RATIONALE CONTENT SHAPES WHAT A DETECTOR LEARNS: CONTROLLED SUPERVISION FOR PERSON-CENTRIC AI-IMAGE FORENSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Detectors trained to generate rationales alongside authenticity decisions may learn different cues depending on the evidence they are taught to explain. This study focuses on consistency between gaze and attended targets in person-centric AI-generated images and analyzes how incorporating such relational gaze information into rationales changes detection performance and output behavior. We conduct experiments with Custom Gaze, 23,415 identity-preserving real/fake pairs that differ only in a generated eye band, and find that gaze-centered fine-tuning outperforms FakeVLM origin across three person-centric evaluation sets. The detector also maintains high balanced accuracy on eye edits from three held-out inpainters, and the largest point-estimate gain under person-count stratification appears in the stratum with two or more detected faces. It achieves the strongest aggregate metrics among the representative detectors compared, while paired-data training produces the same directional effect on the vision-only Effort detector. Finally, with the data, five-block structure, optimization, and checkpoint rule fixed, gaze-centered rationales outperform anatomy-centered and context-centered alternatives on COCOAI Person by 4.2–4.6 and 2.1–2.4 balanced-accuracy points, respectively. These results show that the evidence emphasized in a rationale is a substantive training factor that changes detector behavior.

1 INTRODUCTION

Vision-language fake-image detectors are trained to explain their evidence alongside authenticity decisions. FakeVLM (Wen et al., 2025), for example, generates an authenticity label together with a free-form description of artifacts observed in the image. Yet generating a rationale alone does not determine which evidence the model should attend to.

Person-centric images provide a controlled domain in which to examine this question. Eyes in AI-generated images can appear locally realistic yet conflict with eye direction, head pose, another person, or an attended target. We call supervision that organizes generated eye-region evidence jointly with the face and scene **relational gaze supervision**. It complements detectors based on frequency anomalies (Frank et al., 2020), upsampling traces (Durall et al., 2020; Tan et al., 2024), and generator-specific residuals (Ojha et al., 2023; Wang et al., 2020; Yan et al., 2025a).

Testing the role of rationale content requires both a localized intervention and a content-controlled comparison. Custom Gaze supplies the first: each real/fake pair shares source identity, scene, illumination, compression history, and global composition, while the fake member contains a regenerated eye band. Five-block captions provide the second: they keep the decision–scene–method–evidence–conclusion order fixed, identify whether the scene contains multiple people, and vary the semantic content among gaze-, anatomy-, and context-centered alternatives.

We first compare gaze-centered fine-tuning with FakeVLM origin. The gaze-centered detector performs better across three person-centric evaluation sets. The effect extends to three held-out inpainters, while person-count stratification shows the largest point-estimate gain in the 4,442-image stratum with two or more detected faces. These results show that the learned detection behavior extends beyond a single editing model and appears in the same direction in multi-person scenes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Against representative existing detectors, our model achieves the strongest aggregate metrics across the three evaluation sets, and paired-data training produces the same directional performance change on the vision-only Effort detector. Finally, in a controlled experiment that changes only the rationale content, gaze-centered rationales improve balanced accuracy on COCOAI Person by 4.2–4.6 points over anatomy-centered rationales and by 2.1–2.4 points over context-centered rationales. Replacing decision-only targets with the five-block gaze format yields a smaller observed difference of 1.15–1.23 points.

The paper makes three contributions. First, it introduces Custom Gaze, 23,415 identity-preserving real/fake pairs with controlled eye-band edits and five-block rationales. Second, it evaluates the intervention progressively through within-family comparisons, held-out inpainters, person-count strata, existing-detector comparisons, and cross-architecture transfer. Third, it analyzes how the evidence emphasized during rationale training affects detector behavior through fixed-structure replacements of gaze-, anatomy-, and context-centered content and an output-level audit.

2 Related Work

Generalizable synthetic-image detection. Early detectors exposed generator artifacts through spectral mismatches (Frank et al., 2020; Durall et al., 2020) or features learned from a fixed family of synthetic images (Wang et al., 2020). Later work targets robustness across generators and post-processing: UnivFD uses transferable visual representations (Ojha et al., 2023), NPR focuses on upsampling-related residuals (Tan et al., 2024), and AIDE, SAFE, and Effort study complementary routes to cross-domain generalization (Yan et al., 2025a; Li et al., 2025; Yan et al., 2025b). These methods establish strong classification baselines, but their supervision does not expose which semantic evidence a detector is encouraged to verbalize. Our study is therefore complementary: it holds the detector family and output structure fixed and intervenes on the content of its training rationales.

Multimodal detection and explanation supervision. Vision-language detectors extend authenticity prediction with natural-language evidence. FakeVLM pairs decisions with artifact explanations (Wen et al., 2025), SIDA combines detection, localization, and explanation (Huang et al., 2025), and LOKI broadens evaluation of large multimodal models on synthetic-data detection (Ye et al., 2025). In these systems, however, the rationale wording, training examples, and task formulation are typically introduced together. Comparisons against binary classifiers therefore cannot isolate whether a gain comes from producing a structured explanation or from the particular evidence named in that explanation. We address this distinction with a fixed five-block format and content-matched alternatives centered on gaze, facial anatomy, and scene context.

Ocular evidence in visual forensics. Person-centric forensics has long used local physiological or optical irregularities. Eye-blink dynamics reveal temporal inconsistencies in generated face videos (Li et al., 2018); irregular pupil shapes (Guo et al., 2022) and inconsistent corneal highlights (Hu et al., 2021) provide image-level cues. These approaches treat the eye primarily as a local forensic region. Our formulation instead asks whether generated eye evidence is compatible with the surrounding face and scene. It does not replace local artifact cues or assume that gaze alone is sufficient.

Gaze and social relations. Gaze research supplies the relational variables used in our supervision. LAEO-Net and OI Mutual Gaze model whether people look at one another (Marín-Jiménez et al., 2019; Doosti et al., 2021), while gaze-following methods connect a head and eye direction to an attended target (Recasens et al., 2015; Chong et al., 2020). Gaze360 and ETH-XGaze further document how gaze estimation interacts with head pose and unconstrained viewing conditions (Kellnhofer et al., 2019; Zhang et al., 2020). Those works predict attention or gaze rather than image authenticity. We use their scene-level primitives as a vocabulary for supervising forensic explanations, then test the resulting detector under pose and person-count strata.

Controlled paired data. Dataset documentation clarifies provenance, composition, and intended use (Gebru et al., 2021), but observational datasets alone leave many visual factors entangled. Custom Gaze pairs each authentic image with a localized eye-region intervention while preserving the source identity and scene, and keeps the pair atomic across train, validation, and test splits. The paired construction supports within-identity evaluation; a separate rationale-content replacement

holds the data and optimization fixed. Together, these controls distinguish the effect of the edited evidence from the effect of how that evidence is described during training.

3 Custom Gaze Construction and Training Pipeline

The pipeline constructs paired eye-region edits, converts each example into five-block supervision, and fine-tunes on the result mixed with FakeClue. Figure 1 shows the overall flow.

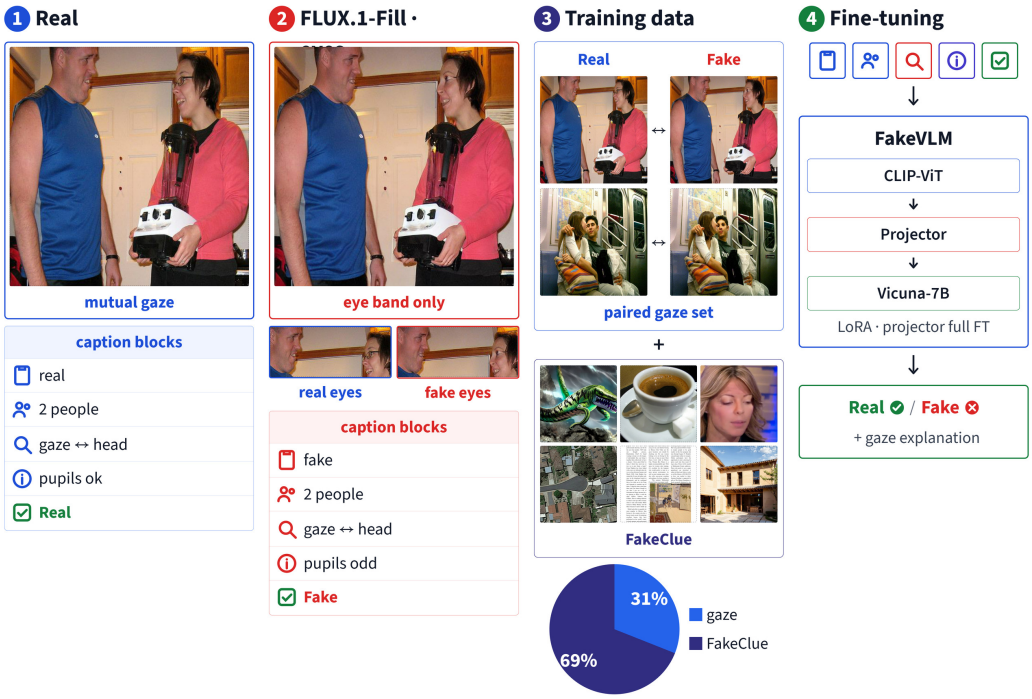

Figure 1: Overview of the proposed data and supervision pipeline.

3.1 Identity-Preserving Eye-Region Pairs

We select source interactions from the mutual-gaze-positive portion of OI Mutual Gaze (Doosti et al., 2021). Multi-box unpacking yields 23,415 effective participant pairs from 8,972 retained source images. Each record provides bounding boxes for participants A and B. In each pair we regenerate one participant's eye band. For a face box with upper-left coordinate (x_{min}, y_{min}), width w, and height h, the edited region is

$$M_{eye} = \{(i, j) : y_{min} + 0.25h \le i \le y_{min} + 0.55h,\ x_{min} + 0.05w \le j \le x_{max} - 0.05w\}.$$

The alpha mask is softened with a Gaussian blur of radius 3. FLUX.1-Fill-dev (Black Forest Labs, 2024) regenerates the selected eye region with the prompt *eyes looking to the side, avoiding eye contact, natural eye movement, same person, photorealistic*. Images are resized to a maximum side of 1,024 pixels and aligned to a 16-pixel grid before generation.

Each real/fake pair is assigned to a split as one atomic unit. The real and edited versions therefore share base identity, scene, lighting, and global composition, while every validation and test identity remains outside the training set. Custom Gaze contains 46,830 images with an 8:1:1 train/validation/test split.

Mask coverage was audited against both eye landmarks from InsightFace over 102,138 detected faces. The band contains both eye centers for 97.5% of faces, with coverage of at least 95.8% in every face-size quartile and at least 96.5% across frontality strata. This measurement is the primary validation of the mask geometry.

3.2 FIVE-BLOCK CAPTION SUPERVISION

Each Custom Gaze response follows a five-block sequence. The first block states the authenticity decision, the second identifies the scene configuration, the third names the inspection strategy, the fourth describes the observed evidence, and the fifth connects the evidence back to the decision. The second block is not merely a scene summary; it establishes when relational gaze reasoning applies. Because mutual gaze is meaningful when multiple people are identified, the model first checks whether the scene is multi-person and then compares head–eye alignment with plausible attended targets. As Figure 2 shows, the design fixes the progression from scene condition to inspection method and evidence while varying the wording within each block.

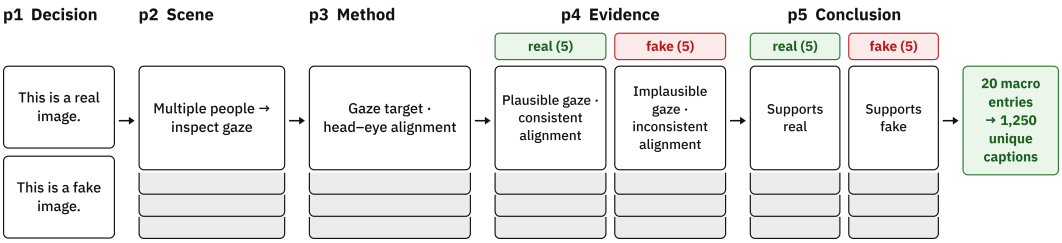

Figure 2: Five-block caption schema. Each response follows Decision, Scene, Method, Evidence, and Conclusion; the Scene block first identifies when relational gaze reasoning applies. Wording varies within each block.

The decision sentence is label-deterministic. The remaining four positions draw from five paraphrases per position, with label-specific evidence and conclusion pools. The resulting caption space contains 1,250 unique captions in total. All captions preserve the same reasoning order and vary only in lexical realization. A representative fake response is:

> This is a fake image. More than one person appears in the scene, so I examined their gaze relationship. I compared pupil placement and head-eye alignment with plausible interaction targets. The visible eye directions appear inconsistent with the depicted exchange. Overall, the eye-region evidence supports a fake image.

3.3 FAKEVLM AND EFFORT FINE-TUNING CONFIGURATIONS

We use separate fine-tuning paths for the rationale-generating and vision-only detectors. The FakeVLM-based model learns from Custom Gaze pairs together with five-block rationales, whereas Effort (Yan et al., 2025b) learns authenticity labels from the same Custom Gaze training split. Table 1 compares the settings needed to interpret and reproduce these two paths.

Table 1: Fine-tuning configurations for FakeVLM and Effort. Full settings are reported in Appendices B.1 and B.2.

Configuration	FakeVLM-based model	Effort
Base model	CLIP-ViT-L/14-336 + Vicuna-v1.5-7B	CLIP-ViT-L/14
Training data	Custom Gaze 31% + FakeClue 69%; 151,173 images total	Custom Gaze training split; 37,464 images
Supervision	Authenticity decision + rationale	Authenticity label
Checkpoint	Maximum development-mixture BA	Maximum validation BA

For FakeVLM, we adapt the language model with LoRA (Hu et al., 2022) and fully train the vision-language projector so that the model can learn both the paired data and the rationale structure. Effort does not generate rationales, so the second path tests whether the effect of Custom Gaze transfers to another detector at the label level. Both checkpoints are selected by validation balanced accuracy rather than training loss. For FakeVLM, the BA-optimal and token-loss-optimal checkpoints differ by 1,200 steps, so generation loss alone is not used to choose the detector checkpoint.

4 EXPERIMENTS

We first compare the overall effect of gaze-centered supervision with FakeVLM origin and examine transfer across generators and inpainters. We then place the result among existing detectors and another architecture, and finally isolate the contribution of rationale content through a fixed-structure ablation. **Common setup.** Custom Gaze evaluates paired partial eye edits. Of the two evaluation views constructed from MS COCOAI (Roy et al., 2026), COCOAI Person evaluates person-centric full synthesis from five generators, while COCOAI Interaction provides a small multi-person interaction view. The subset of COCOAI Person with two or more detected faces is used as the larger multi-person stratum. Table 2 summarizes the benchmark sizes.

Table 2: Evaluation benchmarks.

Dataset	Total	Real	Fake
Custom Gaze test	4,684	2,342	2,342
COCOAI Person	15,720	2,620	13,100
COCOAI Interaction	198	33	165
COCOAI Person, 2+ detected faces	4,442	842	3,600

The primary metrics are balanced accuracy (BA) (Brodersen et al., 2010), macro-F1, and the Matthews correlation coefficient (MCC). For model-to-model comparisons, we resample paired image identities within each class and use 10,000 bootstrap resamples, paired label-swap permutation tests for BA, and exact McNemar tests on correctness disagreements. Results are reported separately for sampling seeds 42, 47, and 53. Appendix B.3 gives the parsing, decoding, checkpoint, and threshold rules.

4.1 HOW DOES GAZE-CENTERED SUPERVISION CHANGE DETECTION?

We first compare FakeVLM origin with the gaze-supervised checkpoint within the same model family. This comparison changes both the paired Custom Gaze data and its rationale content; Section 4.3 separately analyzes the contribution of rationale semantics.

The central result in Table 3 is that the improvement is not confined to the localized eye edits closest to the training distribution: the same direction appears on full person synthesis and multi-person interaction scenes. The effect is largest on Custom Gaze at roughly 13 BA points. On COCOAI Person it is smaller at roughly 1.2 points, but remains consistent across seeds, and every seed-wise confidence interval excludes zero on both datasets. The macro-F1 and MCC gains on COCOAI Person are larger than the BA gain, indicating that the decision profile becomes less class-asymmetric rather than merely producing more correct predictions overall. Thus, gaze-centered supervision has two observed effects: it strongly separates the targeted edits and contributes to more balanced decisions under distribution shift.

Table 3: FakeVLM origin versus gaze-centered fine-tuning. Values are mean ± standard deviation over seeds 42, 47, and 53.

Dataset	Model	BA	Macro-F1	MCC
COCOAI Interaction	FakeVLM origin	68.3 ± 0.4	64.9 ± 0.8	31.6 ± 1.2
	Ours	$\mathbf{71.1 \pm 0.3}$	$\mathbf{71.9 \pm 0.2}$	$\mathbf{43.8 \pm 0.3}$
COCOAI Person	FakeVLM origin	83.1 ± 0.1	78.4 ± 0.2	58.4 ± 0.3
	Ours	$\mathbf{84.3 \pm 0.1}$	$\mathbf{83.8 \pm 0.1}$	$\mathbf{67.6 \pm 0.2}$
Custom Gaze	FakeVLM origin	86.5 ± 0.2	86.3 ± 0.3	75.7 ± 0.4
	Ours	$\mathbf{99.9 \pm 0.0}$	$\mathbf{99.9 \pm 0.0}$	$\mathbf{99.8 \pm 0.1}$

Figure 3 relates this change to the number of people visible in a scene. The gain persists at 0.93–1.24 pp in the zero-detected-face stratum and 0.93–1.04 pp in the one-detected-face stratum, while every seed has its largest point estimate, 1.72–1.89 pp, in the two-or-more-face stratum. This pattern

is consistent with the supervision becoming more useful when a relation between people and their attended targets can be formed, rather than only the local appearance of an eye being available. The stable improvement in the one-face stratum also suggests that the learned evidence is not restricted to multi-person relations and includes person-level cues such as head–eye alignment.

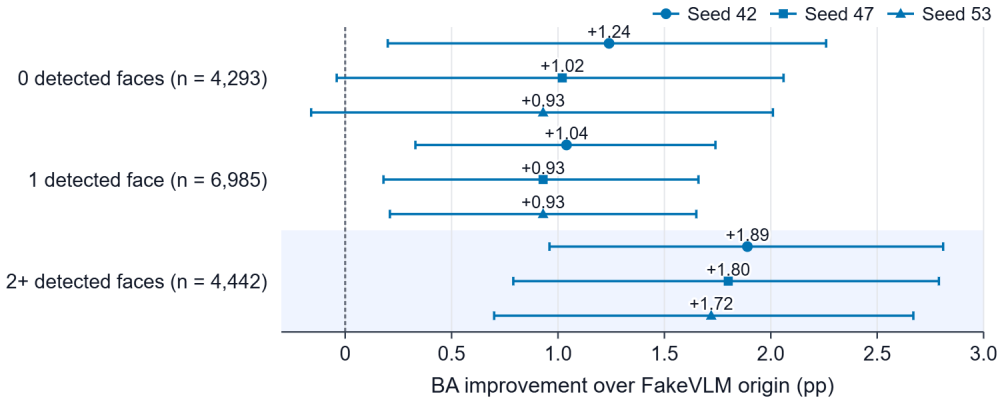

Figure 3: Seed-wise paired BA improvements over FakeVLM origin on COCOAI Person, stratified by detected-face count. Points and whiskers show the paired BA difference and its 95% bootstrap confidence interval.

Figure 4 tests whether the effect is tied to a particular generator or editor. In panel (a), all five full-image generators and every seed move in the same direction, with the three-seed mean fake-recall gain ranging from 4.4 pp on SD2.1 to 9.6 pp on DALL-E 3. This shows that the average improvement is not produced by only a subset of generator families. In panel (b), the gaze-supervised model retains 98.1–99.1% BA when the eye regions are regenerated with SDXL Inpaint (Podell et al., 2024), SD3 Inpaint (Esser et al., 2024), and RePaint (Lugmayr et al., 2022), none of which is used for training. Although the editors change the visual form of the intervention, the large gap from FakeVLM origin recurs. The observed effect therefore accords more closely with cues recurring across generated eye regions than with a single FLUX-specific editing signature.

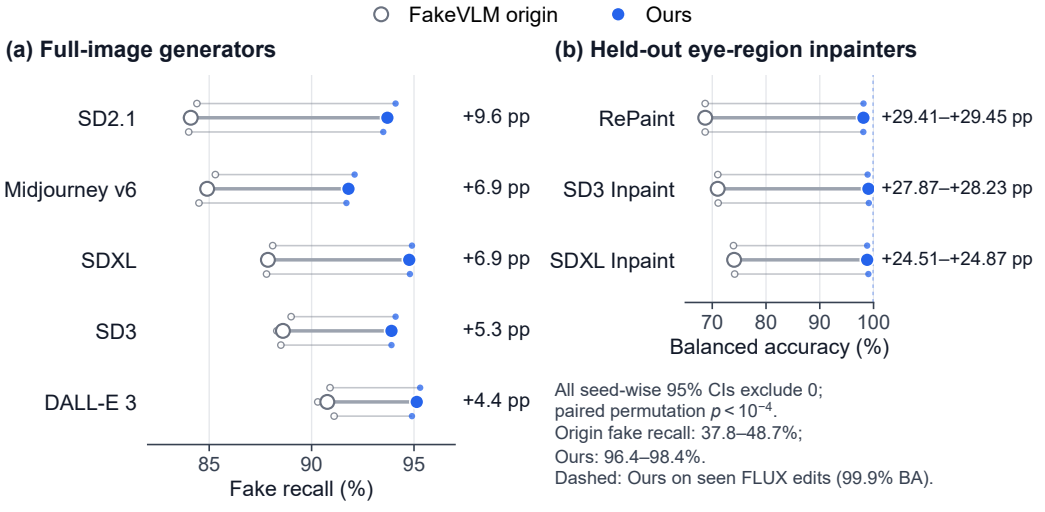

Figure 4: Generalization across full-image generators and held-out eye-region inpainters. (a) Fake recall on COCOAI Person; right-hand labels give the three-seed mean gain. (b) Balanced accuracy on paired held-out inpainter pools, each containing 2,324 real and 2,324 fake images; right-hand labels give the seed-wise gain range. Small points denote individual seeds and large points their mean.

4.2 How Does the Model Compare with Existing Detectors?

We compare gaze-centered fine-tuning with representative low-level, classifier, and multimodal detectors on the same evaluation views. FakeVLM origin is the most direct controlled baseline because it retains the same backbone and inference family; Appendix B.3 gives the complete protocol.

Figure 5 shows that detectors centered on low-level traces remain near chance or strongly class-asymmetric on the person-centric views, whereas models using semantic information are comparatively stable. Our model has the strongest aggregate metrics across the three evaluation sets, with the largest separation on Custom Gaze, where the intervention directly matches the training signal. The advantage also persists on COCOAI Person and Interaction, indicating that the effect is not limited to memorizing the paired edit. Its smaller cross-distribution margin positions relational gaze as a complement to general-purpose detectors rather than a universal replacement. On LOKI (Ye et al., 2025), the gaze-supervised checkpoint scores 84.5%, similar to the published FakeVLM value of 84.3%, so this specialization does not reduce the prior evaluation capability measured there.

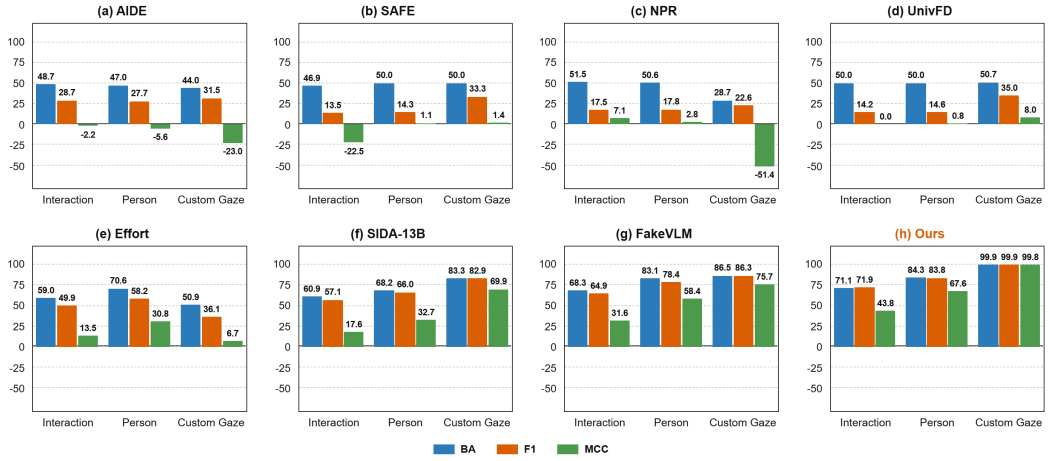

Figure 5: External baseline comparison. Each panel shows one detector's BA, macro-F1, and MCC on Interaction, Person, and Custom Gaze. FakeVLM origin and Ours are the same three-seed means as in Table 3; the remaining methods use fixed released checkpoints.

To test whether the gain depends on the rationale-generating VLM, the second path in Section 3.3 fine-tunes the vision-only Effort detector on the same Custom Gaze training split. Effort emits a scalar rather than an explanation, separating the role of the paired data from rationale supervision. In Table 4, Custom Gaze BA rises from 50.9 to 99.9, while the two COCOAI views also move in the same direction. Paired examples that are identical outside the local eye region therefore provide an effective training signal without a rationale-generation process; the role of rationale content is isolated next by holding the structure and data fixed.

Table 4: Cross-architecture transfer on Effort.

Model	COCOAI Interaction BA / F1 / MCC	COCOAI Person BA / F1 / MCC	Custom Gaze BA / F1 / MCC
Effort origin	59.0 / 49.9 / 13.5	70.6 / 58.2 / 30.8	50.9 / 36.1 / 6.7
Effort + Custom Gaze	**62.4 / 58.1 / 20.0**	**75.2 / 66.2 / 39.5**	**99.9 / 99.9 / 99.9**

4.3 What Does Rationale Content Change?

To distinguish the contributions of the training data and rationale content that change together in the origin comparison, we conduct controlled ablations. The decision-only comparison removes the four explanatory blocks. The content comparison keeps the five-block structure fixed and rewrites only Blocks 2–5 around relational gaze, generic facial anatomy, or general scene context.

The key result in Table 5 is that the evidence a rationale is required to organize matters more than the mere presence of a longer explanation. On COCOAI Person, adding the five-block gaze structure yields about 1.2 BA points over decision-only supervision. With the structure fixed, gaze-centered content yields 4.2–4.6 points over anatomy-centered content and 2.1–2.4 points over context-centered content; every seed-wise confidence interval excludes zero for both contrasts. The same ordering holds on the other two evaluation sets, making a simple explanation based on learning more output tokens less consistent with the results. This ablation directly supports the study's central hypothesis that the evidence named in an explanation helps shape the detector's decision criterion.

Table 5: Balanced accuracy (BA, %) for the rationale ablation. Values are three-seed mean ± standard deviation.

Supervision variant	Custom Gaze	COCOAI Person	COCOAI Interaction
Decision only	99.8 ± 0.0	83.2 ± 0.1	68.8 ± 1.1
Five-block anatomy	98.3 ± 0.3	80.0 ± 0.2	65.5 ± 0.4
Five-block context	98.1 ± 0.2	82.1 ± 0.1	66.5 ± 0.8
Five-block gaze	$\mathbf{99.9 \pm 0.0}$	$\mathbf{84.3 \pm 0.1}$	$\mathbf{71.1 \pm 0.3}$

The performance differences are accompanied by changes in how the model expresses its evidence. The output audit defines a gaze card as a response that contains both a multi-person phrase (*multiple people*, *more than one person*, or *several people*) and a gaze-related term (*gaze*, *eye contact*, *eye direction*, or *pupil*). In Table 6, FakeVLM origin never emits this card, whereas the gaze-supervised model links scene composition and gaze evidence on all three evaluation sets. The relational structure supplied during training therefore remains visible in the generated reasoning, not only in the final label. Texture-related vocabulary also remains frequent at 52–82%, suggesting that relational evidence augments rather than replaces the original artifact descriptions.

Table 6: Gaze-card usage rate. Ranges summarize seeds 42, 47, and 53.

Model	Custom Gaze	COCOAI Person	COCOAI Interaction
FakeVLM origin	0.0%	0.0%	0.0%
Ours	**69.9–70.8%**	**15.3–18.4%**	**20.7–25.3%**

Table 7 asks whether this change in explanatory vocabulary is also associated with a different decision profile. Within the two-or-more-face stratum of COCOAI Person, gaze-card-active outputs treat the two classes comparatively evenly and reach approximately 0.83 BA. Gaze-card-inactive outputs are more strongly biased toward fake predictions and reach approximately 0.77 BA. The higher real recall of card-active outputs recurs in the one-face stratum and on COCOAI Interaction. Thus, the gaze card marks not only the surface reproduction of learned vocabulary but also a decision mode associated with relational scene reasoning. Appendix E reports the corresponding error concentration analysis.

Table 7: Conditional gaze-card behavior on COCOAI Person. Ranges summarize three seeds.

Stratum and output state	Approx. n per seed	Fake recall	Real recall	Within-stratum BA
2+ faces, card active	1,100	0.72–0.76	**0.89–0.90**	**approx. 0.83**
2+ faces, card inactive	3,300	**0.97–0.98**	0.55–0.58	approx. 0.77
1 face, card active	Variable	0.70–0.75	**0.87–0.88**	approx. 0.79
1 face, card inactive	Variable	**approx. 0.97**	0.62–0.66	approx. 0.80

5 Discussion and Limitations

The fixed-structure ablation provides the paper's most direct result. Detector performance changes when the semantic content of the rationale changes, even with the data, rationale form, optimization, and checkpoint rule fixed. On COCOAI Person, the observed effect size of the content contrasts is larger than that of the contrast between five-block and decision-only targets. This is not a separate significance test between contrast magnitudes, and the conclusion is scoped to the evaluated rationale-supervision setting.

Performance on SDXL, SD3, and RePaint weakens an explanation based only on a FLUX-specific signature. At the same time, all held-out fake images contain generated eye regions, so the current design does not fully separate gaze semantics from general generated-eye texture. The multi-person result rests primarily on the 4,442-image COCOAI Person stratum with two or more detected faces; overlapping stratum-wise intervals do not establish a difference by face count itself. COCOAI Interaction is a supplementary view with 33 real images, and the zero-face stratum denotes face-detector non-detection rather than the absence of people.

The scope is aligned with person-centric images and the evaluated rationale structure. In the output audit, 52.6–56.6% of fake→real errors invoke the multi-person gaze card. This result suggests a role for scene conditioning, although a gating mechanism is not evaluated in this study.

6 Conclusion

We analyzed whether the evidence emphasized during rationale training affects the performance and output behavior of person-centric AI-image detectors. Gaze-centered fine-tuning outperforms FakeVLM origin across three evaluation sets, and the effect extends to held-out inpainters and multi-person scenes. Existing-detector comparisons and transfer to Effort show the same directional benefit from gaze-centered paired-data training. With the rationale structure fixed, gaze-centered content outperforms anatomy- and context-centered alternatives, while the output audit confirms supervision-dependent changes in vocabulary. Under the evaluated conditions, rationale content therefore acts not merely as an output format but as a training-design factor that adjusts detector performance and output behavior.

AI Use Statement

In this work, the authors used ChatGPT to translate into English parts of a manuscript they had written in Korean. Generative AI tools were also used to improve the grammar, clarity, and readability of the manuscript and to render and visually refine selected figures according to layouts designed by the authors. The layouts and scientific content of those figures, including all data and numerical results, were determined and verified by the authors. The authors take full responsibility for the final content of this paper.

Reproducibility Statement

The main paper specifies the paired data construction, caption design, optimization, checkpoint selection, evaluation metrics, and statistical tests. The appendix reports dataset sources and licenses, exact split counts and rules, training recipes for FakeVLM and Effort, per-baseline inference settings, label parsing, paired resampling and significance tests, and caption-ablation construction. We also report the generation prompt, mask geometry, hardware, and random seeds needed to reproduce the principal comparisons.

Ethics Statement

This work uses publicly available person-centric images and generates localized eye-region edits for research on synthetic-image detection. Source licensing, derivative-use constraints, and provenance are retained in the dataset documentation, and the resulting resource is intended for non-commercial research. Because the images depict people, model outputs should not be used to infer identity,

intent, or mental state, or as the sole basis for consequential authenticity decisions. The evaluation concerns aggregate detection behavior and does not establish demographic fairness.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023.

Black Forest Labs. Flux.1 fill [dev]. Technical report, Black Forest Labs, 2024.

K. H. Brodersen et al. The balanced accuracy and its posterior distribution. In *International Conference on Pattern Recognition*, 2010.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

B. Doosti et al. Boosting image-based mutual gaze detection using pseudo 3d gaze. In *AAAI Conference on Artificial Intelligence*, 2021. arXiv:2010.07811.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, 2020.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, Hal Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu. Eyes tell all: Irregular pupil shapes reveal GAN-generated faces. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

E. J. Hu et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. arXiv:2106.09685.

Shu Hu, Yuezun Li, and Siwei Lyu. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2500–2504, 2021.

Z. Huang et al. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2412.04292.

P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2405–2414, 2025. doi: 10.1145/3690624.3709392.

Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

M. J. Marín-Jiménez et al. Laeo-net: Revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

U. Ojha et al. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. arXiv:2302.10174.

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. In *International Conference on Learning Representations*, 2024.

Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, 2015.

R. Rombach et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. arXiv:2112.10752.

R. Roy et al. A comprehensive dataset for human vs. ai generated image detection. *arXiv preprint arXiv:2601.00553*, 2026.

C. Tan et al. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. arXiv:2312.10461.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Siwei Wen et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Advances in Neural Information Processing Systems*, 2025.

Shilin Yan et al. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, 2025a. arXiv:2406.19435.

Z. Yan et al. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *International Conference on Machine Learning*, 2025b. Oral, arXiv:2411.15633.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

J. Ye et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. In *International Conference on Learning Representations*, 2025. Spotlight, arXiv:2410.09732.

X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Proceedings of the European Conference on Computer Vision*, 2020.

A CUSTOM GAZE AND COCOAI DATA DETAILS

A.1 CUSTOM GAZE DATASHEET

We report the dataset composition, provenance, license, and intended use following the datasheet practice of Gebru et al. (2021).

Table A.1: Custom Gaze summary.

Item	Value
Total samples	46,830
Pair structure	23,415 real/fake pairs
Real-image source	OI Mutual Gaze (Doosti et al., 2021)
Fake construction	FLUX.1-Fill-dev eye-band regeneration
Splits	Train 37,464 / validation 4,682 / test 4,684
Split rule	Atomic grouping by base pair; seed 42
Caption space	1,250 five-block captions
Image license	CC-BY-NC-4.0 derivative terms; upstream provenance pointers retained
Caption-pool license	CC0

Table A.2: Custom Gaze split counts.

Split	Pairs	Real	Fake	Total
Train	18,732	18,732	18,732	37,464
Validation	2,341	2,341	2,341	4,682
Test	2,342	2,342	2,342	4,684
Total	23,415	23,415	23,415	46,830

OI Mutual Gaze contains binary annotations for mutual eye contact. We retain the positive records and unpack multiple face pairs annotated on the same source image into distinct pair candidates. The fake-A variant edits participant A's eye band while retaining the partner and the rest of the image.

The fixed eye band was audited over 102,138 detected faces using the five-point landmarks returned by InsightFace buffalo_l. Both eye centers fall within the band for 97.5% of faces. Coverage by increasing face-size quartile is 99.8%, 98.1%, 96.4%, and 95.8%, and coverage across high, medium, and low frontality groups is 98.4%, 96.8%, and 96.5%. The normalized eye-center height has a median of 0.378 and a 5th–95th percentile range of 0.300–0.461.

The dataset retains per-image upstream license pointers. The derivative license restricts commercial use and prohibits distribution that misrepresents the actual attention or interaction of a depicted person. The intended uses are research on image authenticity, localized edits, and person-centric robustness. Because the source distribution inherits the geographic and demographic composition of Open Images, population coverage should be audited for any deployment context.

A.2 COCOAI PERSON AND COCOAI INTERACTION CONSTRUCTION

Both evaluation views are constructed from MS COCOAI (Roy et al., 2026). They begin with COCO captions containing at least one clearly human noun from {man, men, woman, women, people, person, boy, girl}. COCOAI Person pairs 2,620 real images with five generated counterparts per source. The generators are DALL-E 3 (Betker et al., 2023), Midjourney v6, SD2.1 (Rombach et al., 2022), SD3 (Esser et al., 2024), and SDXL (Podell et al., 2024), which produces 2,620 real and 13,100 fake images.

COCOAI Interaction applies the same human-noun filter, removes caption overlaps with earlier extraction rounds, partitions samples by generator label, and balances the partitions to form a 198-image evaluation view. The canonical paired view contains 33 real and 165 fake images.

For person-count analysis, InsightFace buffalo_l is run with a 640 x 640 detection canvas and the default confidence threshold. On COCOAI Person, zero detected faces cover 650 real and 3,643 fake images, one detected face covers 1,128 real and 5,857 fake images, and two or more cover 842 real and 3,600 fake images. The zero-detection group is interpreted as a detector-defined stratum; small or occluded human faces may still be present.

Table A.3: Face-count stratification on COCOAI Person.

Detected faces	Total	Real	Fake	Seed-wise BA difference	Permutation p
0	4,293	650	3,643	+0.93–+1.24	0.019–0.096
1	6,985	1,128	5,857	+0.93–+1.04	0.006–0.015
2+	4,442	842	3,600	**+1.72–+1.89**	$\leq$ **0.0005**

B Training Setup, Baseline Configurations, and Inference Protocol

B.1 FakeVLM Mixed Fine-Tuning Recipe

Table B.1: FakeVLM mixed fine-tuning configuration.

Item	Value
Base model	`lingcco/fakeVLM`
Vision encoder	CLIP-ViT-L/14-336
Language model	Vicuna-v1.5-7B
Trainable parameters	Approximately 45M of 7B
LoRA	rank 16, alpha 32, dropout 0.05
Optimizer	AdamW, beta=(0.9, 0.999), weight decay 0
Peak learning rate	2e-5
Schedule	Cosine, warmup ratio 0.03
Precision	BF16 + TF32
Distributed training	DeepSpeed ZeRO-2, 8$\times$ RTX A6000 48GB
Per-device batch / accumulation	1 / 4
Effective batch size	32
Maximum sequence length	1,024
Epochs / maximum steps	2 / 7,558
Evaluation interval	50 steps
Development mixture	Custom Gaze 470 + FakeClue 500
Selected checkpoint	Step 1,650, BA 0.9990
Minimum-loss checkpoint	Step 2,850, loss 0.2252

The content-controlled caption variants reuse this recipe. Their input images, labels, five-block order, sentence count, optimization settings, and BA-based selection are identical, and only the semantic vocabulary of Blocks 2–5 differs. Anatomy captions emphasize bilateral facial structure and the integrity of local features. Context captions emphasize object layout, lighting, and global scene coherence. Both variants exclude gaze, pupil, eye-direction, mutual-attention, and equivalent relational terms.

B.2 Effort Fine-Tuning Recipe

Effort uses CLIP-ViT-L/14 with the top 1,023 singular directions of each attention matrix frozen, and trains the residual singular component plus a linear 1,024-to-1 head on Custom Gaze. The DeepfakeBench-style recipe (Yan et al., 2023) uses BCE-with-logits, AdamW (Loshchilov & Hutter, 2019), a learning rate of 2e-4, batch size 32, 256-pixel loading, 224-pixel crops, horizontal flips during training, and center crops at evaluation. The selected model is the best validation-BA checkpoint from two epochs on one RTX 4080 SUPER.

B.3 Per-Baseline Inference and Statistical Protocol

Table B.2: Baseline inference settings.

Baseline	Backbone or checkpoint	Decision rule
AIDE	ResNet-50, GenImage checkpoint	Score ≥ 0.5
SAFE	Public ImageNet-pretrained detector	Score ≥ 0.5
NPR	ResNet-50, resolution-preserving + flip TTA	Score ≥ 0.5
UnivFD	CLIP-ViT-L/14 + linear head	Score ≥ 0.5
Effort	CLIP-ViT-L/14 SVD-residual detector	Score ≥ 0.5
SIDA-13B	Public multimodal checkpoint	REAL vs. FULLSYNTHETIC/TAMPERED
FakeVLM origin	`lingcco/fakeVLM`	First decision keyword
Ours	FakeVLM + checkpoint 1,650	First decision keyword

In the paired FakeVLM comparisons, an output counts as parseable when the generated response contains a valid first decision keyword. BA, F1, and MCC are computed on each model's parseable subset, while McNemar treats an unparsable response as an error. Pairwise uncertainty is estimated with 10,000 class-stratified bootstrap samples, applying the same sampled image indices to both models. The BA permutation test swaps the two models' predictions independently per image with probability 0.5. Seeds are analyzed separately rather than pooled.

For Figure 5, stratified bootstrap intervals were computed where per-sample predictions are publicly available. SIDA's intervals were reconstructed by class-wise binomial resampling of its confusion matrix. Representative BA intervals are as follows: on COCOAI Interaction, SAFE 47.0 [42.4, 50.0], NPR 51.5 [50.3, 53.0], and SIDA 60.9 [51.8, 70.0]; on COCOAI Person, AIDE 47.0 [46.1, 47.9]; and on Custom Gaze, SIDA 83.4 [82.4, 84.3].

C Caption-Level Diagnostics

C.1 Output Diversity and Training Dynamics

Table C.1: Caption-diversity diagnostics.

Interval	Unique-output rate	Modal-template rate	Mean generation length
Step 1,650	0.9308	0.5258	89.15 tokens
Final step 7,550	0.9638	0.5238	89.30 tokens
Mean over the last 50 evaluations	0.956	0.524	89.30 tokens
Mean over all post-warmup evaluations	0.943	0.524	89.28 tokens

The stable length and unique-output ratio show faithful use of the five-block format with lexical variation. Checkpoint selection also changes which evidence cards dominate. Later checkpoints increasingly favor generic real-side descriptions, whereas the BA-best checkpoint preserves a more balanced decision profile.

C.2 Supervision-Dependent Rationale Vocabulary

Table C.2: Keyword frequencies in Custom Gaze outputs.

Keyword family	FakeVLM origin	Five-block gaze	Decision only
Gaze / eye direction / pupil	0.0%	**70.6%**	0.0%
Multi-person / multiple people	0.0%	**70.6%**	0.0%
Texture / appearance	71.2%	**87.2%**	0.3%
Gaze–multi-person co-occurrence	0.0%	**70.6%**	0.0%

The decision-only model usually emits a bare label sentence. The five-block model uses the scene and evidence vocabulary supplied during training, and its joint multi-person/gaze frequency matches the design of the caption pool. General texture vocabulary remains active in the five-block model as well.

D Caption Ablation: Output Statistics and Two-View Accuracy

D.1 Paired-View Evaluation

The canonical ablation uses the same real-plus-fake paired view as Tables 3 and 5. This avoids the class-asymmetric interpretation produced by a fake-only view. At each variant's BA-best checkpoint, the five-block model exceeds decision-only by 1.15–1.23 points on COCOAI Person and by 1.08–3.23 points on COCOAI Interaction, while both exceed 99.8% on Custom Gaze.

Table D.1 reports fake-only recall as a class-specific diagnostic alongside the paired-view evaluation used for the primary comparison.

Table D.1: Fake-only diagnostics and Custom Gaze BA.

Variant	COCOAI Interaction fake recall	COCOAI Person fake recall	Custom Gaze BA
Five-block gaze, step 1,650	0.844	0.881	0.9998
Decision only, step 1,500	0.815	0.907	0.9996

The decision-only model's high fake recall is paired with low real recall, which the paired-view BA, F1, and MCC metrics capture. This is why the ordering changes between the two evaluation views.

D.2 Content-Controlled Captions

The anatomy and context variants match the gaze variant in block order, sentence count, decision prefix, approximate length, training images, labels, mixture ratio, LoRA configuration, optimizer, evaluation cadence, and checkpoint rule. The variants differ only in the semantic content of Blocks 2–5.

The ranking is stable across the three datasets. Gaze-centered captions perform best, context captions form the stronger generic semantic alternative, and anatomy captions trail on the cross-

Table D.2: Content-controlled BA differences, gaze minus alternative. Custom Gaze reports the pooled range across the two controls; COCOAI values are separated by alternative.

Dataset	Gaze − anatomy	Gaze − context	Confidence-interval summary
Custom Gaze	Positive; pooled range +1.4–+1.9	Positive; pooled range +1.4–+1.9	All seed-wise 95% CIs > 0
COCOAI Person	+4.2–+4.6	+2.1–+2.4	All seed-wise 95% CIs > 0
COCOAI Interaction	+5.3–+6.0	+3.7–+5.7	Anatomy: all CIs > 0; context: one CI > 0

distribution person-centric evaluations. This result identifies the content of the rationale as the primary explanatory factor in the caption ablation.

E Discussion, Limitations, and Additional Analyses

E.1 Controlled-Evidence Map

The evidence can be organized by what varies while the rest of the pipeline stays fixed.

Table E.1: Controlled-evidence map.

Analysis	Factor varied	Factors fixed	Key observation
Unseen inpainters	Eye-generator family	Source image, target face, eye band, frozen detector	BA reaches 98.1–99.1 on SDXL, SD3, and RePaint
Real-eye control	Original vs. regenerated eye content	Crop, resize, JPEG, mask blur, compositing	Real recall remains 99.6–99.9% under real-eye processing
Face-count stratification	Available inter-person relations	Dataset, generator, model, inference	Largest BA gain in the 2+ face stratum
Caption-content replacement	Gaze vs. anatomy vs. context	Five-block structure, data, optimizer, selection	Gaze content ranks first on all three evaluations
Explanation audit	Gaze card active vs. inactive	Face-count stratum	Active branch has higher real recall and more balanced behavior
Cross-architecture transfer	FakeVLM vs. Effort	Custom Gaze training pairs	Same-direction gains in both detector families

These analyses address different parts of the interpretation. The content-controlled comparison isolates rationale semantics, the held-out inpainters test transfer beyond FLUX, and the authentic-eye control tests matched processing. Together, they provide complementary evidence about rationale content, editor transfer, and processing effects.

E.2 Scope and Robustness

The evaluation is constructed around person-centric imagery: Custom Gaze covers localized eye edits, whereas COCOAI covers fully generated people and interactions. We therefore interpret robustness across these two person-centric distributions.

Training-time mask construction uses face boxes, while inference consumes only the image and the text prompt. The fixed band covers both detected eye centers for 97.5% of audited faces and remains above 95.8% across face-size quartiles. Pose-stratified evaluation provides a complementary behavioral robustness view.

We stratify COCOAI images by the absolute yaw of the largest detected face: frontal below 25 degrees, semi-profile from 25 to 45 degrees, and profile above 45 degrees. Values are means over seeds 42, 47, and 53, with unknown predictions counted as incorrect.

Performance is at or above the origin model in every pose bucket. The smaller gain in profile scenes is consistent with the reduced visibility of bilateral eye evidence, and the pose table summarizes behavioral robustness under changes in viewing condition.

E.3 Additional Controls and Error Analysis

The authentic-eye transformation retains the original eyes while reproducing resizing, JPEG encoding, blurred-mask compositing, and crop placement. Band-internal mean absolute error ranges from 2.4 to 7.6, compared with approximately 2 outside the band, confirming a visible processing inter-

Table E.2: Pose-stratified BA, FakeVLM origin → ours.

Dataset	Pose	Images	Origin BA	Ours BA	BA difference
COCOAI Person	Frontal	6,326	82.3	**83.2**	+0.8
	Semi-profile	2,465	80.3	**81.9**	+1.6
	Profile	2,636	85.2	**85.5**	+0.3
COCOAI Interaction	Frontal	46	65.9	**67.5**	+1.6
	Semi-profile	57	79.3	**80.6**	+1.4
	Profile	85	63.9	**64.1**	+0.2

vention. The model retains 2,201–2,212 parseable images as real, with 3–9 fake calls depending on the seed.

We also applied a transformation that mirrors the authentic eye patch horizontally within the same pipeline. The transformed set receives a fake-call rate of 0.1–0.4%, almost identical to the processing control. An independent gaze-change proxy estimates a median induced change of 4.7 degrees, with 78.2% of valid faces below 10 degrees. Given this change magnitude, we interpret the mirror result as a supplementary processing control rather than as a direct intervention on gaze geometry.

On COCOAI Person, errors concentrate in outputs that combine an interaction-style scene description with a natural-gaze conclusion. Across seeds, 404/768, 429/762, and 428/756 fake-to-real errors contain the gaze card, or 52.6%, 56.3%, and 56.6%. Other recurring errors involve smooth skin, cosmetic retouching, extreme expressions, and generic natural-lighting rationales.

This distribution makes scene gating a plausible calibration target, especially for single-person frames. No gating mechanism is implemented or evaluated here, so the error concentration is reported as a diagnostic rather than as a demonstrated mitigation.

For reproducibility, Custom Gaze generation uses FLUX.1-Fill-dev in bfloat16 with 20 inference steps and guidance scale 25. Held-out SDXL edits (Podell et al., 2024) use 30 steps, guidance 8.0, and strength 0.99 at 1,024 pixels before downsampling to 512. SD3 (Esser et al., 2024) uses 28 steps and guidance 7. RePaint (Lugmayr et al., 2022) uses 50 steps with jump length and jump count 5 at 256 pixels before resizing. Each held-out set contains 2,324 real references and 2,324 regenerated eye-band counterparts; 18 source cases without a usable face box are omitted during construction.

Primary model comparisons use frozen checkpoints and three independent decoding seeds. Statistical scripts operate on per-sample predictions, preserve class ratios during bootstrap, and report seed-specific intervals. Dataset identifiers, source pointers, face boxes, target-face selections, mask coordinates, generation seeds, and prediction records should accompany the released artifact.