ICML

International Conference
On Machine Learning

Robust Vision-Language Models via Manifold-Adversarial Adapters

Hao Li, Zeyu Xiao, Junhao Zhou, Peng Liu,
Yang Zhao, Wei Jia

NUS
National University
of Singapore

Background

Modern VLMs have made remarkable progress on clean, high-quality visual inputs.

Yet in deployment, images often contain corruptions such as noise, blur, compression and low resolution.

Robustness benchmarks show that these degradations can turn visual understanding into wrong answers or hallucinated reasoning.

How can we make VLMs robust to such corruptions?

Question: What is the weather like?

Reference Image

LLM: Sunny.

Degraded Image

LLM: Rainy.

Ground Truth: Clear

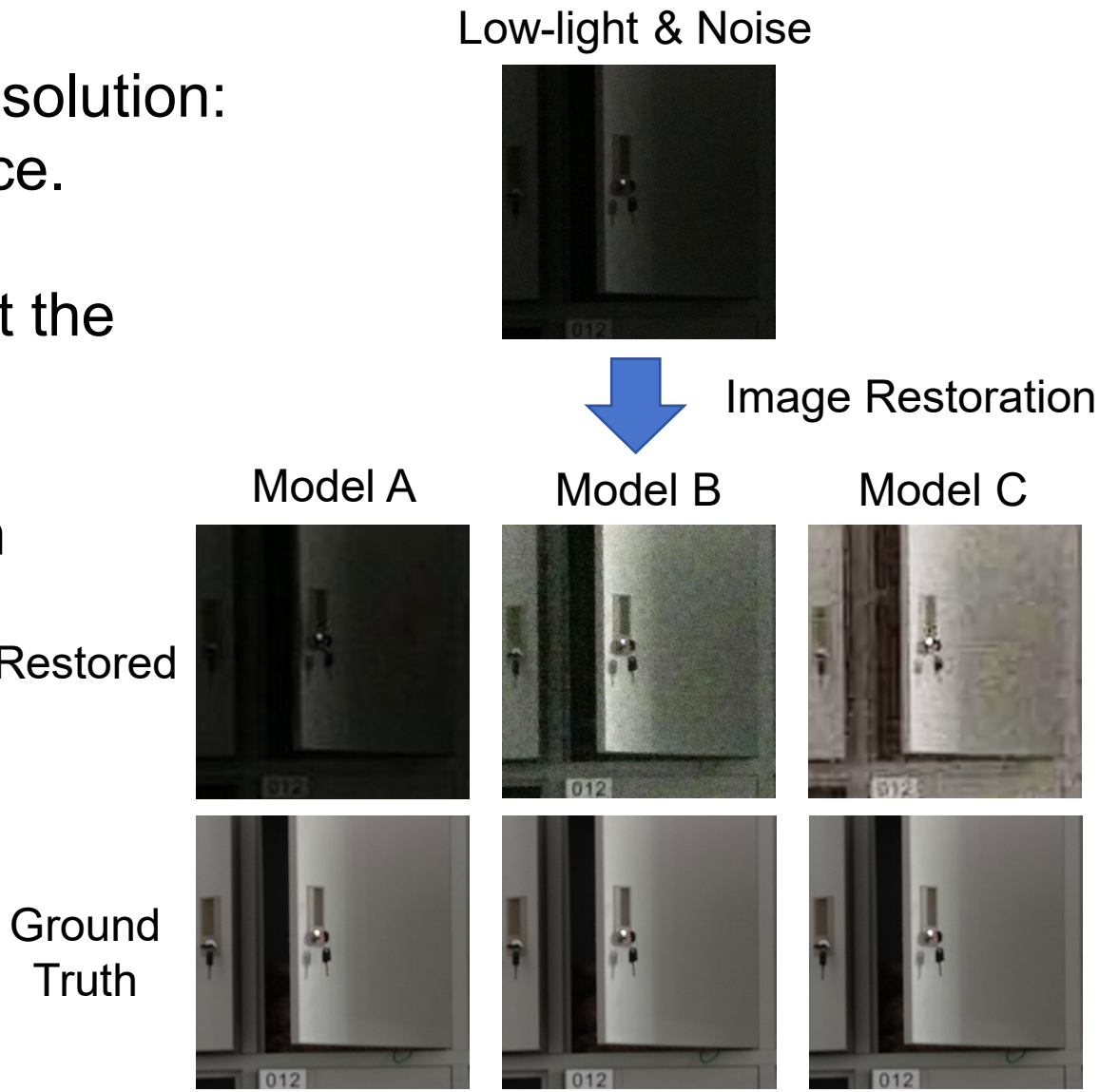
The obvious fix: restore the image first

Image restoration is a natural preprocessing solution: recover a cleaner image before VLM inference.

But IR models optimize pixel-level fidelity, not the semantic representations used by VLMs.

As shown on the right, different restorers can generate different textures, contrast, and artifacts from the same degraded input.

Cleaner pixels do not necessarily mean better VLM reasoning.



Restoration is costly and not consistently helpful

If image restoration were a reliable solution, it should improve corrupted-image performance without hurting general capability.

Preprocessing	R-Bench-Dis	MMVet	Latency
None, Base VLM	**58.79**	**42.33**	**213 ms**
AirNet	56.97 ↓1.82	40.23 ↓2.10	1024 ms 4.8× slower
Real-ESRGAN	57.17 ↓1.62	38.12 ↓4.21	985 ms 4.6× slower
RAM-PromptIR	57.37 ↓1.42	40.73 ↓1.60	1081 ms 5.1× slower
MoCE-IR	58.38 ↓0.41	40.96 ↓1.37	1013 ms 4.8× slower

Scores are higher-is-better. Red arrows show drops relative to the base VLM.

However, our results show the opposite trend: external IR models bring negative gains across VLM benchmarks.

Meanwhile, restoration turns inference into a two-stage pipeline, adding substantial latency before the VLM even starts reasoning.

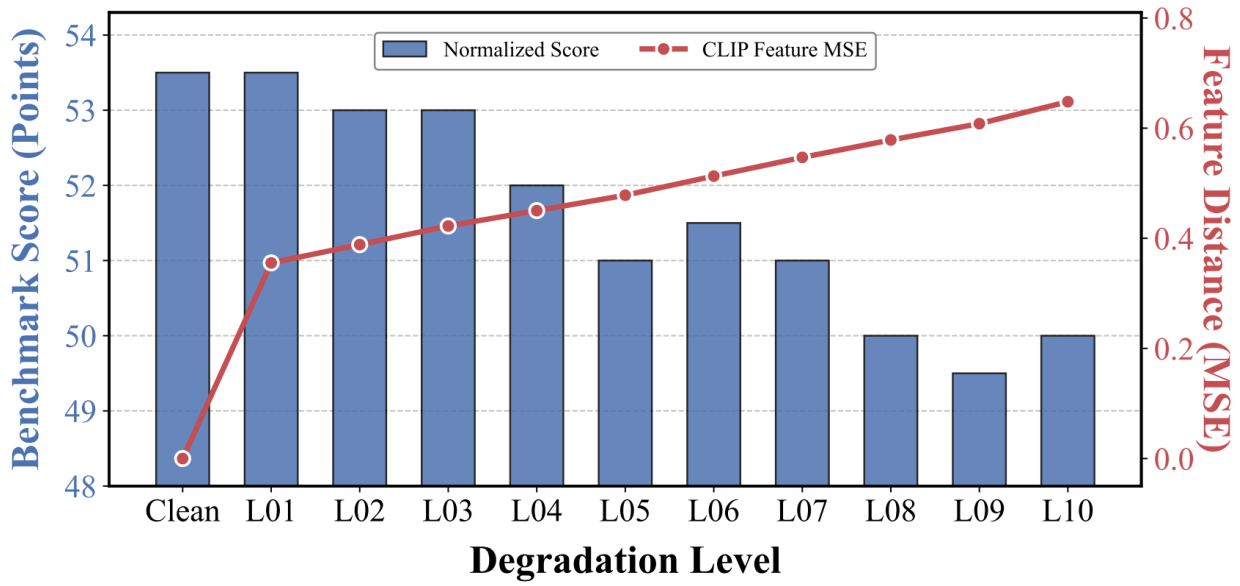
Image restoration adds computation, but does not guarantee better VLM performance.

From Pixel Restoration to Feature-Level Repair

ICML International Conference On Machine Learning

Instead of restoring pixels, we target the VLM's visual feature space.

On Flickr2K, we apply progressively stronger corruptions and measure
(i) MSE between corrupted and clean visual features.
(ii)downstream VLM performance.



The trend is clear: stronger corruptions lead to larger feature drift and lower task performance.

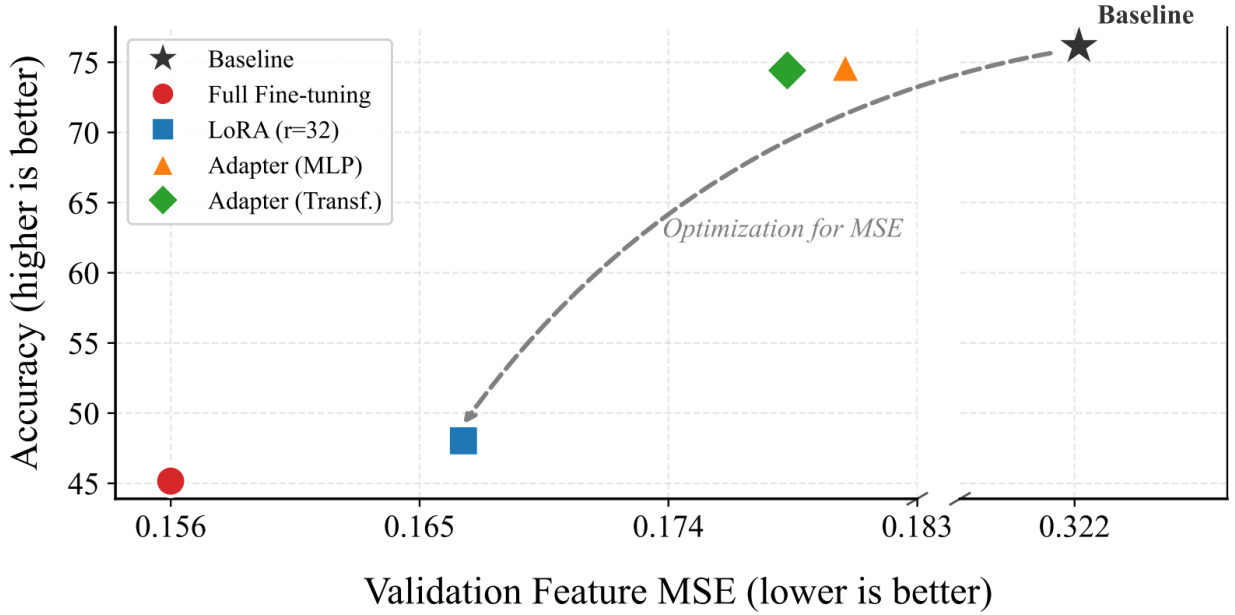
So, can we directly train a module to minimize feature MSE?

Does Feature MSE Alignment Solve the Problem?

ICML International Conference On Machine Learning

We train on 75K clean–degraded image pairs using feature MSE as the only objective.

We test several adaptation strategies: full fine-tuning, LoRA, MLP adapters, and Transformer adapters.



Although these methods reduce feature MSE, they consistently hurt downstream VLM performance.

In particular, invasive updates such as full fine-tuning and LoRA can severely damage the pretrained visual representation.
Lower feature distance does not guarantee better reasoning.

A Paradox: Lower Feature MSE, Worse Reasoning

Earlier, we observed a clear correlation:

stronger corruptions → larger feature MSE → lower VLM performance.

So we tried the most direct intervention:

minimize feature MSE between corrupted and clean images.

However, after optimization, feature MSE decreases while downstream performance also drops.

Why does the same metric give opposite intuition?

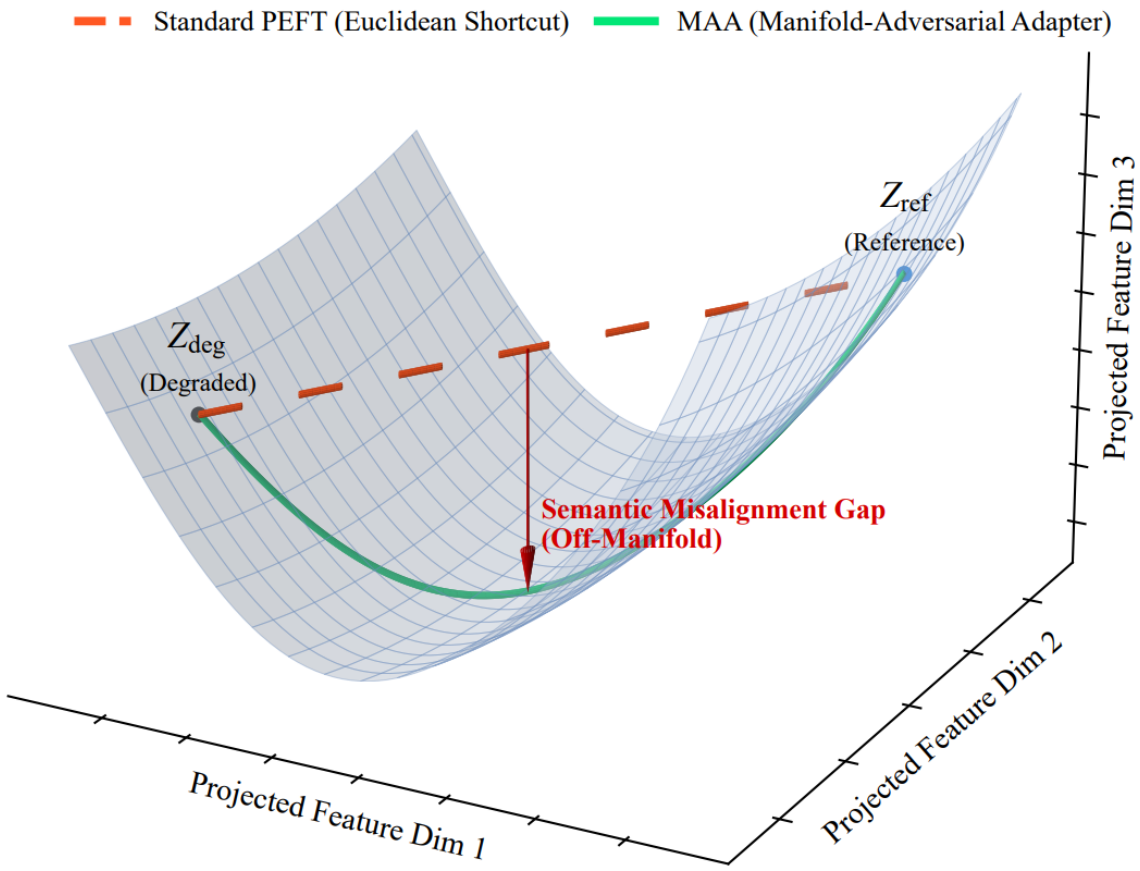
What Does MSE Ignore?

ICML
International Conference
On Machine Learning

To explain this paradox, we look beyond pairwise feature distance.

A pretrained VLM does not operate on arbitrary visual features; its reasoning is calibrated to a particular distribution of visual tokens.

MSE only pulls each corrupted feature toward its clean reference in Euclidean space.



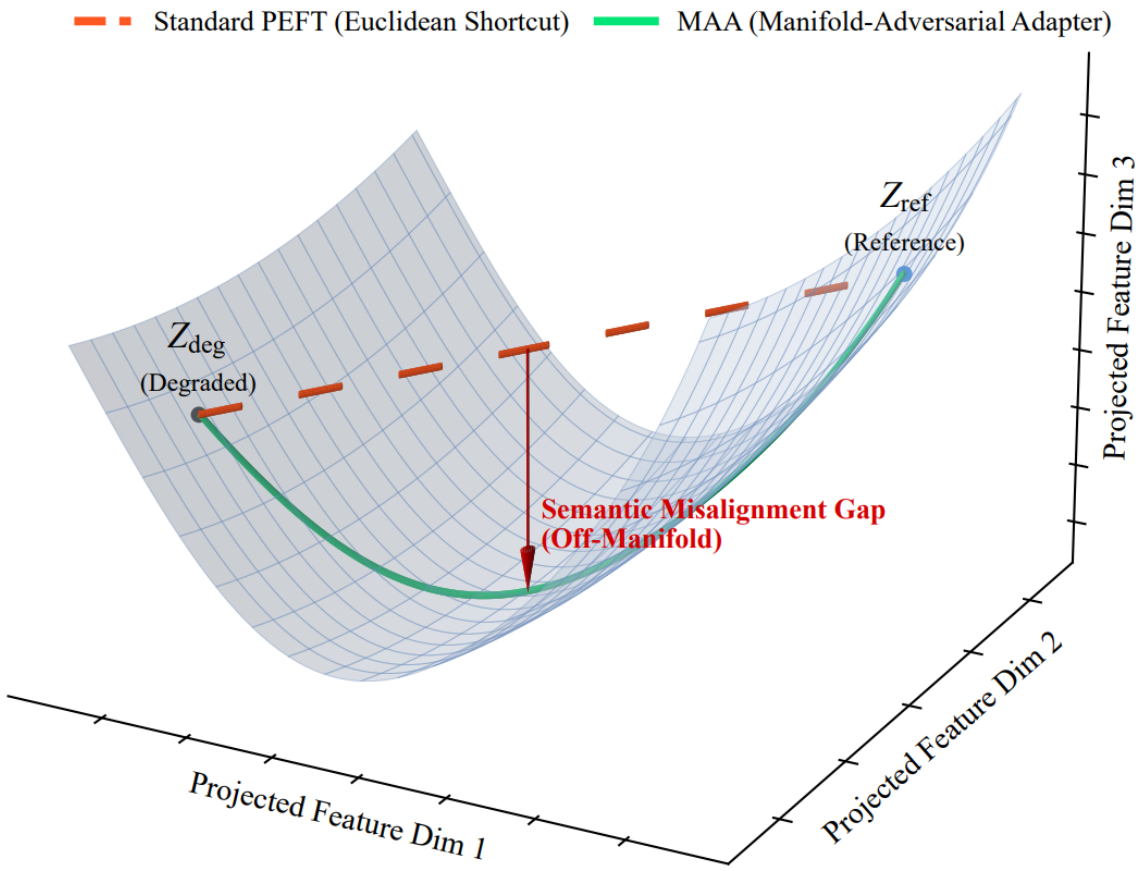
It does not ensure that the corrected feature remains in a semantically valid region of the feature space.

Semantic Misalignment Gap

We call this failure the Semantic Misalignment Gap.

MSE-only adaptation can reduce feature distance while pushing representations away from the in-distribution semantic support.

As a result, the corrected feature becomes geometrically closer but less compatible with VLM reasoning.



The goal is not just to reduce feature distance, but to restore features on the semantic manifold.

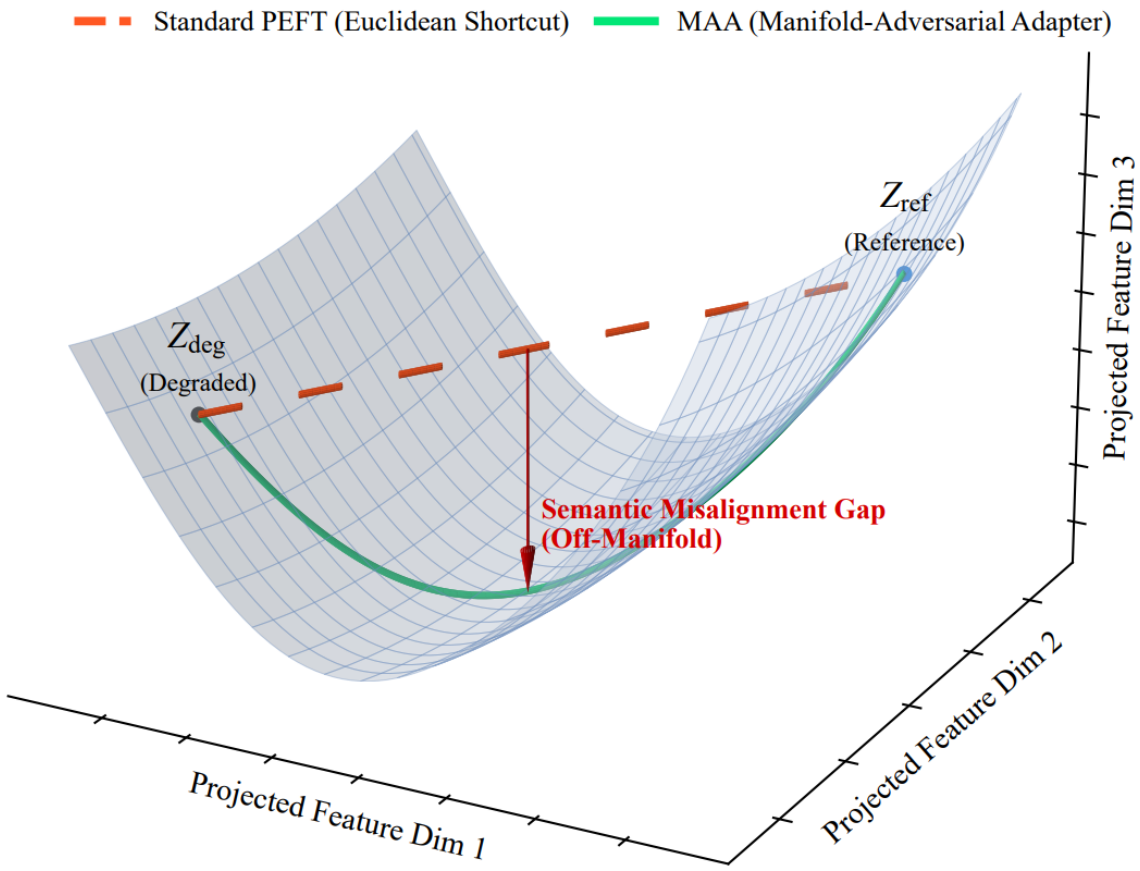
How to Avoid Off-Manifold Feature Correction?

ICML International Conference On Machine Learning

To prevent corrected features from drifting into low-density, semantically invalid regions, we introduce an adversarial manifold constraint.

A discriminator is trained to distinguish:
(i) Clean teacher features as real
(ii) Corrected corrupted features as fake

The adapter is then optimized to make corrected features indistinguishable from clean features.
This turns feature repair from point-wise matching into distribution-aware alignment.

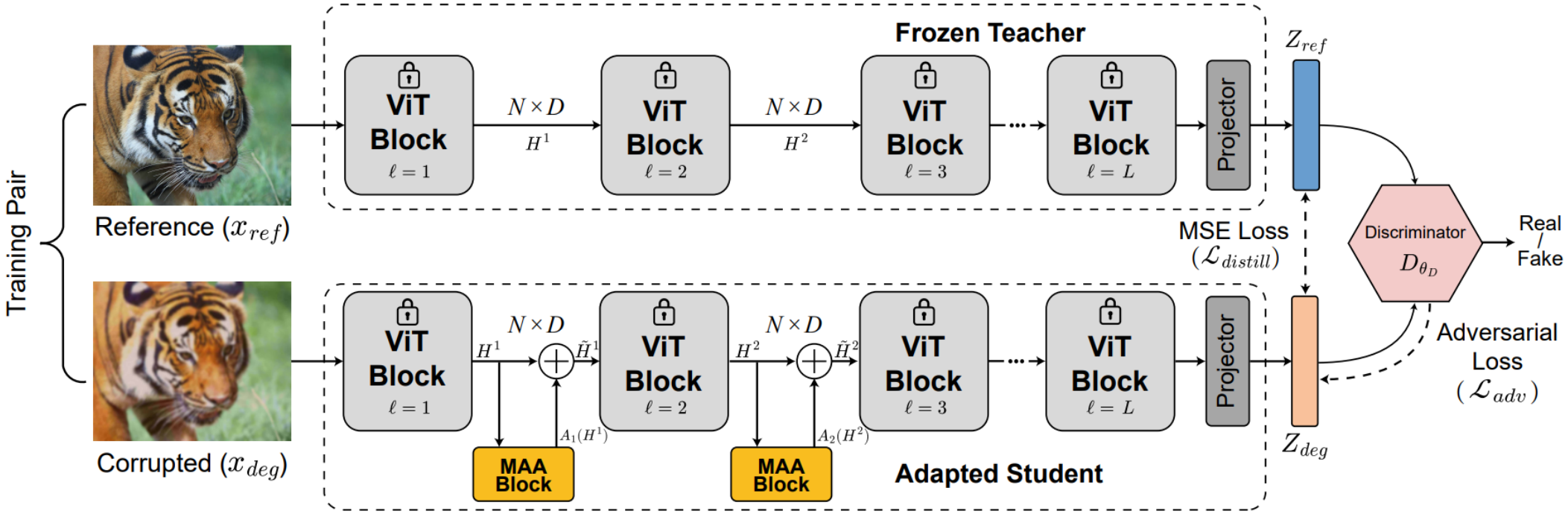


Manifold-Adversarial Adapter

We insert lightweight adapters into each transformer block of the frozen vision encoder.
Each adapter applies a small residual correction to the intermediate visual tokens.
The original VLM backbone and projector remain frozen; only the adapters are trained.
Training combines two objectives:

$$\min_{\theta_A} \mathcal{L}_{distill} + \lambda \mathcal{L}_{adv} \ ,$$

At inference time, the discriminator is discarded, and only the adapters are retained.



Three Complementary Correction Paths

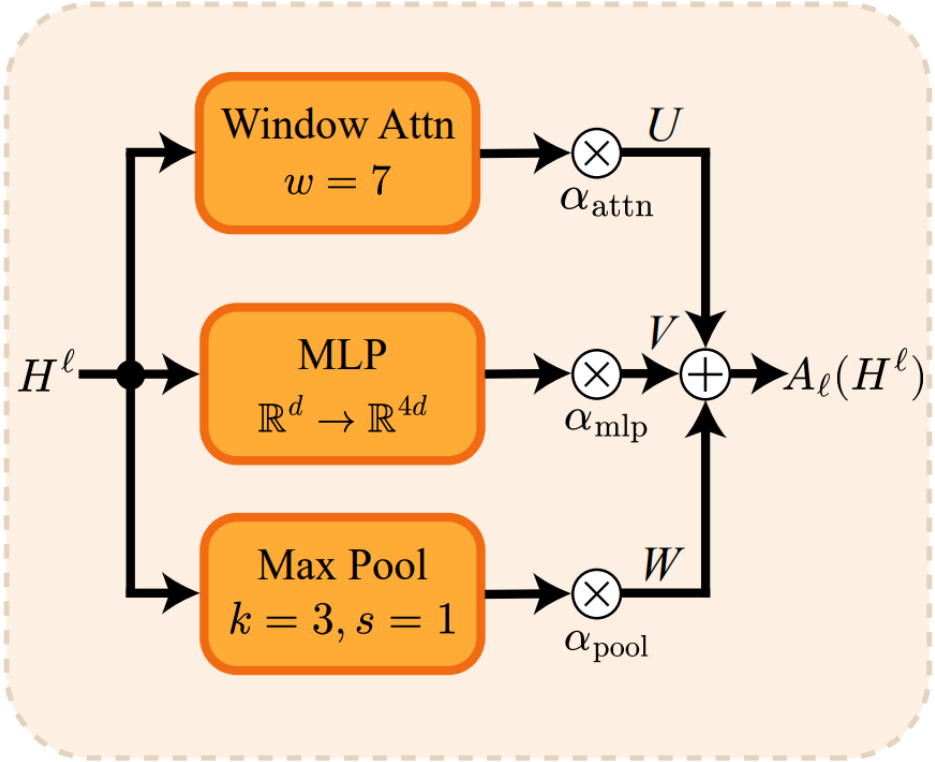
Real-world corruptions are spatially non-uniform and statistically diverse.

MAA combines:

Window Attention for local structural repair

MLP for feature-statistic recalibration

Max Pooling for robust local aggregation

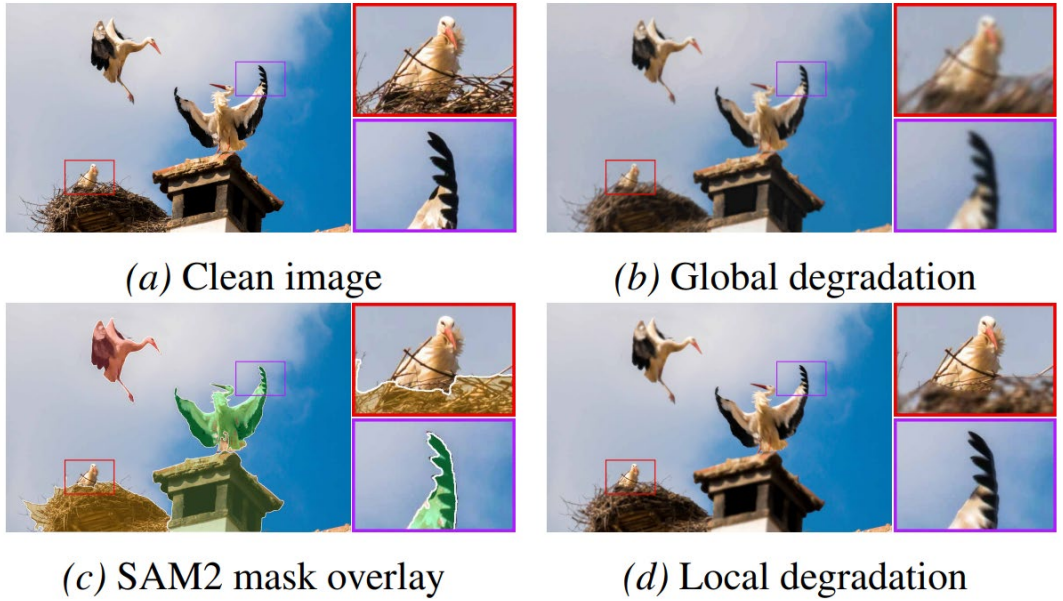


Gated residual branches start from identity, enabling conservative corrections.

Training with Realistic and Diverse Corruptions

We train MAA on paired clean–degraded images from multiple sources.

The dataset combines high-quality images, web images, and OCR-heavy images to improve generalization.



(a) Clean image *(b)* Global degradation

(c) SAM2 mask overlay *(d)* Local degradation

Beyond global corruptions, we use SAM2 masks to apply degradations locally to semantic regions.

This produces spatially heterogeneous artifacts that better match real-world image corruptions.

MAA Improves Robustness

ICML International Conference On Machine Learning

Method	R-Bench-Dis	R-Bench-Ref	MMBench	MMVet	Hallusion
Base VLM	58.79	58.91	76.14	42.33	50.47
MSE-only Adaptation	57.98 ↓0.81	58.30 ↓0.61	74.63 ↓1.51	42.02 ↓0.31	50.36 ↓0.11
+ Real-ESRGAN	57.17 ↓1.62	58.70 ↓0.21	75.45 ↓0.69	38.12 ↓4.21	50.26 ↓0.21
+ MoCE-IR	58.38 ↓0.41	59.31 ↑0.40	76.00 ↓0.14	40.96 ↓1.37	50.79 ↑0.32
MAA (Ours)	**59.39** ↑0.60	**59.92** ↑1.01	**76.30** ↑0.16	**42.61** ↑0.28	**51.31** ↑0.84

We compare MAA against the base VLM, MSE-only adaptation, and restoration-based preprocessing.

MAA achieves the best R-Bench-Dis score, improving 58.79 → 59.39.

It also preserves clean/reference and general benchmark performance, unlike naive adaptation or external restoration.

Single-Stage Robustness with Minimal Overhead

External restoration requires a two-stage pipeline: restore the image first, then run VLM inference.

This increases latency to nearly 1 second per image, about 5× slower than the base VLM.

Method	GFLOPs	Latency	Speedup	Tokens
Base (LLaVA-1.6)	34,594	213	1.00×	4.91
+ Real-ESRGAN	57,865	985	0.22×	4.87
+ AirNet	43,067	1,024	0.21×	5.14
+ RAM-PromptIR	38,599	1,081	0.20×	4.91
+ MoCE-IR	63,956	1,013	0.21×	4.89
MAA (Ours)	**34,765**	**230**	**0.93**×	5.03

MAA is integrated into the vision encoder, adding only 17 ms over the base model.

At inference, no restoration model or discriminator is needed.

Ablation: Adversarial Alignment Matters

MSE-only adaptation underperforms the base VLM, confirming that feature matching alone is not sufficient.

Adding adversarial alignment already improves the MLP-only adapter to 59.19 / 60.12 on R-Bench.

Removing the adversarial loss from full MAA drops both robustness and reference performance.

The full model achieves the best overall corrupted-image score.

Method	Components			R-Bench	
	Adv	W-Attn	Pool	Dis	Ref
Base (Zero-shot)	-	-	-	58.79	58.91
Naive Adaptation	×	×	×	57.98	58.30
MAA (MLP-Only)	✓	×	×	59.19	60.12
MAA w/o $\mathcal{L}_{adv}$	×	✓	✓	58.70	57.37
MAA w/o W-Attn	✓	×	✓	58.79	**60.53**
MAA w/o Pooling	✓	✓	×	**59.39**	59.11
MAA (Full)	✓	✓	✓	**59.39**	59.92

Mechanism Verification

Naive adaptation reduces paired feature distance, but worsens distribution-level alignment with clean features.

Method	MMD^2	KID	Energy	CKA
Base (Corrupted)	0.0248	0.0062	0.2777	0.8252
Naive Adaptation	0.0506	0.0083	0.4823	0.8092
MAA (Ours)	**0.0235**	**0.0059**	**0.2683**	**0.8467**

This matches the Semantic Misalignment Gap: closer point-wise, but farther from the clean feature distribution.

MAA reduces MMD^2, KID, and Energy distance, while improving CKA similarity.

This supports that MAA improves robustness by keeping corrected features distributionally compatible with clean visual tokens.

ICML

International Conference
On Machine Learning

Thank You