AWAKENING VERBALIZER FORENSICS: TRAINING-FREE READOUT ALIGNMENT FOR AI-GENERATED IMAGE DETECTION IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) show promise for AI-generated image detection, but existing approaches mainly fine-tune them with binary real/fake supervision that, even with explanatory objectives, easily overfits to training fake patterns and fails to generalize. We find that frozen MLLM representations already contain transferable forensic information that the pretrained verbalizer, *i.e.*, vocabulary projection, does not fully exploit. We introduce Forensic Readout Alignment (FRA), which estimates a forensic direction with a boundary offset using real images and single generator's fake images, and encodes them into the vocabulary projection weights of the real/fake answer tokens, without updating the backbone or any gradient-based training. FRA preserves the original distance and midpoint between the two answer rows, keeping the fitted boundary calibrated against the rest of the vocabulary. Using the same training data as the evaluated methods, FRA achieves the best overall cross-generator detection performance, outperforming LoRA-based adaptations, and remains effective across different training generators, training-set sizes, and MLLM backbones. Overall, FRA provides a lightweight way to effectively awaken pretrained forensic capabilities, while preserving the backbone's general language and visual understanding capabilities.

1 INTRODUCTION

Detecting AI-generated images requires generalizing to generators unseen during training. Detectors that learn synthesis artifacts often overfit to a fixed generator family, degrading under new generation methods or post-processing (Wang et al., 2020; Zhu et al., 2023; Li et al., 2025). A natural remedy is to build detection on top of broadly pretrained visual representations, which capture generator-agnostic patterns and improve cross-generator transfer (Ojha et al., 2023; Liu et al., 2024; Yan et al., 2025b). Multimodal large language models (MLLMs) push this idea further: on top of strong visual priors, they offer a language interface that couples detection with human-readable explanation. To turn this potential into forensic performance, recent work adapts MLLMs through specialized supervision, feature integration, and auxiliary tasks (Zhou et al., 2025; Lin et al., 2026; Xu et al., 2026). Nevertheless, real/fake data is far narrower than MLLM pretraining corpora, so gradient-based adaptation readily overfits the training fake patterns and erodes the generalization that MLLMs were meant to provide. Before committing to further adaptation, however, **one may ask a more fundamental question**: *how much forensic capability already lives inside a frozen MLLM, and how can we effectively harness this latent potential for detection while keeping the model's general capabilities intact?*

To address this question, we probe a frozen Qwen3-VL, comparing the forensic information encoded in its representations with that exposed by its original readout. In fact, the resulting forensic score is the inner product of the final hidden state and a readout vector given by the fake-minus-real difference between the answer-token vectors in the vocabulary output layer. Figure 1(a) projects frozen features from unseen benchmarks onto a two-dimensional score space defined by the original readout direction and an orthogonal reference direction. Notably, real and fake samples overlap along the original readout direction, and the corresponding decision boundary (dashed line) fails to separate them. However, the two-dimensional view reveals class structure, suggesting that discriminative information is retained in the representations but poorly captured by the original readout.

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.

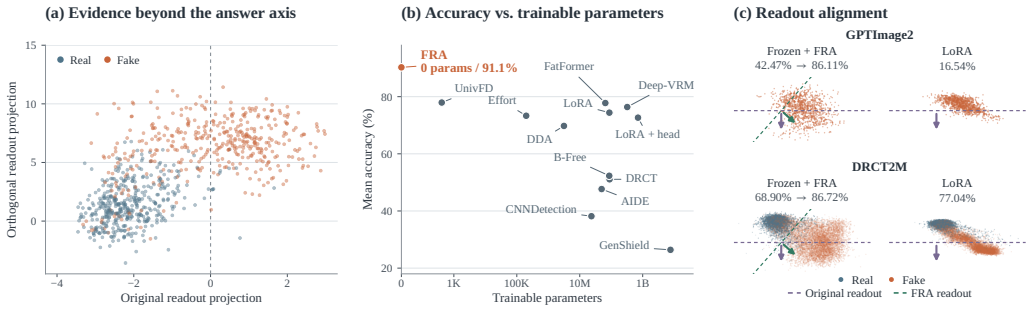

Figure 1: **Frozen MLLMs may already carry forensic evidence, and a training-free readout realignment can turn it into strong detection.** (a) Frozen Qwen3-VL features projected onto the original answer direction and an orthogonal direction estimated from the training data. The dashed line marks the original decision boundary. (b) Mean accuracy across seven benchmarks versus trainable parameters using the same training data. (c) Frozen features with the original and FRA readouts (left) and LoRA-adapted features with the original readout (right) on GPTImage2 and DRCT2M. Arrows indicate the directions of increasing fake scores.

We refer to this mismatch as **forensic verbalizer misalignment**: the pretrained verbalizer's readout direction is poorly aligned with a forensic discriminative direction in the frozen MLLM representation space. To resolve this mismatch with negligible impact on the MLLM's general capabilities, we propose **Forensic Readout Alignment (FRA)**, a lightweight training-free method that estimates a forensic direction and a boundary offset from a small set of real images and fake images from a single generator, and encodes them into the verbalizer's answer-token rows while leaving the backbone frozen. As shown in Figure 1(b) and Table 1, FRA achieves higher mean accuracy across seven cross-generator benchmarks than the state-of-the-art detectors, using the same training data but requiring zero trainable parameters and no gradient-based training. To reveal the underlying mechanism, Figure 1(c) compares the feature geometries and decision boundaries of FRA and LoRA on GPTImage2 and DRCT2M. Although parameter-efficient, LoRA still modifies the backbone, adjusting the learned representations to fit training fake patterns, potentially overfitting to these patterns when data are limited. In contrast, FRA preserves the pretrained informative representation space and uses the training data only to estimate a forensic decision boundary. This readout-level alignment is less data-demanding than reshaping the underlying representations, making FRA better suited to generalization from a narrow range of training forgeries to unseen generators.

Overall, FRA turns the forensic evidence available in a frozen MLLM into a detector through its existing verbalizer. This shows that MLLMs can serve as effective cross-generator detectors by simply realigning their vocabulary readouts, without learning new forensic representations. Our main contributions are listed as follows:

- We identify forensic verbalizer misalignment: frozen MLLM representations already contain transferable forensic information, but their pretrained vocabulary readouts do not fully exploit it. This finding motivates readout alignment rather than learning new forensic representations.
- We propose FRA, a novel verbalizer-based approach to forensic detection with frozen MLLMs. Without gradient-based training, FRA fits a forensic direction and boundary offset from real images and single-generator fakes in closed form, encoding both into the real/fake token rows while preserving their distance and midpoint.
- We show that FRA achieves the best overall performance across seven benchmarks covering diverse unseen generators. Further experiments with different training generators, training-set sizes, and MLLM backbones validate its robustness and computational efficiency.

2 Related Work

2.1 Traditional AI-generated image detection

Traditional detectors typically distinguish real from generated images using task-specific decision rules or classifiers. Early methods learn synthesis artifacts with convolutional networks and data aug-

mentation (Wang et al., 2020). Reconstruction-based methods exploit pretrained generative models: DIRE (Wang et al., 2023) uses diffusion reconstruction error, while AEROBLADE (Ricker et al., 2024) uses autoencoder reconstruction error to detect latent-diffusion images without detector training. Recent methods use pretrained representations. UnivFD (Ojha et al., 2023) and Gaussian discriminant analysis (Wang et al., 2024) use frozen CLIP features with linear and statistics-based readouts, respectively. FatFormer (Liu et al., 2024) introduces forgery-aware adapters and vision–language alignment; Effort (Yan et al., 2025b) adapts representations through orthogonal subspaces. AIDE (Yan et al., 2025a) combines OpenCLIP features with local artifacts and noise cues. DRCT (Chen et al., 2024) uses diffusion reconstruction and contrastive training, while DINOv2-based B-Free (Guillaro et al., 2025) and DDA (Chen et al., 2025) improve generalization through bias-reduced data construction and pixel–frequency alignment, respectively. Despite these advances, detection cues may miss semantic inconsistencies in objects, scene structure, or contextual relationships.

2.2 MLLM-BASED AI-GENERATED IMAGE DETECTION

MLLM-based methods use pretrained semantic knowledge to assess image authenticity. AIGI-Holmes (Zhou et al., 2025) combines visual experts with instruction tuning, preference optimization, and collaborative decoding. Deep-VRM (Lin et al., 2026) injects low-level visual features into intermediate language-model layers to improve forensic signal perception. GenShield (Xu et al., 2026) combines detection with artifact correction through joint training and a visual reasoning curriculum. Such feature integration and multi-stage training increase adaptation complexity, while task-specific supervision may encourage overfitting to training forgery patterns and limit transfer to unseen generators. FRA instead aligns the readout by editing only the existing real/fake token rows, with the backbone frozen. This uses pretrained representations without extensive task-specific training or additional forensic modules, reducing the risk of overfitting to training forgery patterns.

3 FORENSIC READOUT ALIGNMENT

3.1 OVERVIEW

As shown in Figure 2, FRA adapts the real/fake readout of a frozen MLLM while keeping image representations fixed. For an image x and a fixed detection prompt, let $\hat{\boldsymbol{h}}(x) \in \mathbb{R}^D$ denote the final-layer hidden state input to the frozen MLLM's vocabulary head, and $\boldsymbol{w}_{\text{real}}$ and $\boldsymbol{w}_{\text{fake}}$ its answer-token rows. Decoding restricted to these tokens predicts fake if

$$s(x) = \boldsymbol{w}^\top \hat{\boldsymbol{h}}(x) > 0, \qquad \boldsymbol{w} = \boldsymbol{w}_{\text{fake}} - \boldsymbol{w}_{\text{real}}. \tag{1}$$

The verdict depends on a single readout direction $\boldsymbol{w}$. The bias-free output layer places the boundary through the origin, so scaling $\boldsymbol{w}$ alone cannot shift it.

FRA modifies this readout to capture forensic cues and calibrate the threshold while keeping image representations fixed:

$$\boldsymbol{w}' = \boldsymbol{w} + \eta \|\boldsymbol{w}\| \hat{\boldsymbol{u}} - b\boldsymbol{v}. \tag{2}$$

Here, $\hat{\boldsymbol{u}}$ is a unit forensic direction estimated from real and fake images, and η controls its contribution. The carrier direction $\boldsymbol{v}$ has nearly constant projection $\boldsymbol{v}^\top \hat{\boldsymbol{h}}(x)$ across training images, so $-b\boldsymbol{v}$ approximates a threshold shift without an explicit bias. Both η and b are calibrated using real images.

FRA positively rescales the complete readout and writes it into the real/fake vocabulary rows, preserving their original midpoint and distance while keeping all other weights frozen. The following subsections detail the computation of these quantities and the readout's encoding into the verbalizer.

3.2 FORENSIC DIRECTION ESTIMATION

Given frozen hidden states $\{(\hat{\boldsymbol{h}}_i, y_i)\}_{i=1}^n$ from real ($y_i = 0$) and fake ($y_i = 1$) training images, we define $\boldsymbol{D}$ as the diagonal matrix of per-dimension standard deviations across these features. In standardized coordinates, μ_0 and μ_1 denote the class means, and S_w is the pooled within-class covariance. We estimate a regularized Fisher direction (Fisher, 1936) from these labeled features and map it back to the original coordinates:

$$\tilde{\boldsymbol{u}} = \boldsymbol{D}^{-1} \left(S_w + \lambda \tfrac{\operatorname{tr} S_w}{D} \boldsymbol{I} \right)^{-1} (\mu_1 - \mu_0), \qquad \hat{\boldsymbol{u}} = \tilde{\boldsymbol{u}} / \|\tilde{\boldsymbol{u}}\|. \tag{3}$$

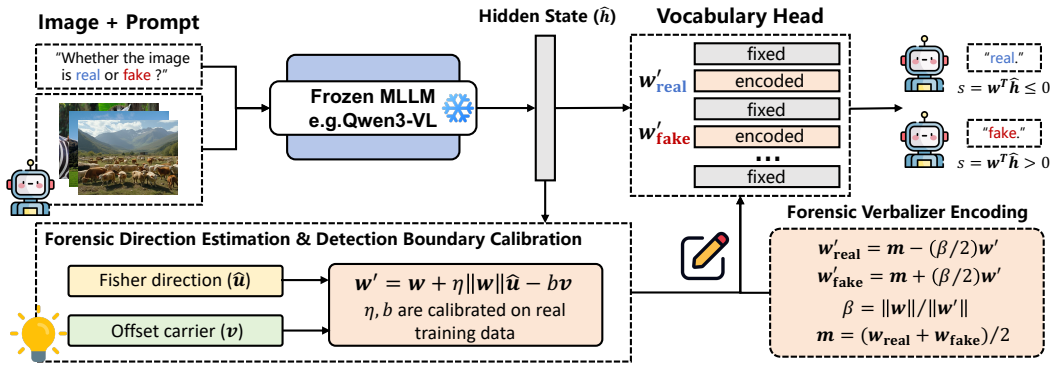

Figure 2: **FRA overview.** The upper path performs inference with a frozen MLLM and a modified vocabulary head. The lower branch estimates the forensic direction, carrier, and calibration coefficients once from training data. The complete readout is positively scaled and encoded into the answer tokens, preserving their original midpoint and distance while leaving all other weights fixed.

This separates the class means while accounting for within-class variation. Scaling shrinkage by the average eigenvalue of S_w makes λ scale-free. The factor $\boldsymbol{D}^{-1}$ restores the original coordinates before normalization. Notably, only the readout direction is estimated, with the backbone frozen.

3.3 Detection Boundary Calibration

To encode an offset through the existing output layer, we seek a carrier whose projection stays close to one. Given mean $\mu \neq 0$ and positive-definite covariance Σ in readout coordinates, the direction minimizing projection variance subject to unit mean is

$$\boldsymbol{v} = \underset{\boldsymbol{a}:\, \boldsymbol{a}^\top \mu = 1}{\arg\min}\, \boldsymbol{a}^\top \Sigma \boldsymbol{a} = \frac{\Sigma^{-1}\mu}{\mu^\top \Sigma^{-1}\mu}. \tag{4}$$

FRA uses its ridge-regularized empirical counterpart, whose fitted projections concentrate near their unit mean and remain positive on training images (Appendix A.5). A linear term along this direction therefore approximates a constant readout offset.

We set η by matching the standard deviations of the original and added forensic score terms on real training images:

$$\eta = \frac{\mathrm{sd}\left(\boldsymbol{w}^\top \hat{\boldsymbol{h}}\right)}{\|\boldsymbol{w}\|\ \mathrm{sd}\left(\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{h}}\right)}. \tag{5}$$

To compute the boundary offset b, we first normalize the combined score by the carrier projection for each of the n_0 real training images:

$$\zeta_i = \frac{(\boldsymbol{w} + \eta\|\boldsymbol{w}\|\hat{\boldsymbol{u}})^\top \hat{\boldsymbol{h}}_i}{\boldsymbol{v}^\top \hat{\boldsymbol{h}}_i}.$$

We then set b to an upper-tail order statistic of these normalized scores:

$$b = \zeta_{(k)}, \qquad k = \lceil (n_0 + 1)(1 - \alpha) \rceil. \tag{6}$$

Here $\zeta_{(1)} \leq \cdots \leq \zeta_{(n_0)}$, $k \leq n_0$, and α sets the upper-tail level for real training images. Since the fitted carrier projections are positive on these images, subtracting $b\boldsymbol{v}$ from the combined readout weights makes the aligned score positive exactly when $\zeta_i > b$. The direction, carrier, and coefficients are all estimated from the training data, without validation-based search or use of additional data.

3.4 Forensic Verbalizer Encoding

We encode the fitted forensic readout into the answer-token weights. Since its magnitude affects competition with the rest of the vocabulary, we first rescale it to match the original distance between

the answer-token weight vectors while preserving their midpoint:

$$\boldsymbol{m} = \frac{\boldsymbol{w}_{\text{fake}} + \boldsymbol{w}_{\text{real}}}{2}, \qquad \beta = \frac{\|\boldsymbol{w}\|}{\|\boldsymbol{w}'\|}. \tag{7}$$

The encoded weights for the real and fake answer tokens are

$$\boldsymbol{w}'_{\text{real}} = \boldsymbol{m} - \frac{\beta}{2}\boldsymbol{w}', \qquad \boldsymbol{w}'_{\text{fake}} = \boldsymbol{m} + \frac{\beta}{2}\boldsymbol{w}'. \tag{8}$$

Their logit difference is $\beta \boldsymbol{w}'^{\top} \hat{\boldsymbol{h}}(x)$. Scaling the entire readout, including its carrier offset, by a positive factor preserves the fitted boundary and score ordering. Among all weight pairs with difference $\beta \boldsymbol{w}'$, retaining the original midpoint minimizes the total squared change.

The midpoint and scale follow directly from the original verbalizer weights and fitted readout. All other vocabulary head weights and the backbone remain fixed. Fitting requires two $D \times D$ linear solves, while encoding uses only vector operations and requires no gradient computation. When $\eta = b = 0$, the original verbalizer weights are recovered.

At a fixed prefix, hidden states and all non-answer token logits remain unchanged. However, changing the answer-token logits affects probabilities across the vocabulary through softmax normalization and may influence subsequent generation. We therefore empirically assess answer behavior preservation in Section 4.5.

4 Experiments

4.1 Experimental setup

Datasets and metrics. We use $2{,}500$ class-balanced training images from GenImage (Zhu et al., 2023), with Stable Diffusion v1.4 (SD1.4) as the training generator. We evaluate on sampled subsets of seven benchmarks: AIGINow (Chen et al., 2026), DRCT2M (Chen et al., 2024), EvalGEN (Chen et al., 2025), GPTImage2 (Zewde et al., 2026), MNW (Roca et al., 2026), T2ICoReBench (Li et al., 2026), and Chameleon (Yan et al., 2025a). We report accuracy (ACC), equivalent to fake recall on generated-only subsets. Mean averages subset scores equally within each benchmark, then equally across the benchmarks in each comparison. The main detector comparisons use all seven benchmarks; other tables and figures specify their evaluation scope.

Implementation details. FRA fits the frozen Qwen3-VL-8B-Instruct (Bai et al., 2025) readout in closed form, with default Fisher shrinkage $\lambda = 0.1$, carrier ridge $\rho = 10^{-3}$, and tail level $\alpha = 0.05$ on real training images. All detection parameters follow the estimation rules in Section 3. Detection comparisons use the fitted linear readout; answer-behavior experiments use full-vocabulary outputs after write-back, with scale fixed by Eq. (7). Trainable baselines in Table 1 share the training images and binary labels. DRCT (Chen et al., 2024), DDA (Chen et al., 2025), and B-Free (Guillaro et al., 2025) detection variants (†) omit additional reconstruction- or synthesis-based training-data construction. Deep-VRM and GenShield (‡) use binary supervision, retaining two-stage adaptation and the Only Det. setting, respectively.

We train two rank-32 LoRA variants on on the attention and MLP projections of Qwen3-VL's language model, with frozen (LoRA) or jointly trained (LoRA + head) unembedding. Both are trained for one epoch with answer-token cross-entropy, using AdamW with a learning rate of 2×10^{-5} and an effective batch size of 32.

Figure 1(a) uses the Fisher component estimated from training data, orthogonal to the original answer-row difference, with normalization using real training images and equal samples per class and benchmark. Figure 1(b) reports the mean performance across seven benchmarks and the number of parameters optimized by gradient descent, as listed in Table 1. For Deep-VRM, we sum the parameter counts across both stages. Figure 1(c) projects frozen features onto the plane spanned by the original and FRA readouts. The LoRA variant retains the original vocabulary head.

4.2 Generalization evaluation

Performance comparison using the same training data. We compare FRA with CNNDetection (Wang et al., 2020), UnivFD (Ojha et al., 2023), FatFormer (Liu et al., 2024), AIDE (Yan

Table 1: **Performance comparison using the same training data.** Trainable baselines and FRA use the same 2,500 training images and binary labels. Scores (%) follow the benchmark aggregation in Section 4.1; Mean equally averages the seven benchmarks. Train. params counts gradient-optimized parameters. Bold and underlined scores mark the best and second-best results. † marks detection variants without additional training-data construction; ‡ marks binary-supervision adaptations.

Method	Train. params	AIGINow	DRCT2M	EvalGEN	GPTImage2	MNW	T2ICoReBench	Chameleon	Mean ↑
CNNDetection	23.510 M	52.58	67.87	47.73	7.73	13.85	22.08	55.36	38.17
UnivFD	769	82.14	74.90	96.14	<u>77.30</u>	<u>77.66</u>	83.26	54.21	<u>77.94</u>
FatFormer	64.975 M	72.17	72.31	95.91	71.43	71.51	85.41	75.68	77.77
AIDE	49.548 M	65.69	65.02	57.73	15.75	27.63	40.96	60.89	47.67
DRCT†	88.618 M	63.75	71.21	65.45	8.02	37.91	49.84	61.39	51.08
Effort	0.199 M	88.00	<u>82.17</u>	**99.32**	29.75	71.69	81.45	60.85	73.32
B-Free†	86.527 M	55.17	39.44	94.32	2.84	45.66	76.90	51.90	52.32
DDA†	3.147 M	70.50	80.60	95.23	23.19	66.52	82.18	70.07	69.76
Deep-VRM‡	437.985 M	96.64	79.88	<u>97.95</u>	26.22	67.63	<u>97.61</u>	68.69	76.37
GenShield (Only Det.)‡	7.633 B	50.00	65.62	12.73	0.00	2.09	3.42	50.94	26.40
Qwen3-VL-8B (LoRA)	87.294 M	**98.08**	77.04	94.55	16.54	63.38	94.13	<u>77.10</u>	74.40
Qwen3-VL-8B (LoRA + head)	709.624 M	<u>98.03</u>	76.18	94.55	13.31	59.69	92.75	74.07	72.65
Qwen3-VL-8B (frozen)	0	81.72	68.90	69.32	42.47	48.06	79.10	75.11	66.38
FRA (ours)	**0**	92.94	**86.72**	<u>97.95</u>	**86.11**	**90.15**	**99.59**	**84.26**	**91.10**

et al., 2025a), DRCT (Chen et al., 2024), Effort (Yan et al., 2025b), B-Free (Guillaro et al., 2025), DDA (Chen et al., 2025), Deep-VRM (Lin et al., 2026), and GenShield (Xu et al., 2026), the frozen Qwen3-VL-8B reference, and two LoRA controls under identical downstream training images and binary supervision to compare their use of the training set.

FRA achieves the highest mean acc, ahead of UnivFD, and improves over frozen Qwen3-VL-8B on all seven benchmarks (Table 1). It leads on DRCT2M, GPTImage2, MNW, T2ICoReBench, and Chameleon, where the original MLLM is already competitive. These gains arise entirely from readout alignment, making existing representations more useful for cross-dataset detection.

The same-backbone LoRA controls perform strongly on AIGINow, but FRA scores higher on the remaining benchmarks, particularly GPTImage2 and MNW. Under this protocol, direct readout fitting thus provides broader transfer than the tested answer-token fine-tuning configurations while keeping the backbone frozen. The corresponding MNW breakdown in Appendix A.6 further shows gains across the displayed generator families.

Performance comparison with released checkpoints. We evaluate released checkpoints on the same seven benchmarks using native preprocessing and default decision rules. They retain their original supervision; Table 2 lists the generator counts. GenShield is omitted because a trained detector checkpoint was unavailable for this comparison. AIDE's count follows its published SD1.4 protocol, although its released checkpoint does not identify the training subset.

FRA also achieves the highest Mean against released weights (Table 2), leading on GPTImage2, MNW, and T2ICoReBench. Deep-VRM leads on AIGINow, EvalGEN, and Chameleon, where FRA ranks second; B-Free leads on DRCT2M. Released checkpoints thus retain benchmark-specific strengths despite FRA's higher aggregate performance.

Transfer across MLLM backbones. We apply the same fitting procedure to each of the six MLLM backbones' frozen hidden states, using identical training images and fixed hyperparameters. Table 3 compares their original and aligned readouts on all seven benchmarks, including Chameleon.

FRA improves Mean for every backbone and scores on every benchmark for DeepSeek, Kimi, and GLM, including those with low native generated-image recall. These gains beyond Qwen and InternVL show that weak native detection can coexist with transferable forensic information. Transfer still depends on the backbone: Qwen3-VL has the highest Mean, GPTImage2 and MNW scores vary across models, and InternVL3.5 remains slightly below its frozen reference on GPTImage2.

Table 2: **Performance comparison with released checkpoints.** Scores (%) on the same seven benchmarks as Table 1; Mean equally averages them. #Gen. counts generators used for forensic supervision, including reconstruction models; § marks an inferred count. Bold and underlined values mark the best and second-best results.

Method	#Gen.	AIGINow	DRCT2M	EvalGEN	GPTImage2	MNW	T2ICoReBench	Chameleon	Mean ↑
CNNDetection	1	49.36	49.98	4.77	0.10	0.31	1.17	49.98	22.24
UnivFD	1	52.36	61.86	22.95	4.79	9.11	15.82	50.98	31.12
FatFormer	1	51.67	51.99	40.00	7.24	12.06	20.32	51.06	33.48
AIDE	1§	67.42	62.56	97.05	4.89	29.72	84.25	61.74	58.23
DRCT	2	65.31	90.59	74.32	9.78	57.23	64.44	68.31	61.43
Effort	1	64.83	69.88	97.27	18.40	61.11	87.01	59.89	65.48
B-Free	1	61.92	**99.21**	84.55	22.99	69.29	77.22	76.72	70.27
DDA	1	70.06	<u>97.99</u>	95.00	20.45	82.65	92.80	81.94	77.27
Deep-VRM	4	**95.53**	87.51	**100.00**	<u>47.65</u>	<u>86.15</u>	<u>99.22</u>	**93.85**	<u>87.13</u>
Qwen3-VL-8B (frozen)	0	81.72	68.90	69.32	42.47	48.06	79.10	75.11	66.38
FRA	1	<u>92.94</u>	86.72	<u>97.95</u>	**86.11**	**90.15**	**99.59**	<u>84.26</u>	**91.10**

Table 3: **Generalization across MLLM backbones.** Scores (%) on the seven benchmarks evaluated for all backbones. Mean equally averages these benchmarks. Bold marks the better readout per backbone.

Backbone	Readout	AIGINow	DRCT2M	EvalGEN	GPTImage2	MNW	T2ICoReBench	Chameleon	Mean ↑
Qwen3-VL-8B	Frozen	81.72	68.90	69.32	42.47	48.06	79.10	75.11	66.38
	FRA	**92.94**	**86.72**	**97.95**	**86.11**	**90.15**	**99.59**	**84.26**	**91.10**
Qwen3.5-9B	Frozen	50.00	50.00	0.00	0.00	0.00	0.00	50.00	21.43
	FRA	**95.53**	**78.12**	**99.55**	**77.40**	**86.83**	**99.50**	**82.10**	**88.43**
InternVL3.5-8B	Frozen	68.61	58.42	39.77	**47.06**	17.42	53.51	62.34	49.59
	FRA	**97.75**	**80.74**	**94.55**	45.89	**52.00**	**97.67**	**69.38**	**76.85**
deepseek-vl2-small	Frozen	52.11	55.37	3.86	1.66	2.15	2.51	55.28	24.71
	FRA	**94.67**	**85.78**	**98.41**	**79.35**	**79.08**	**99.21**	**79.12**	**87.95**
Kimi-VL-A3B-Instruct	Frozen	52.08	54.41	2.27	2.45	1.78	2.53	59.23	24.96
	FRA	**91.47**	**83.07**	**91.36**	**82.09**	**68.00**	**94.70**	**80.60**	**84.47**
GLM-4.6V-Flash	Frozen	50.03	50.01	0.00	0.00	0.00	0.01	50.01	21.44
	FRA	**96.00**	**83.88**	**95.45**	**71.23**	**77.54**	**97.73**	**75.89**	**85.39**

Table 4: **Component ablations.** Qwen3-VL-8B accuracy (%) on three benchmarks. Mean averages the three displayed benchmarks. Bias denotes an explicit scalar offset; bold marks column maxima.

Variant	AIGINow	DRCT2M	EvalGEN	Mean ↑
Frozen	81.72	68.90	69.32	73.31
w/o direction	83.06	70.08	72.05	75.06
w/o covariance	84.83	71.74	75.45	77.34
w/o offset	63.11	54.51	36.59	51.40
Linear probe	92.31	83.13	92.59	89.34
Fisher + bias	87.78	82.37	94.17	88.10
$\boldsymbol{w}$ + Fisher + bias	87.92	83.18	94.95	88.68
FRA	**92.94**	**86.72**	**97.95**	**92.54**

Table 5: **Effect of the training generator.** FRA scores (%) with different training generators. Mean averages the four displayed benchmarks; bold marks column maxima.

Generator	AIGINow	DRCT2M	EvalGEN	T2ICoRe Bench	Mean ↑
SD1.4	92.94	86.72	**97.95**	99.59	94.30
SD1.5	93.53	**86.92**	97.73	99.51	**94.42**
ADM	88.11	83.47	90.68	92.55	88.70
GLIDE	85.97	86.15	95.68	98.32	91.53
VQDM	88.75	84.44	92.05	89.72	88.74
Wukong	93.50	85.59	97.05	99.10	93.81
Midjourney	**94.28**	85.99	97.73	**99.64**	94.41
BigGAN	88.92	86.08	87.27	92.16	88.61

4.3 Ablation Study

Table 4 tests the roles of direction, covariance, and offset on Qwen3-VL-8B, with learned linear-probe and explicit scalar-bias references. Table 5 varies the training generator used to fit the readout.

Table 6: **Adaptation time.** Baseline training and FRA feature extraction plus solve-and-write time. FRA's stages are listed below its total.

Method	Epochs	Time (s)
UnivFD	17	318.5
FatFormer	6	421.1
Effort	25	480.5
Deep-VRM	2 + 4	1464.1
LoRA	1	490.7
LoRA + head	1	491.6
Linear probe	–	71.2
FRA	–	**68.7**
Feature extraction	–	68.3
Solve + write	–	0.4

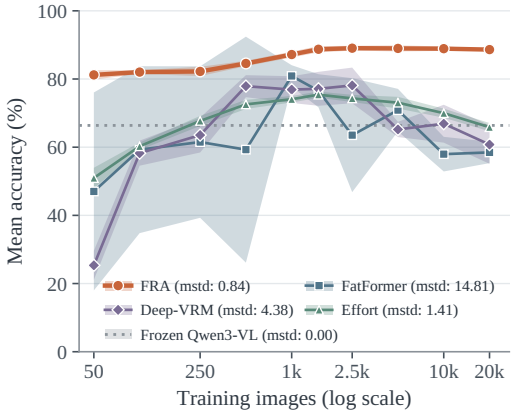

Figure 3: **Data efficiency.** Mean accuracy over seven benchmarks.

Direction and boundary placement. Without the forensic direction, performance stays close to the frozen model, so offset adjustment alone does not recover FRA's gains. Omitting covariance weakens performance across benchmarks; removing the offset causes the largest drop, below the frozen reference. Thus, adding a direction without shifting the boundary can impair detection, while fitting both yields the strongest readout among these variants.

Alternative readouts. A learned linear probe and the explicit-bias controls improve over the frozen readout. Retaining the original direction $\boldsymbol{w}$ modestly improves the explicit-bias construction. FRA outperforms both the linear probe and this stronger $\boldsymbol{w}$ + Fisher + bias control on all three benchmarks, supporting the complete carrier-based readout among the evaluated implementations.

Sensitivity to the training generator. In Table 5, SD1.4 (the default), SD1.5, Midjourney, and Wukong yield similarly strong aggregate performance; GLIDE, ADM, VQDM, and BigGAN score lower. DRCT2M is comparatively stable, while EvalGEN and T2ICoReBench vary more, with BigGAN and VQDM respectively giving the weakest results. Effective alignment therefore extends across training generators, although generator choice still affects cross-benchmark transfer.

4.4 Adaptation efficiency

We analyze the computational cost of fitting the readout and how it scales with training-set size, assessing the efficiency of the fitting procedure as more training data become available.

Computation cost. Table 6 includes feature extraction, the closed-form solve, and the two-row write. FRA has the lowest reported adaptation time, with a total close to that of linear probing. Most of its cost comes from extracting frozen source features; fitting the readout and matching the answer-row distance add little overhead.

Data efficiency. Figure 3 compares seven-benchmark mean accuracy for FRA, FatFormer, Deep-VRM, and Effort at matched training-data budgets. Lines and bands show means and standard deviations over five runs per budget; legend values average these deviations across budgets, in percentage points. The dotted line is the fixed frozen Qwen3-VL reference. FRA re-estimates the statistics from each training set with fixed estimation rules.

FRA has the highest mean accuracy throughout the evaluated range, improving with training-set size before reaching a broad plateau, and the lowest budget-averaged standard deviation among adapted methods. FatFormer varies widely across runs, especially with limited data; Deep-VRM's mean fluctuates across budgets. Effort is more consistent across runs, but its early gains diminish as the training set grows. Additional sensitivity results in Appendix A.7 show stable performance across the tested carrier ridges and a plateau around the scale estimated from the training data.

Table 7: **Answer behavior preservation.** Agreement (%) with each model's frozen backbone: first-token selections on 90 questions and exact generation matches on 300 questions.

Method	First token	Generation
Frozen	100.00	100.00
LoRA	96.67	54.00
LoRA + head	96.67	42.00
Deep-VRM	55.56	71.00
FRA	**100.00**	**100.00**

Table 8: **Full MLLM versus vision-only features.** Scores (%) with FRA on three benchmarks. Each cell reports MLLM / ViT; bold marks the better result per pair. Backbone sizes follow Table 3.

Backbone	GPTImage2	MNW	Chameleon
Qwen3-VL	**86.11** / 30.82	**90.15** / 75.82	**84.26** / 79.19
Qwen3.5	**77.40** / 24.07	**86.83** / 83.88	**82.10** / 76.45
InternVL3.5	**45.89** / 43.74	52.00 / **60.31**	**69.38** / 66.34
DeepSeek-VL2	**79.35** / 22.50	**79.08** / 64.49	**79.12** / 70.09
Kimi-VL	**82.09** / 37.87	68.00 / **85.97**	**80.60** / 75.28
GLM-4.6V	**71.23** / 31.80	77.54 / **90.34**	**75.89** / 70.44

4.5 Answer behavior preservation

Evaluation settings. We compare the highest-scoring first token over the full vocabulary with the frozen model on 90 non-forensic questions across six categories under a fixed image context. Agreement measures consistency with the frozen model, not correctness. We also measure exact greedy-generation agreement on 300 ordinary questions across the same six categories, generating at most 64 new tokens. On 120 forensic prompts, we check whether the highest-scoring first token is real or fake without restricting the vocabulary. All questions are used only for evaluation. FRA and the LoRA controls use frozen Qwen3-VL-8B as their reference. Deep-VRM is compared with its own frozen Qwen2.5-VL-7B reference for both agreement metrics.

Preserving ordinary answers. FRA matches the frozen model on all 90 ordinary first-token selections and all 300 short generations (Table 7). Generation agreement is 54.0% for LoRA, 42.0% for LoRA + head, and 71.0% for Deep-VRM; these methods also change some first-token selections. These results concern only the tested questions and generation lengths.

Competition with the full vocabulary. FRA selects real or fake on all 120 forensic prompts without restricting the vocabulary. This measures answer-format compliance, not detection accuracy. At a fixed prefix, FRA leaves hidden representations and non-answer token logits unchanged. However, the edited answer tokens may outrank the original top token. The quantitative results above assess ordinary answer preservation, with additional qualitative comparisons provided in Appendix A.11.

4.6 Full MLLM versus its vision encoder

Table 8 compares FRA on each frozen MLLM's final representations and its frozen vision-encoder features. Full MLLMs perform better on GPTImage2 and Chameleon for every backbone, while vision-only features perform better on MNW for InternVL, Kimi, and GLM. These results suggest that language-model processing can improve visual features for detection, with benefits varying across data distributions. This motivates aligning the answer verbalizer to better access forensic signals in MLLM. Table 10 in Appendix A.9 provides the full seven-benchmark comparison.

5 Conclusion and Limitations

Frozen MLLM features contain transferable forensic signals that FRA accesses through a task-specific readout. Adjusting the readout direction and decision boundary yields strong aggregate detection performance without updating the backbone. FRA encodes aligned detector into the vocabulary while matching the original answer-row distance, preserving ordinary answers on the evaluated prefixes and short generations. These findings support direct readout alignment as an effective approach to adapting frozen MLLMs for detection. These findings also point to directions for further study. Evaluating answer preservation beyond the current question set and generation lengths would help assess its broader applicability. Further analysis could clarify factors underlying the observed transfer gains, while more efficient feature extraction and deployment could reduce MLLM inference costs.

AI USE STATEMENT

We used generative AI tools for language polishing and writing assistance, including improving clarity and organization, formatting tables and references, and checking consistency between the text and experimental results. The authors are responsible for the final manuscript, including its claims and interpretations.

ETHICS STATEMENT

This work studies the detection of AI-generated images using existing benchmarks. Such detectors can support authenticity assessment, but errors and distribution shifts may lead to incorrect judgments. Detection scores should therefore be treated as supporting evidence rather than definitive proof, with additional verification before consequential use.

REPRODUCIBILITY STATEMENT

Section 3 describes the readout construction and estimation rules, while Section 4.1 specifies the source data, evaluation benchmarks, metrics, and implementation settings. Appendix A provides supporting derivations and fitted coefficients for the default detector.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24ay.html.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=C39ShJwtD5.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Align, don't blend: Orthogonal supervision disentanglement as the key to generalizable AI-generated image detection. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026. URL https://huggingface.co/datasets/Gaffeyzz/AIGI-Now.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *CVPR*, pp. 18685–18694, 2025.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage,

but not direct the play? In *ICLR*, 2026. URL https://openreview.net/forum?id=iqAFhWistW.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *Advances in neural information processing systems*, 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fb693c67f61e5321746ffce8b6fdd2d0-Abstract-Datasets_and_Benchmarks_Track.html.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, and Shouhong Ding. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models. *arXiv preprint arXiv:2606.15880*, 2026. URL https://arxiv.org/abs/2606.15880.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, pp. 10770–10780, 2024.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *CVPR*, pp. 9130–9140, 2024.

Thomas Roca, Marco Postiglione, Chongyang Gao, Isabel Gortner, Zuzanna Wojciak, Pengce Wang, Mahsa Alimardani, Shirin Anlen, Kevin White, Juan Lavista Ferres, Sarit Kraus, Sam Gregory, V. S. Subrahmanian, and San Murugesan. The Microsoft-Northwestern-WITNESS Benchmark for Deepfake Detection . *IEEE Intelligent Systems*, 41(02):15–23, March 2026. ISSN 1941-1294. doi: 10.1109/MIS.2026.3668398. URL https://doi.ieeecomputersociety.org/10.1109/MIS.2026.3668398.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Js5PJPHDyY.

Zhipei Xu, Xuanyu Zhang, Youmin Xu, Qing Huang, Shen Chen, Taiping Yao, Shouhong Ding, and Jian Zhang. GenShield: Unified detection and artifact correction for AI-generated images. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.16122.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *ICLR*, volume 2025, pp. 70702–70720, 2025a.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 2025b. URL https://proceedings.mlr.press/v267/yan25b.html.

Kidus Zewde, Simiao Ren, Xingyu Shen, Jiaqi Wu, Yuchen Zhou, Tommy Duong, Zikang Zhang, Ethan Traister, and Kewen Xie. Gpt-image-2 in the wild: A twitter dataset of self-reported ai-generated images from the first week of deployment. *arXiv preprint arXiv:2604.25370*, 2026.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *ICCV*, pp. 18746–18758, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu YAN, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 77771–77782. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f4d4a021f9051a6c18183b059117e8b5-Paper-Datasets_and_Benchmarks.pdf.

A Supporting Readout Analysis

This appendix provides the loss decomposition, carrier characterization, and answer-row construction that support Section 3.

A.1 What answer-token supervision observes

Let $y = 0$ denote real images and $y = 1$ generated images. Write $\varepsilon(x)$ for the probability mass outside the two answer verbalizers at the evaluated prefix. Conditioning on the answer pair separates its relative preference from competition with the rest of the vocabulary.

Proposition 1 (What the objective can observe)**.** *The answer-token cross-entropy satisfies*

$$\mathcal{L}(x, y) = \ell\Big(\boldsymbol{w}^\top \hat{\boldsymbol{h}}(x), y\Big) - \log(1 - \varepsilon(x))\,, \tag{9}$$

where $\ell(s, y) = \log(1 + e^s) - ys$ is binary logistic loss. At a fixed readout, the conditional loss is invariant to any $\Delta\boldsymbol{h} \perp \boldsymbol{w}$. If the leakage before and after the perturbation is at most $\bar{\varepsilon} < 1$, the change in full-vocabulary loss has magnitude at most $-\log(1 - \bar{\varepsilon}) = O(\bar{\varepsilon})$.

Proof. The probability of the correct token is the mass on the answer pair, $1 - \varepsilon(x)$, multiplied by the conditional probability of that token. The conditional fake probability is $\sigma(\boldsymbol{w}^\top \hat{\boldsymbol{h}}(x))$; taking the negative logarithm gives Eq. (9). An orthogonal perturbation leaves this scalar score unchanged. The remaining term lies in $[0, -\log(1 - \bar{\varepsilon})]$ before and after the perturbation, which bounds its change. □

When leakage is small, the scalar term approximates the full answer loss. With a frozen output layer, LoRA must modify the representation to route information through the inherited direction. FRA instead estimates that direction from the frozen features and adjusts the boundary. The decomposition explains this design choice without asserting that a fixed readout prevents representation adaptation from succeeding.

A.2 The carrier and its source variation

For a nonzero source mean μ and positive-definite covariance Σ, Eq. (4) minimizes projection variance under a unit-mean constraint. This characterization describes when the linear carrier can approximate a constant offset.

Proposition 2 (When an offset is available)**.** *The minimizer in Eq. (4) has $\mathbb{E}[\boldsymbol{v}^\top \hat{\boldsymbol{h}}] = 1$ and*

$$\mathrm{Var}\Big(\boldsymbol{v}^\top \hat{\boldsymbol{h}}\Big) = \frac{1}{\mu^\top \Sigma^{-1} \mu}, \qquad \mathrm{cv}\Big(\boldsymbol{v}^\top \hat{\boldsymbol{h}}\Big) = \frac{1}{\sqrt{\mu^\top \Sigma^{-1} \mu}}, \tag{10}$$

where cv *denotes the coefficient of variation.*

Proof. The stationary condition for minimizing $\boldsymbol{a}^\top \Sigma \boldsymbol{a}$ subject to $\boldsymbol{a}^\top \mu = 1$ gives $\boldsymbol{a}$ proportional to $\Sigma^{-1}\mu$. The constraint fixes the normalization in Eq. (4), and positive definiteness makes this minimizer unique. Substitution into the quadratic form yields the variance. The mean is one, so its standard deviation is also its coefficient of variation. □

The projection is stable when the mean is far from the origin relative to the source spread, as measured by its Mahalanobis norm. The term $-b\boldsymbol{v}^\top \hat{\boldsymbol{h}}$ then approximates $-b$, with source root-mean-square variation $|b|/\sqrt{\mu^\top \Sigma^{-1} \mu}$. The implemented carrier uses ridge-regularized empirical statistics; the population expression characterizes the construction under its stated covariance assumptions.

A.3 Default fitted readout

For the default Qwen3-VL-8B detector, the training set contains $2{,}500$ class-balanced images, including $n_0 = 1{,}250$ real photographs. With $\alpha = 0.05$, the fitted coefficients in Eqs. (5) and (6) are $\eta = 40.14$ and $b = -45.43$. The write-back uses $\beta = 0.0248874249$, determined by the original answer-row distance and the norm of the complete fitted readout. The output layer contains $151{,}936$ rows; FRA edits the two declared answer rows using the $4{,}096$-dimensional difference in Eq. (8).

A.4 MINIMUM-CHANGE ANSWER ROWS

The write-back in Eq. (8) is the unique solution to

$$\min_{\boldsymbol{a}_{\text{fake}}-\boldsymbol{a}_{\text{real}}=\beta\boldsymbol{w}'} \|\boldsymbol{a}_{\text{fake}}-\boldsymbol{w}_{\text{fake}}\|^2+\|\boldsymbol{a}_{\text{real}}-\boldsymbol{w}_{\text{real}}\|^2. \tag{11}$$

The stationary condition gives $\boldsymbol{a}_{\text{fake}}+\boldsymbol{a}_{\text{real}}=\boldsymbol{w}_{\text{fake}}+\boldsymbol{w}_{\text{real}}$, so the optimal pair retains the original midpoint. Combining this identity with the prescribed difference gives Eq. (8); strict convexity ensures uniqueness. The scale in Eq. (7) also retains the original distance between the answer rows. Together, these properties determine the write-back directly from the fitted readout and the original answer rows.

A.5 TRAINING FIT DIAGNOSTICS

Figure 4(a) shows that both LoRA variants reduce the answer-token training loss to similarly low levels over five independent runs. Jointly training the vocabulary head follows a similar optimization trajectory to keeping it frozen. Despite their similarly low training losses, both variants transfer less effectively than FRA in Table 1.

Figure 4(b) verifies that the carrier $\boldsymbol{v}$ provides an approximately constant projection on training images, allowing a readout offset to be encoded through the existing weights. The ridge-regularized empirical counterpart of Eq. (4) fixes the training mean of $\boldsymbol{v}^\top\hat{\boldsymbol{h}}$ to one while minimizing a regularized variance objective. The horizontal axis plots this projection for real and fake training images. When the projection stays close to one, adding $-b\boldsymbol{v}$ to the readout weights contributes

$$(-b\boldsymbol{v})^\top\hat{\boldsymbol{h}}=-b(\boldsymbol{v}^\top\hat{\boldsymbol{h}})\approx -b.$$

Thus the carrier implements an approximately constant bias through the existing weight matrix, without an additional bias parameter. Notably, the two class distributions overlap tightly near one, supporting the carrier's role as a stable offset coordinate. The measured coefficient of variation is 0.0018 (0.18%) overall and 0.0016 on real images. The minimum projection is 0.9917. The carrier ranks the training classes near chance and is nearly orthogonal to the forensic direction. Its near-constant behavior is measured on the training distribution; target-domain behavior depends on the target features.

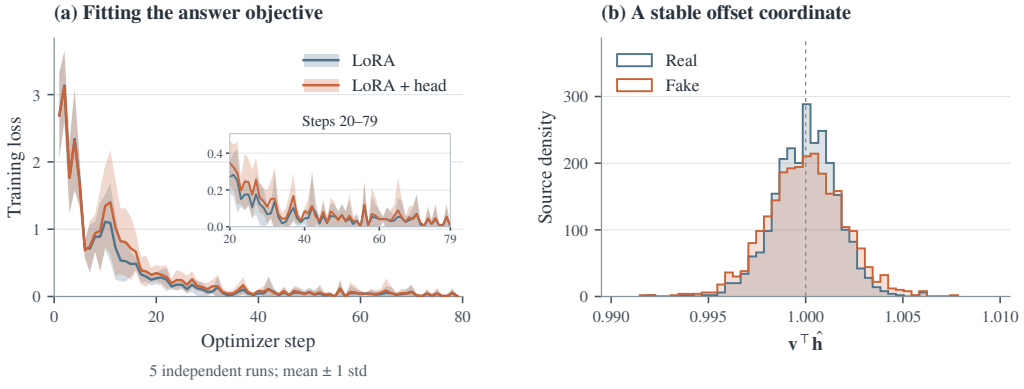

Figure 4: **Training fit and carrier stability.** (a) Answer-token training loss for LoRA with a frozen vocabulary head (LoRA) and a jointly trained vocabulary head (LoRA + head), using the same $2{,}500$ training images. Curves and shaded bands show the mean and one standard deviation over five independent runs; the inset enlarges steps 20–79. (b) Distributions of the carrier projection $\boldsymbol{v}^\top\hat{\boldsymbol{h}}$ for real and fake training images. Both distributions concentrate near one, with an overall coefficient of variation of 0.18%, supporting the use of $\boldsymbol{v}$ to encode an approximately constant readout offset.

A.6 TRANSFER ACROSS GENERATOR FAMILIES

To resolve the transfer results by generator family, Table 9 compares FRA with the SD1.4-trained baselines from Table 1 and the frozen Qwen3-VL reference on MNW. We group its subsets by named

model series or providers, retaining their versions and task variants. Each displayed group averages its constituent subsets equally, while Mean averages all subsets; Others collects the remaining subsets.

Table 9: **Fine-grained detection on MNW using SD1.4 training data.** Fake recall (%) uses the same checkpoints and method variants as Table 1. Columns group named model series or providers; parentheses give subset counts. Others contains the remaining subsets. Each subset contains 25 images. Mean averages all 65 subsets equally; bold marks the best displayed value per column.

Method	Firefly (3)	Seedream (2)	FLUX (6)	OpenAI (8)	Google (3)	Midjourney (4)	SD (6)	Kling (4)	MAI (3)	Others (26)	Mean ↑
Frozen Qwen3-VL	22.67	20.00	32.00	36.00	52.00	46.00	59.33	43.00	34.67	60.15	48.06
UnivFD	64.00	80.00	74.00	58.00	89.33	82.00	82.67	87.00	85.33	80.46	77.66
Effort	72.00	76.00	60.00	54.00	90.67	72.00	62.67	72.00	77.33	78.62	71.69
B-Free[†]	52.00	44.00	44.00	63.50	32.00	48.00	41.33	73.00	12.00	41.85	45.66
DDA[†]	85.33	58.00	36.67	61.00	72.00	67.00	75.33	52.00	56.00	74.31	66.52
Deep-VRM[‡]	61.33	52.00	64.00	52.50	68.00	61.00	60.67	50.00	86.67	78.15	67.63
FRA	**90.67**	**88.00**	**91.33**	**76.50**	**96.00**	**83.00**	**93.33**	**91.00**	**94.67**	**93.23**	**90.15**

FRA improves average fake recall over the compared baselines in the displayed MNW groups. The gains extend across commercial providers and model series beyond Stable Diffusion. The OpenAI group remains more challenging than the other displayed groups, showing variation in transfer across generation systems even after readout alignment.

A.7 Readout sensitivity

Figure 5 shows stable performance across the tested carrier ridges. Fisher shrinkage works best near the default, while the scale estimated from training data lies on a performance plateau. Downweighting the forensic term reduces Mean, consistent with the direction ablation.

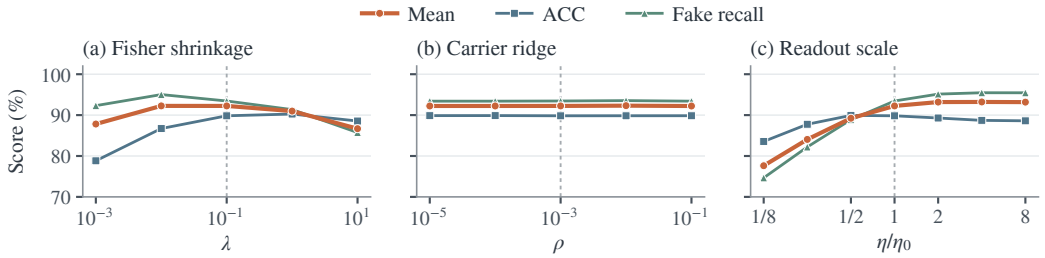

Figure 5: **Readout sensitivity.** One factor varies at a time, with the offset re-estimated on real training images. Mean averages the six benchmarks excluding Chameleon. Vertical dashed lines mark defaults; η_0 is the scale estimated from training data.

A.8 Robustness to image perturbations

We reuse the checkpoints from the main comparison under the unified SD1.4 training setting with 2,500 images, without retraining or threshold adjustment for perturbations. The fixed evaluation panel contains 256 records per benchmark, totaling 1,792 records from 1,791 unique images across all 149 subsets of the seven benchmarks. Sampling is stratified by subset and label, independently of model outputs. We apply Gaussian blur, JPEG compression, color-contrast changes, block noise, and Gaussian noise at five levels each.

On this panel, FRA achieves 90.33% clean accuracy and 85.68% mean accuracy across the 25 perturbed conditions, ranking first among all 13 evaluated methods and reference configurations. Figure 6 displays FRA and five published baselines from this comparison. The strongest published baseline, Deep-VRM, averages 75.32%, giving FRA a margin of 10.36 percentage points.

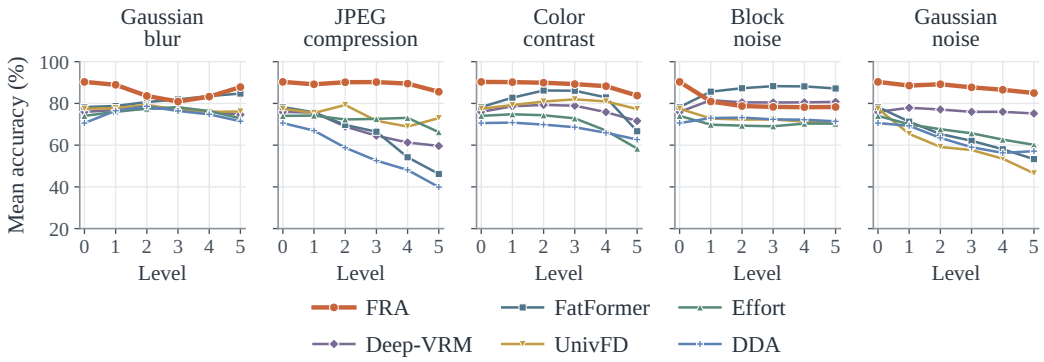

Figure 6: **Robustness to image perturbations.** Mean accuracy on a fixed seven-benchmark panel for FRA and five published baselines. Level zero denotes clean images.

Figure 6 shows that FRA maintains a clear advantage under JPEG compression and Gaussian noise, and remains comparatively stable under color-contrast changes. Block noise causes a larger drop, with FatFormer and Deep-VRM performing better in this setting. The aggregate advantage therefore reflects robustness across the evaluated perturbations, with residual sensitivity to localized corruption.

A.9 Full MLLM versus its vision encoder

Table 10: **Paired comparison of full MLLM and vision-only features.** Scores (%) with FRA and frozen feature extractors on seven shared benchmarks. Mean equally averages the displayed benchmarks. Bold marks the better result within each backbone pair, including ties.

Backbone	Features	AIGINow	DRCT2M	EvalGEN	GPTImage2	MNW	T2ICoRe Bench	Chameleon	Mean ↑
Qwen3-VL-8B	MLLM	92.94	86.72	**97.95**	**86.11**	**90.15**	**99.59**	**84.26**	**91.10**
	ViT	**93.11**	**88.60**	**97.95**	30.82	75.82	79.31	79.19	77.83
Qwen3.5-9B	MLLM	**95.53**	78.12	**99.55**	**77.40**	**86.83**	**99.50**	**82.10**	**88.43**
	ViT	93.72	**84.31**	**99.55**	24.07	83.88	83.76	76.45	77.96
InternVL3.5-8B	MLLM	**97.75**	**80.74**	**94.55**	**45.89**	52.00	**97.67**	**69.38**	**76.85**
	ViT	96.47	76.34	94.09	43.74	**60.31**	94.68	66.34	76.00
deepseek-vl2-small	MLLM	94.67	**85.78**	**98.41**	**79.35**	**79.08**	**99.21**	**79.12**	**87.95**
	ViT	**97.50**	84.86	95.45	22.50	64.49	95.79	70.09	75.81
Kimi-VL-A3B-Instruct	MLLM	91.47	83.07	91.36	**82.09**	68.00	94.70	**80.60**	**84.47**
	ViT	**93.11**	**84.79**	**99.32**	37.87	**85.97**	**97.15**	75.28	81.93
GLM-4.6V-Flash	MLLM	**96.00**	83.88	95.45	**71.23**	77.54	**97.73**	**75.89**	**85.39**
	ViT	82.42	**85.43**	**97.73**	31.80	**90.34**	79.36	70.44	76.79

To assess the contribution of the language-model stack, we compare FRA on each frozen MLLM's final representations with FRA on its frozen vision encoder alone. Table 10 pairs these settings within each backbone on the same seven benchmarks, with full-MLLM results from Table 3.

The full MLLM improves the shared-benchmark mean for every backbone and consistently improves detection on GPTImage2 and Chameleon. The gains vary across targets: vision-only features perform better on MNW for InternVL, Kimi, and GLM, and remain competitive on several other benchmarks. These paired results suggest that processing visual features through the language-model stack can make them more useful for detection, with benefits depending on the backbone and target distribution.

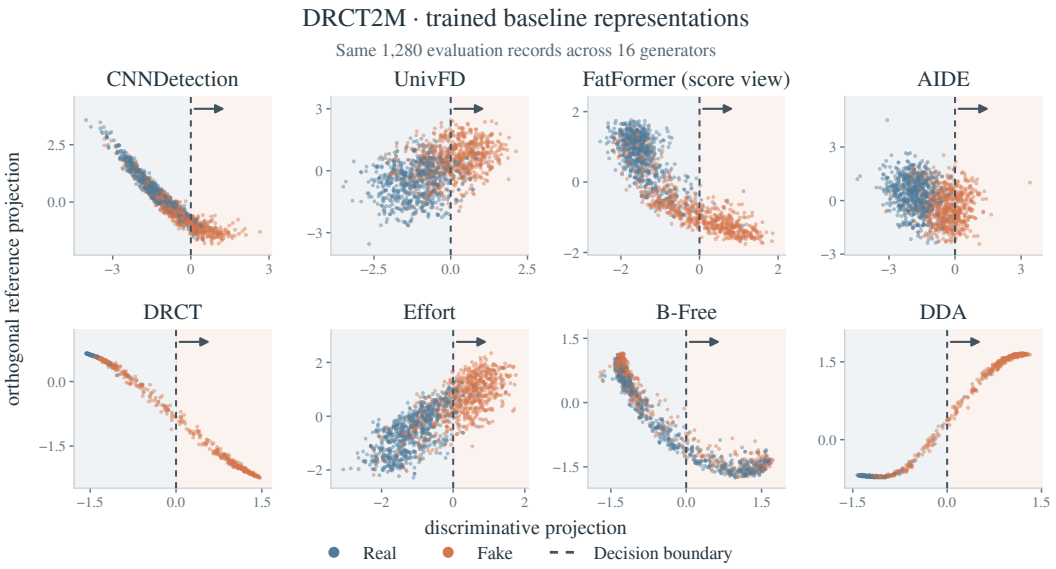

Figure 7: **Baseline detection geometry on DRCT2M.** For each trained baseline detector, we visualize its features in a two-dimensional score space. The horizontal axis is the standardized discriminative projection along the detector's decision direction, and the vertical axis is the projection onto an orthogonal reference direction. The direction of the arrow represents the direction judged as fake. Notably, FatFormer uses a score–principal-component view to accommodate its input-conditioned head. Dashed lines mark the inherited decision boundaries.

A.10 Baseline detection geometry on DRCT2M

Figure 7 compares eight detectors trained on the same 2,500 training images and evaluated on the same 1,280 test images spanning 16 DRCT2M generators. The horizontal axis shows each detector's standardized discriminative projection along its decision direction, and the vertical axis shows the projection onto an orthogonal reference direction. FatFormer uses a score–principal-component view to accommodate its input-conditioned decision head. Overall, CNNDetection, AIDE, and B-Free show substantial class overlap, while UnivFD, Effort, and DDA exhibit clearer separation but still place some fake samples on the real side of the inherited boundary. These views show that real/fake training does not guarantee class separation or well-placed decision boundaries on unseen generators. These observations, together with FRA's results in Table 1, suggest that sufficiently discriminative pretrained features can support competitive detection through hyperplane adaptation alone, which underscores the importance of both representation quality and boundary alignment.

A.11 Qualitative answer preservation

Figures 8–10 provide 15 qualitative comparisons on Qwen2.5-VL-7B, covering ordinary visual questions, object-level meanings of real and fake, and more demanding visual judgments. All four variants share this backbone and receive identical inputs. These examples complement the quantitative FRA and LoRA results on Qwen3-VL-8B in Section 4.5.

Visual question answering after forensic adaptation

Qwen2.5-VL-7B | Identical inputs and greedy decoding | Verbatim answers



Source: DRCT2M

Q: What brand name is printed on the motorcycle?

Frozen	The brand name "Honda" is printed on the motorcycle.
FRA (ours)	The brand name "Honda" is printed on the motorcycle.
LoRA	Honda
Deep-VRM	Honda is the brand name.



Source: DRCT2M

Q: Is the phone to the left or right of the laptop?

Frozen	The phone is to the right of the laptop.
FRA (ours)	The phone is to the right of the laptop.
LoRA	right of the laptop on the table.
Deep-VRM	right



Source: DRCT2M

Q: How many dogs are visible in the image?

Frozen	Two black dogs are visible in the grassy field.
FRA (ours)	Two black dogs are visible in the grassy field.
LoRA	Two dogs are playing with a red frisbee on grass.
Deep-VRM	Two dogs.



Source: MNW

Q: What is the man doing with the instrument?

Frozen	The man is playing an acoustic guitar while seated.
FRA (ours)	The man is playing an acoustic guitar while seated.
LoRA	The man is playing guitar.
Deep-VRM	guitar.



Source: MNW

Q: What color is the person's face mask?

Frozen	The person's face mask is blue.
FRA (ours)	The person's face mask is blue.
LoRA	blue
Deep-VRM	face mask is blue.



Source: DRCT2M

Q: According to the printed caption, what city and US state is the hotel in?

Frozen	The hotel is located in Davenport, Iowa.
FRA (ours)	The hotel is located in Davenport, Iowa.
LoRA	Iowa.
Deep-VRM	The hotel is in Davenport, Iowa.



Source: DRCT2M

Q: Does the seated foreground person wear a patterned headscarf or a plain hat?

Frozen	The seated foreground person wears a patterned headscarf, not a plain hat.
FRA (ours)	The seated foreground person wears a patterned headscarf, not a plain hat.
LoRA	The seated foreground person wears a patterned headscarf, not a plain hat.
Deep-VRM	a patterned headscarf.

Figure 8: **Ordinary visual question answering after forensic adaptation.** Frozen, FRA, LoRA, and Deep-VRM share Qwen2.5-VL-7B and receive identical inputs with greedy decoding. FRA reproduces the frozen model's answers in all seven examples. The motorcycle, desk, dogs, hotel, and headscarf images are from DRCT2M (Chen et al., 2024); the guitar and face mask images are from MNW (Roca et al., 2026).

Object semantics involving real / fake

Qwen2.5-VL-7B | Verbatim answers | Red: incorrect object judgment



Source: DRCT2M

Q: Is the cat standing on the car a real animal or a fake decorative cat? Give one clue.

Reference: Living tabby cat; not a decorative object.

Frozen	The cat's realistic fur and posture suggest it is a real animal, not a decorative one.
FRA (ours)	The cat's realistic fur and posture suggest it is a real animal, not a decorative one.
LoRA	The cat's fur appears slightly unnatural, suggesting it might be a decorative cat rather than a real one.
Deep-VRM	fake.



Source: DRCT2M

Q: Is the figure on top of the clock a real person or a fake human figure? What is it?

Reference: A statue above the clock; not a living person.

Frozen	The figure atop the clock is a statue, not a real person; it's a symbolic representation.
FRA (ours)	The figure atop the clock is a statue, not a real person; it's a symbolic representation.
LoRA	The figure atop the clock is a real real real human figure, likely a statue or sculpture.
Deep-VRM	real person, statue of lady holding lantern



Source: DRCT2M

Q: Is the sheep on the rock a real animal or a fake ornament? Give one visible clue.

Reference: A living sheep; not an ornament.

Frozen	The sheep appears to be real, as it has natural wool and stands on a rocky terrain.
FRA (ours)	The sheep appears to be real, as it has natural wool and stands on a rocky terrain.
LoRA	The sheep's wool appears too uniform and pristine, suggesting it might be a fake ornament rather than a real animal.
Deep-VRM	fake.



Source: DRCT2M

Q: Is this a real full-size train or a fake miniature train? Give one clue about its scale.

Reference: A miniature train layout with model workers.

Frozen	The presence of miniature workers and the detailed, small-scale environment suggest it's a fake miniature train.
FRA (ours)	The presence of miniature workers and the detailed, small-scale environment suggest it's a fake miniature train.
LoRA	real.
Deep-VRM	train.

Figure 9: **Object semantics of real and fake.** All four images are from DRCT2M (Chen et al., 2024). The Qwen2.5-VL-7B variants answer questions about living animals, a human statue, and a miniature train, testing the meanings of *real* and *fake* when describing objects. FRA preserves the frozen model's judgments and explanations in all four cases. Red text marks incorrect or incomplete object judgments from LoRA and Deep-VRM.

Non-forensic visual question answering: harder cases

Qwen2.5-VL-7B | Verbatim answers | Red: incorrect requested fact



Source: MNW

Q: How many people are visible, and how many are holding bowls?

Reference: Four people; three holding bowls

Frozen	Four people are visible, and three are holding bowls.
FRA (ours)	Four people are visible, and three are holding bowls.
LoRA	Three are holding bowls near the grill.
Deep-VRM	Five children; four are holding bowls.



Source: MNW

Q: Which of the speaker's own hands holds the sheet of paper?

Reference: Left hand

Frozen	The speaker holds the sheet of paper in his left hand.
FRA (ours)	The speaker holds the sheet of paper in his left hand.
LoRA	The speaker holds the sheet of paper with his right hand while gesturing with his right.
Deep-VRM	The speaker holds the sheet of paper with his right hand.



Source: DRCT2M

Q: Does the illuminated pedestrian signal show a walking figure or a raised hand?

Reference: Raised hand

Frozen	The illuminated pedestrian signal shows a raised hand, indicating "Do not walk."
FRA (ours)	The illuminated pedestrian signal shows a raised hand, indicating "Do not walk."
LoRA	The illuminated pedestrian signal shows a walking figure, indicating do not walk.
Deep-VRM	The illuminated pedestrian signal shows a walking figure.



Source: MNW

Q: How many numbered subpoints are listed in total under the two numbered sections on the projected slide?

Reference: Six

Frozen	There are six numbered subpoints listed under the two main sections on the projected slide.
FRA (ours)	There are six numbered subpoints listed under the two main sections on the projected slide.
LoRA	There are six numbered subpoints listed under the three main sections on the projected slide.
Deep-VRM	There are seven numbered subpoints listed in total under the two numbered sections.

Figure 10: **Answer preservation on more demanding visual questions.** Qwen2.5-VL-7B variants are compared on conditional counting, a hand–object relation, pedestrian signal recognition, and counting subpoints on a projected slide. The images of people holding bowls, a speaker, and a projected slide are from MNW (Roca et al., 2026); the pedestrian signal image is from DRCT2M (Chen et al., 2024). FRA matches the frozen model's correct responses in all four examples. Red text marks incorrect or incomplete responses from LoRA and Deep-VRM.