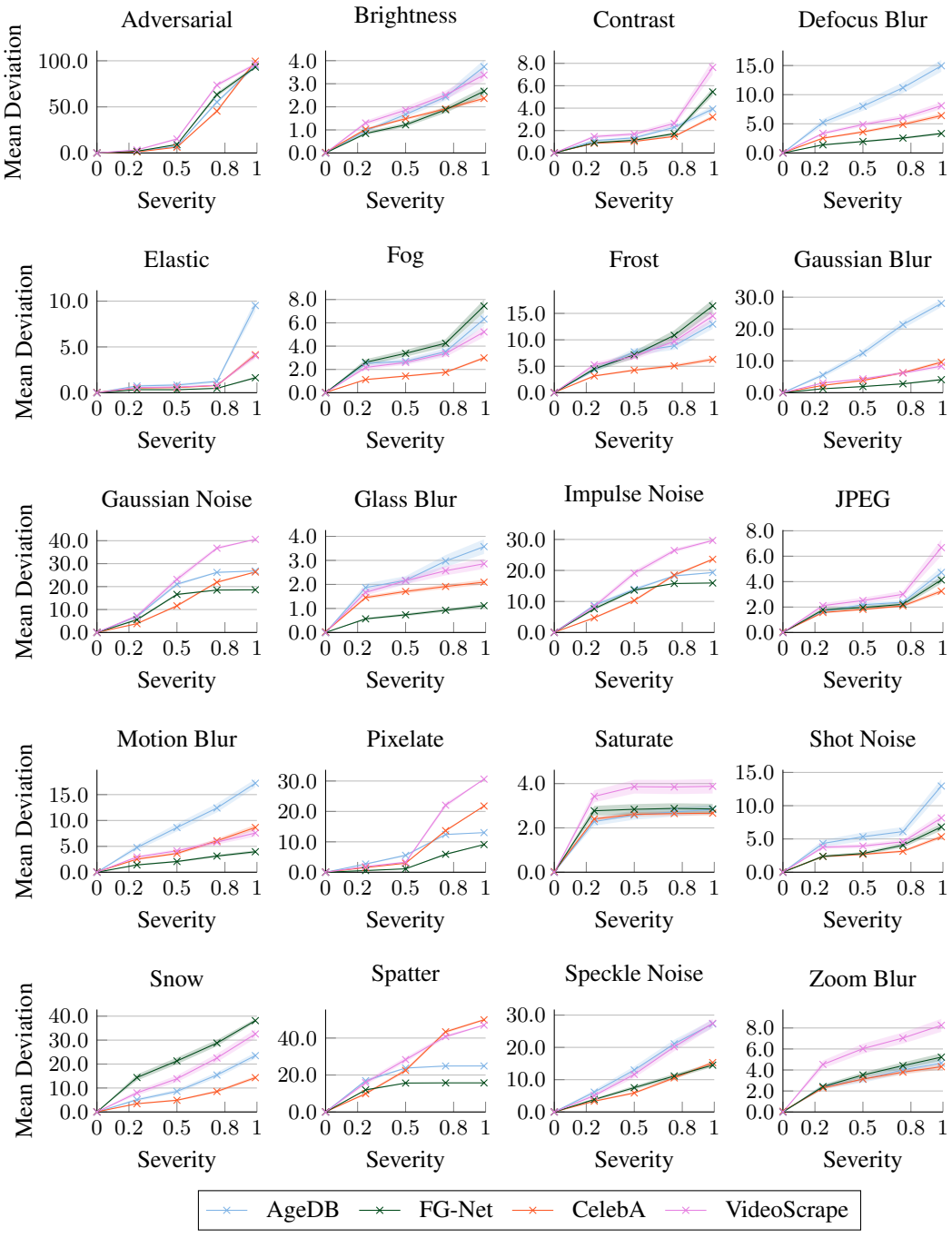


Figure 8: Detailed results for the deviations achieved by each corruption for MiVOLO.

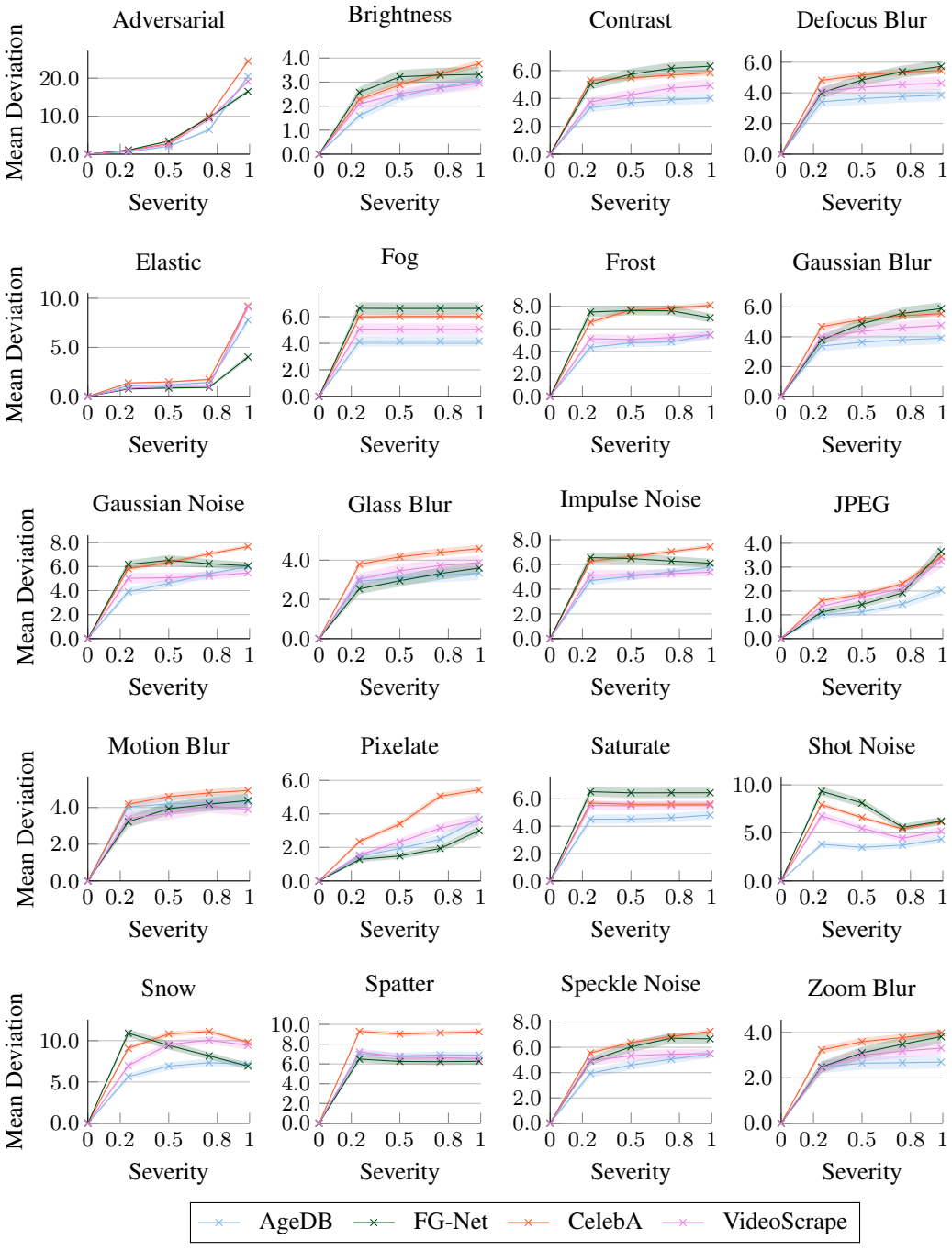


Figure 9: Detailed results for the deviations achieved by each corruption for the CNN.

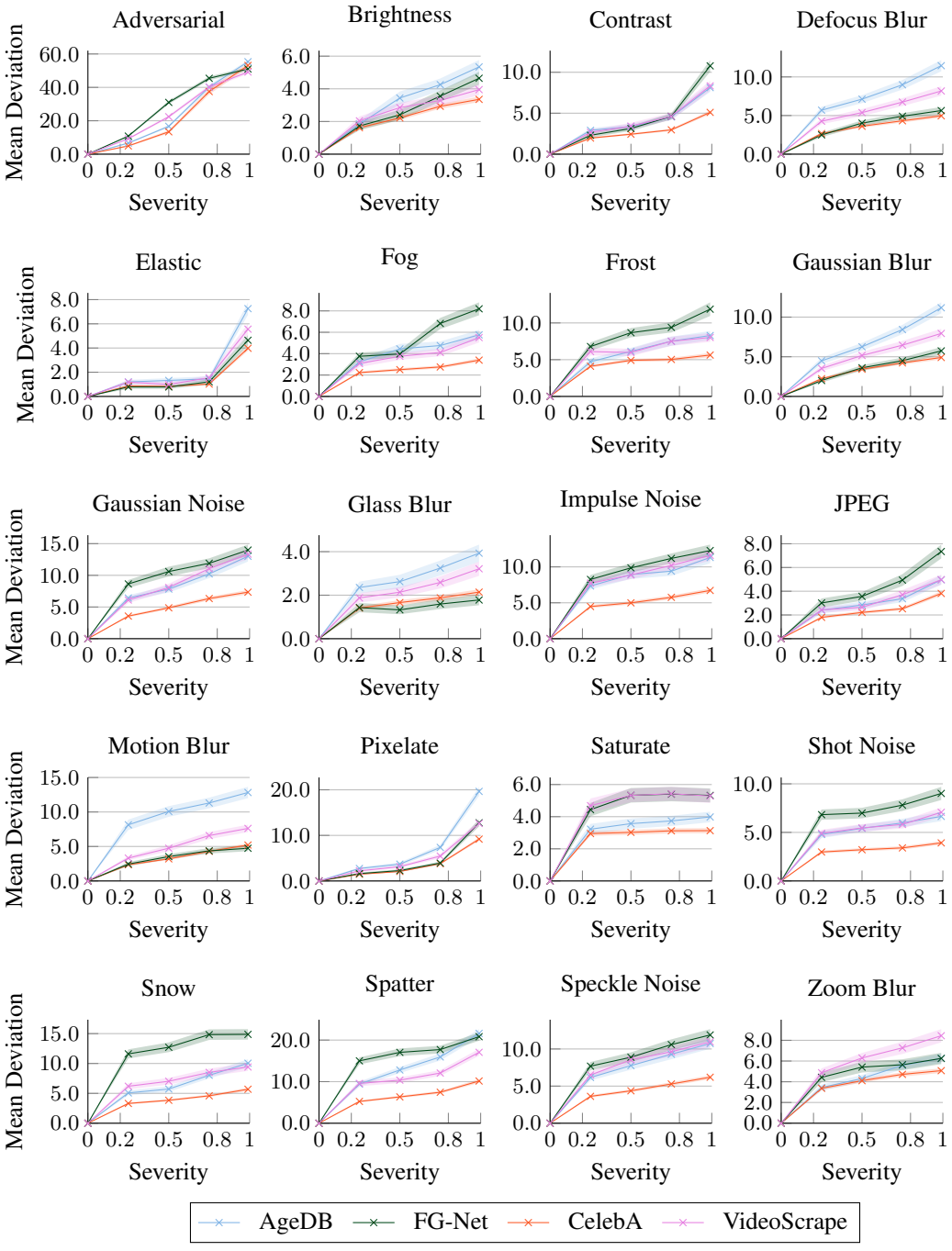


Figure 10: Detailed results for the deviations achieved by each corruption for QwenVL 2.5.

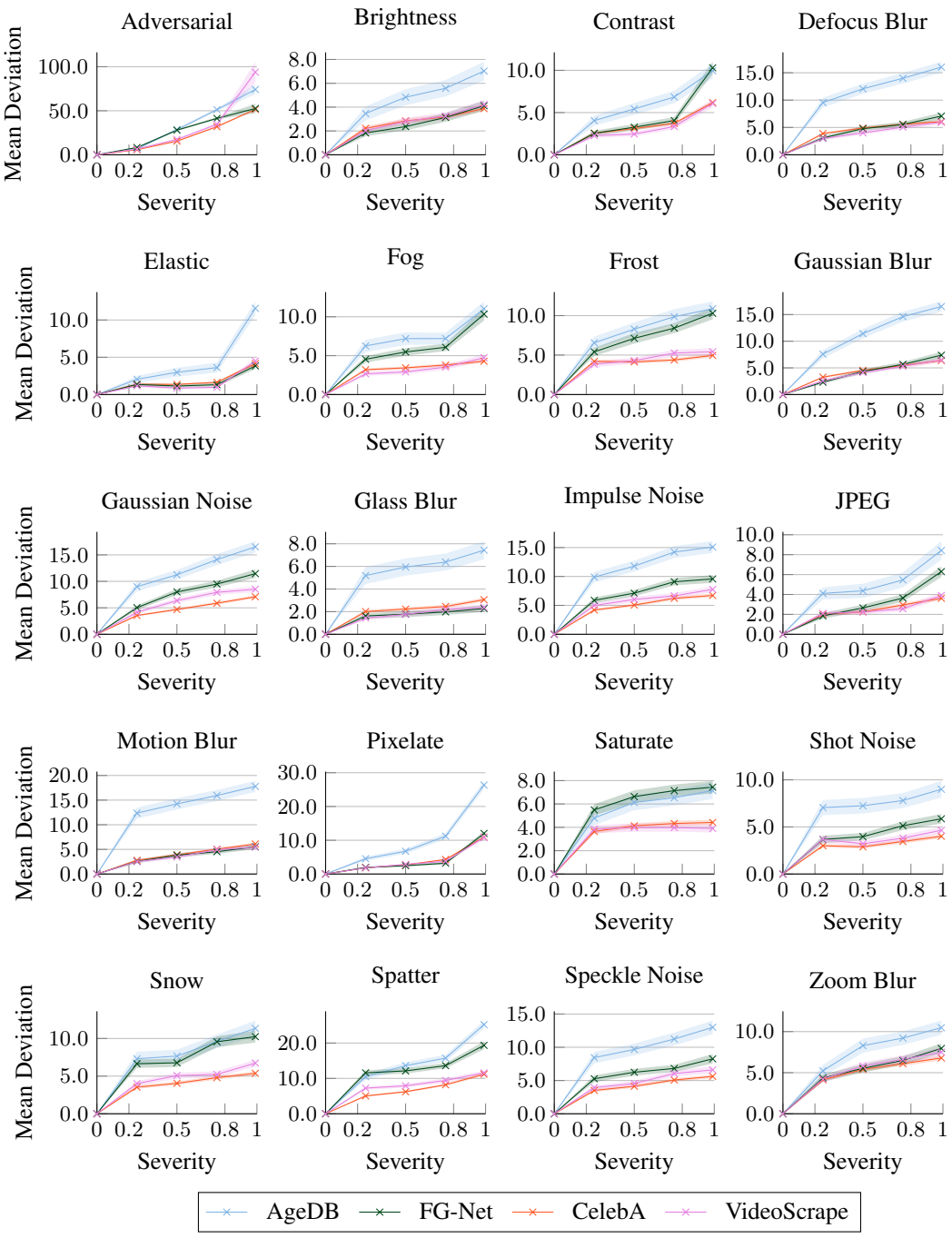


Figure 11: Detailed results for the deviations achieved by each corruption for Qwen 3.5.

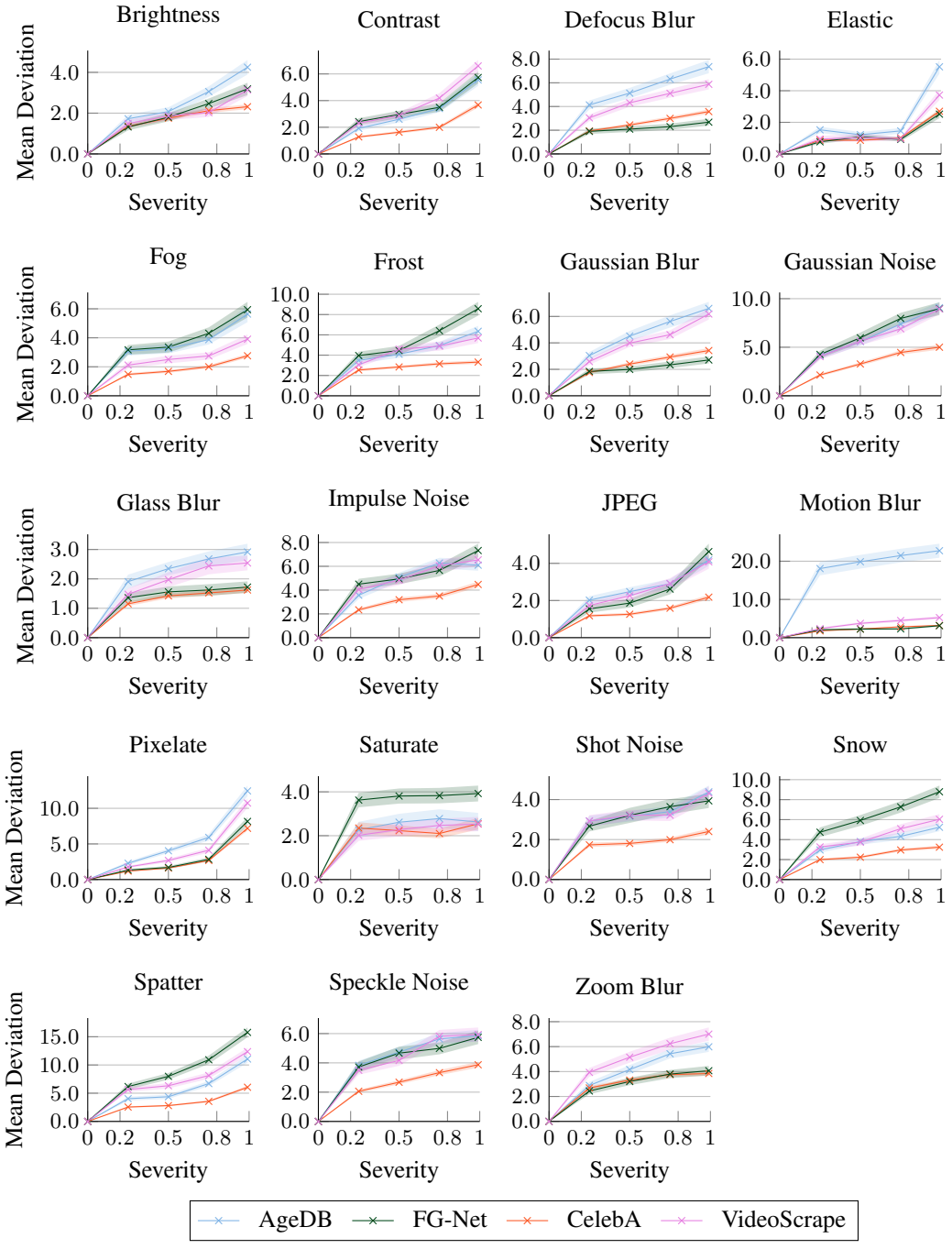


Figure 12: Detailed results for the deviations achieved by each corruption for Gemini 3 Flash.

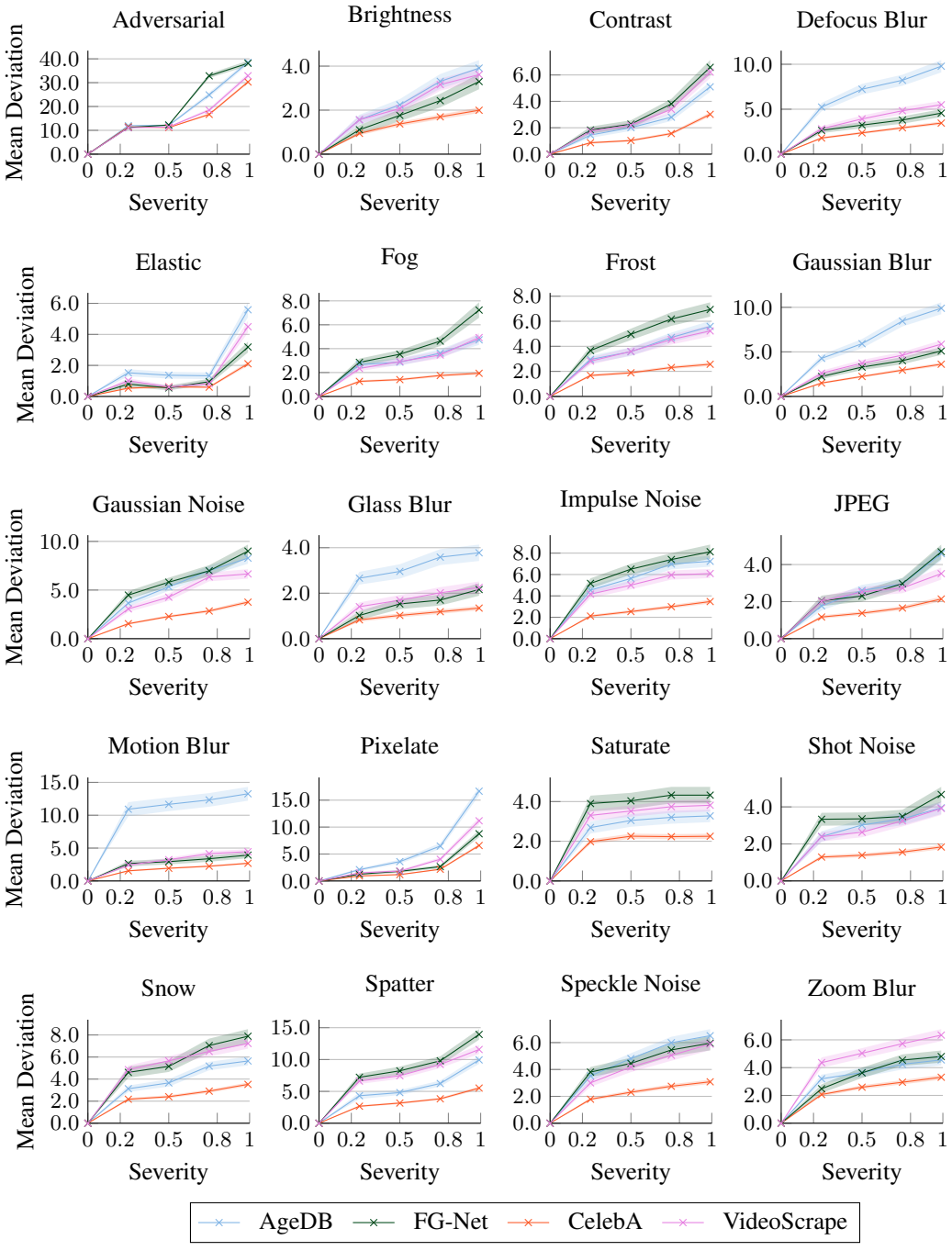


Figure 13: Detailed results for the deviations achieved by each corruption for LLaVa 1.5.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We demonstrate the impact of adversarial and regular noise corruptions on datasets of celebrities and non-celebrities in Section 3.2 and Section 3.4. Our results reveal a discrepancy between the robustness of most evaluated VLMs to noisy inputs of celebrities. We additionally observe that the models tend to *estimate* the age of known identities based on their current age at the knowledge cut off date instead of relying on extracted features, a phenomenon that we introduce as *identity shortcut*. We furthermore suggest a new mitigation strategy in Section 4 improving robustness for age estimation at inference time.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our work are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: Our work does not include formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details regarding our method, the utilized models and image datasets, metrics, and the parameters are provided in the main text, and in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We publish our code alongside instructions to obtain the evaluated datasets. Furthermore, all parameters, attack methods, and metrics to reproduce our results are provided throughout the main text. We make it available at github.com/mlsec-group/identity-shortcut.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the parameters for our experiments and our attack algorithms within the main text of the paper, in the appendix and in the provided code repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the 1-sigma standard error of the mean for all results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information about the computational resources of our experiments in Appendix A.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully judged beneficence by analyzing probable harms and benefits. We limit risks by minimizing both the data we collect and the results we publish. For VideoScrape, we intentionally use at most 10 frames per video and do not save additional metadata such as uploader pseudonyms. We considered alternatives such as manually collecting images guaranteed not to be known by any model. Unfortunately, such alternatives would limit results to a small slice of persons, risking biased representation. We purposely chose not to publish this dataset and delete it upon publication, as privacy harms would outweigh research benefits. Instead, we publish only aggregate statistics on identity shortcut usage. This balances validating our hypotheses in a clean setting against user privacy risks. We gathered only publicly posted videos intended for sharing, observing all applicable terms of use (Section A.4).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper's purpose is to unveil a false sense of robustness of VLMs performing age detection skewed to images of celebrities. We also discovered that VLMs *identity shortcuts* to predict a memorized current age of a celebrity based on the time the model was trained instead of analyzing the picture independent of it. If VLMs are utilized widely for age estimation analyzing images of humans, they are not robust towards adversarial perturbations or noisy corruptions, as discussed in Section 3.2 and Section 3.4, ultimately raising awareness about the security and privacy of VLMs. We, additionally, introduce a countermeasure for our findings that increases robustness towards these corruptions in Section 4. Furthermore, we discuss the broader societal impact of our findings in Section 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We do not upload any of the images in our VideoScrape dataset and are deleting them with the publication of this paper. All other image datasets utilized in this paper are already publicly available and available either under the MIT license or for non-commercial research purposes. We are not publishing any models ourselves.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We fully respect the licensing terms or all publicly available datasets and models, including the name of the license and citing them correctly, see Section 2.2 and Table 7, Table 5. We additionally provide the Terms of Service and Privacy Policy of ZeroZero Robotics in Table 7. We utilized their Hover app to manually scrape images of human subjects to collect a custom dataset, VideoScrape.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We publish the code repository accompanied by the documentation (see github.com/mlsec-group/identity-shortcut).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used for text editing only.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.