BREAKING THE HISTORICAL ACTION SHORTCUT: MITIGATING VISUAL MODALITY COLLAPSE IN GAMING VLA MODELS VIA HISTORICAL ACTION NOISE INJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interleaved multimodal inputs—streams of multi-frame images alternating with historical action tokens—have become a primary data format for Vision-Language-Action (VLA) models used in Embodied AI and gaming agents. Under Supervised Fine-Tuning (SFT), we identify a failure mode of this paradigm: because expert gameplay produces action sequences with strong temporal correlation and low conditional entropy, the autoregressive transformer tends to route gradient updates through the low-entropy historical pathway, which can starve the high-dimensional visual encoder—a phenomenon we term *visual modality collapse*. Unlike the classical copycat/inertia shortcut in unimodal behavioral cloning, this failure is (i) *conditional*—it persists even after perfect marginal class balancing because $P(A_t \mid A_{1:t-1}) \to 1$ within any trajectory, and (ii) *cross-modal*—both modalities are read by the same causal transformer, so the shortcut manifests as gradient starvation of the visual pathway rather than a data-imbalance artifact. We provide a first-principles account of this collapse under standard simplifying assumptions from information-bottleneck theory and gradient dynamics, and introduce two directly measurable diagnostics—*Modality Information Gain* (IG) and *KL Divergence Shift*—that isolate which modality actually drives predictions. Building on this analysis, we propose *Historical Action Noise Injection* (HANI): with probability p, historical action tokens are replaced by uniformly sampled tokens from the discrete action vocabulary, keeping the ground-truth target intact. This discrete, semantically mismatched signal is harder to recover from than continuous noise, which drives up the historical pathway's prediction error and routes larger corrective gradients back onto the visual encoder. On the extended MiniWorld WallGap benchmark with Qwen3-VL-8B-Instruct as backbone, HANI outperforms Standard SFT, NEFTune, and Modality Dropout on task success and modality diagnostics (IG_{action}: $0.3104 \to 0.0159$; IG_{vision}: $0.1121 \to 0.2624$). Additional experiments on three visually richer 3D environments—ViZDoom (Deadly Corridor), Minecraft (MineRL Navigate), and RoboTHOR (ObjectNav)—suggest that the shortcut is not tied to a low-texture testbed, and that HANI's relative improvement is largest in the environment (Minecraft) whose expert data most closely matches the low-entropy regime our analysis targets. HANI is a single-stage, one-line data-loader intervention with zero inference overhead.

1 INTRODUCTION

As the reasoning capabilities of Large Language Models (LLMs) have improved, Vision-Language-Action (VLA) models have become a common approach for building Embodied AI and gaming agents (Sapkota et al., 2025; Kim et al., 2024; Intelligence et al., 2025; Qu et al., 2025; Raad et al., 2024; Magne et al., 2026; Zhou et al., 2026). In complex 3D gaming environments such as Minecraft (Li et al., 2025; Wang et al., 2025) and First-Person Shooters (FPS) (Yue et al., 2026; Zhou et al., 2026), agents typically face Partially Observable Markov Decision Processes (POMDPs) (Li et al., 2025). To address these challenges, previous research has widely adopted a multimodal input paradigm highly aligned with human prior decision-making logic: weaving together multi-frame continuous

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.

images (visual states) and historical action sequences (short-term intent and physical inertia) to autoregressively predict the optimal action for the current frame (Sapkota et al., 2025; Li et al., 2025; Chen et al., 2025; Tan et al., 2025). Intuitively, vision provides immediate environmental feedback, while historical actions help maintain the coherence of long-term tasks. Fusing the two appears to be a reasonable design choice for building capable agents.

However, when performing Supervised Fine-Tuning (SFT) on VLA models with expert gaming data, this multimodal design encounters underlying optimization challenges (Baker et al., 2022; Chen et al., 2025; Wang et al., 2025). Analyses of gaming data distributions report that human player operations exhibit strong temporal correlation and low action-token entropy (Kang et al., 2013; Dubey et al., 2018). For example, in tasks such as long-distance running or continuous shooting, a player may input the same *Forward* or *Fire* command for hundreds of consecutive frames (Liu et al., 2024). As a result, the current frame's action is, with high probability, a repetition of the previous frame's action.

This low-entropy distribution creates favourable conditions for shortcut learning in the optimization dynamics of neural networks (De Haan et al., 2019; Wen et al., 2020; Kim et al., 2026; Yue et al., 2026). Gradient descent tends to follow the path of least resistance (Chi et al., 2025). In the multimodal inputs of VLAs, historical actions are low-dimensional discrete text tokens with minimal semantic extraction costs. Conversely, visual frames are high-dimensional, complex 3D projected pixel arrays, from which extracting effective features like "obstacles" or "enemies" incurs an enormous computational burden. When copying the historical action tokens is sufficient to minimize the cross-entropy loss on most of the training data, the optimizer tends to settle at this local optimum. The prediction error drops, and the gradients backpropagated to the visual encoder become small—a regime known as "Gradient Starvation" (Pezeshki et al., 2021). We refer to the resulting behaviour as *Visual Modality Collapse*: the model comes to rely predominantly on historical actions for inference, and the visual features contribute little to the final prediction.

Why this is a structurally new problem, not a re-branding of copycat/inertia. Action inertia and copycat behavior have been documented for years in behavioral cloning (Wen et al., 2020; De Haan et al., 2019; Bica et al., 2021; Wen et al., 2021). Our claim is not that we discovered the *symptom* but that, inside a modern interleaved VLA transformer, the same symptom is produced by a mechanism that is different in kind and requires a different remedy. First, in classical imitation learning the shortcut is a *marginal-distribution* problem ($P(A_t)$ is skewed and can, in principle, be fixed by reweighting/resampling). In interleaved VLA training the failure is a *conditional* one: even after perfect class balancing, $P(A_t \mid A_{1:t-1}) \approx 1$ within any trajectory, so the historical pathway keeps predicting with near-zero error regardless of *global* balance. Second, in a VLA both modalities are projected into the same embedding space and read by the same causal transformer, so the shortcut is *cross-modal*: the optimizer preferentially updates the low-entropy pathway, producing gradient starvation of the visual encoder—a phenomenon with no counterpart in unimodal MLP/CNN behavioral cloning.

The consequence is a drop in generalization: when a sudden visual event (an enemy around a corner, an obstacle ahead) requires breaking inertia, an agent that relies on the historical pathway tends to keep replaying its previous action. To restore visual grounding without discarding short-term intent, we propose **Historical Action Noise Injection (HANI)**: with probability p, historical action tokens are replaced by uniformly sampled discrete tokens (e.g., *Forward*→*Fire*) while the ground-truth target is left intact. Because the mismatch is discrete and semantically legal, it is not easily removed by the model's continuous denoising capabilities, and larger corrective gradients propagate back to the visual encoder. **Contributions.** (i) We characterize *visual modality collapse* as a cross-modal and conditional failure mode of interleaved VLA SFT, and introduce two modality-level diagnostics—*Modality Information Gain* and *KL Divergence Shift*—that help identify which modality drives predictions. (ii) Under standard simplifying assumptions, we give a stylized theoretical account of why the collapse is likely to occur, from both information-bottleneck and gradient-dynamics perspectives, and show that HANI weakens the identified mechanism. (iii) On MiniWorld WallGap (2–5 rooms) with Qwen3-VL-8B-Instruct, plus ViZDoom, Minecraft MineRL Navigate, and RoboTHOR ObjectNav, HANI improves over Standard SFT, NEFTune, and Modality Dropout on task success and modality diagnostics, and its relative gain is largest in the environment (Minecraft) whose expert data most closely matches the low-entropy regime our analysis targets.

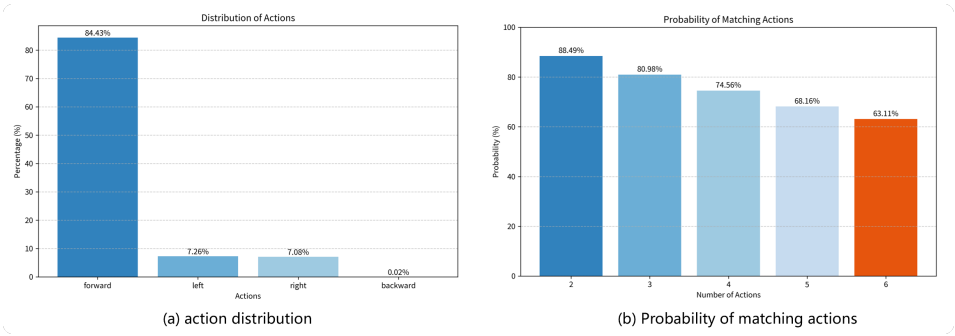
2 THE PROBLEM OF HISTORICAL ACTION SHORTCUT IN GAME VLAS

Interleaved multimodal inputs are a common design choice in recent VLA training pipelines (Sapkota et al., 2025; Li et al., 2025; Chen et al., 2025; Tan et al., 2025). Under standard SFT, however, models converge to high training accuracy but often respond weakly to unexpected visual events at test time—a discrepancy we now analyze.

2.1 PHENOMENON OBSERVATION: SFT-TRAINED MODELS EXHIBIT PERCEPTUAL BLINDNESS AND INERTIAL INFERENCE

Action inertia (high temporal correlation) is a common characteristic of human expert demonstrations in continuous control and sequential decision-making tasks, from Minecraft (VPT (Baker et al., 2022)) to autonomous driving (De Haan et al., 2019) and FPS games (Yue et al., 2026), where operators frequently hold the same key for many consecutive frames. To study this as an *optimization pathology* rather than an environment-specific artifact, we adopt MiniWorld WallGap (Chevalier-Boisvert et al., 2023) first-person navigation as a *diagnostic* testbed (illustrated in Appendix C, Figure 4): confounders such as inventory, hierarchical planning, sparse rewards, and dynamic action spaces are stripped out, leaving only the visual-vs-historical modality bottleneck. Section 4.2.3 later reports evidence that the shortcut and the HANI intervention also apply in three visually richer 3D environments. A qualitative failure case—an SFT agent that keeps predicting *Forward* while stuck against a wall—is shown in Appendix C, Figure 5.

Statistics of the expert trajectories (Figure 1) reveal the two symptoms our theory targets: (a) severe class imbalance—*Forward* dominates at 84.43% while *Backward* is only 0.02%; (b) extreme temporal correlation—88.49% of frames repeat the previous action, and with probability 63.11% the same action holds for six or more consecutive frames.



(a) action distribution (b) Probability of matching actions

Figure 1: Statistical analysis of expert trajectory data demonstrating the "Historical Action Shortcut" conditions. (a) Action distribution exhibiting severe class imbalance, with *Forward* movement dominating at 84.43%. (b) The probability distribution of consecutive identical actions (temporal correlation). There is an 88.49% probability that the current action perfectly mirrors the previous frame, and a >63% chance it mirrors the past five frames, providing a deceptive low-entropy shortcut for autoregressive models.

Since copying the previous action alone already yields near-90% training accuracy, the network has weak incentive to invest in costly visual parsing—we refer to this failure mode as the *Historical Action Shortcut*.

2.2 QUANTITATIVE DIAGNOSIS: MEASURING MODALITY COLLAPSE

Standard accuracy and cross-entropy loss cannot reveal *which* modality drives a prediction. We therefore introduce two directly measurable diagnostics (formal definitions in Appendix D). Let $I_{1:t}$, $A_{1:t-1}$, A_t denote the image history, action history, and target action.

Modality Information Gain (IG). The expected log-likelihood gain of A_t contributed by each modality on top of the other, e.g. $IG_{vision} = \mathbb{E}[\log P(A_t \mid I_{1:t}, A_{1:t-1}) - \log P(A_t \mid A_{1:t-1})]$. A collapsed model shows $IG_{vision} \ll IG_{action}$.

KL Divergence Shift. $D_{KL}(P_{full} \parallel P_{vision})$ vs. $D_{KL}(P_{full} \parallel P_{action})$: whichever divergence is smaller identifies the modality that *alone* already reproduces the full model's predictive distribution, i.e. the modality actually in control.

Empirically on MiniWorld (Base vs. SFT, full table in Appendix D), standard SFT reduces IG_{vision} from 0.4087 to 0.1121 while IG_{action} remains at 0.3104, and $D_{KL}(P_{full} \parallel P_{action})$ decreases from 1.92 to 0.11—consistent with a model whose predictions are driven predominantly by historical text rather than by vision.

2.3 Theoretical Analysis: The Inevitability of Collapse

We now provide a first-principles account of why powerful deep networks are so susceptible to this behavior. We stress upfront that our analysis relies on standard simplifying assumptions—Fano's inequality for the information-theoretic bound and a generalized-linear multimodal head for the gradient analysis—and is intended as a *stylized*, directional account of the collapse rather than a complete characterization of non-convex optimization in 8B autoregressive VLAs. Our empirical diagnostics (Modality IG, KL Shift) then track these predictions directly; the theory would be falsified had the metrics moved the other way. We trace the inevitability to two foundational dimensions: Information Bottleneck Theory and optimization dynamics.

Information-theoretic view. By the chain rule $I(Y; V, H) = I(Y; H) + I(Y; V \mid H) \leq H(Y)$, so $I(Y; V \mid H) \leq H(Y \mid H)$. Fano's inequality then bounds $H(Y \mid H) \leq H_b(\delta) + \delta \log(|\mathcal{A}| - 1)$, where δ is the error probability of the best history-only predictor. When temporal correlation is high, δ becomes small, so the information budget available to the visual pathway is tightly bounded and the marginal utility of the visual encoder for reducing the loss is limited (full derivation in Appendix A.1).

Gradient-dynamics view. Historical action tokens are 1-D discrete features whose semantics the LLM already models well; visual patches are high-dimensional and require deeper encoding to become predictive. When the optimizer reaches the text shortcut, the per-frame prediction error, and hence the gradient flowing to θ_v, becomes small ($\nabla_{\theta_v}\mathcal{L} \approx 0$), a regime we refer to as *Gradient Starvation* of the visual pathway (Appendix A.3). The two views make consistent directional predictions: interventions that (i) enlarge $H(Y \mid H)$ and (ii) keep $|\nabla_{\theta_v}\mathcal{L}|$ non-vanishing would be expected to raise IG_{vision} and, insofar as the visual encoder can be trained further, task success; interventions that reshape only the marginal $P(A_t)$ would not be expected to change IG_{vision} by much. Table 2 is consistent with this contrast between HANI and class-balancing baselines.

In detail, our analysis makes three testable predictions: (a) IG_{vision} should rise when the historical channel is attenuated, (b) IG_{action} should decrease with the semantic mismatch, and (c) task success should track IG_{vision} more closely than IG_{action}. These predictions are supported by our measurements: HANI reduces IG_{action} from 0.31 to 0.016, raises IG_{vision} from 0.11 to 0.26, and yields the largest task-success gain in our comparison, while continuous-space regularization (NEFTune), which does not appreciably change either IG value, yields only a small gain over SFT.

3 Methodology

3.1 Overview of HANI

The central issue is that historical actions offer a low-cost predictive pathway that tends to dominate optimization in standard SFT. HANI perturbs this pathway at the data-loader level. For each training sample we draw $Z \sim \mathrm{Bernoulli}(p)$: if $Z = 1$, every token in the historical action sequence $H = \{A_1, \ldots, A_{t-1}\}$ is independently replaced by a token uniformly sampled from $\mathcal{V}_A \backslash \{A_i\}$ to yield a mismatched $\tilde{H}$; otherwise $H' \leftarrow H$. The ground-truth target $Y = A_t$ and the visual sequence V are *unchanged*. The model then predicts $\hat{P} = \mathcal{M}_\theta(V, H')$ and is trained with the standard cross-entropy loss $\mathcal{L}_{CE} = -\log \hat{P}(Y)$. Under the mismatched branch, H' carries no information about Y, so reducing the loss requires updating the visual pathway—an effect we refer to as "gradient awakening". The full pseudocode is given in Algorithm 1 (Appendix E); HANI adds one line at the data loader and no inference overhead.

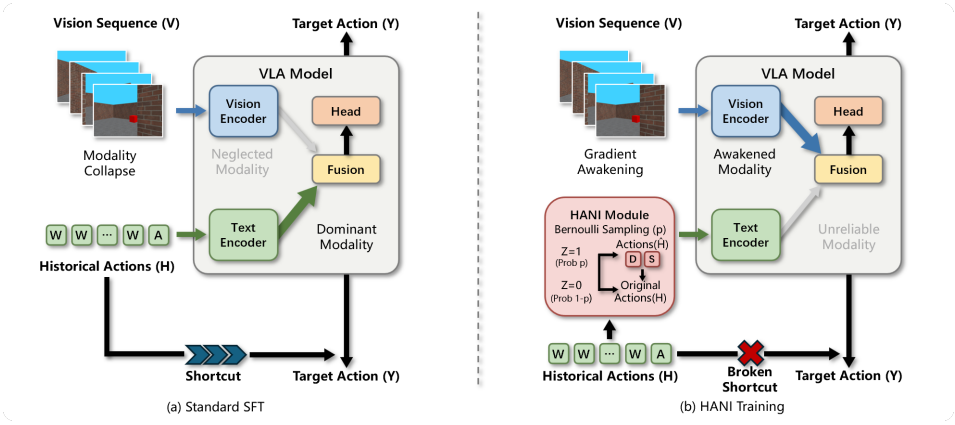


Figure 2: (a) Standard SFT feeds real images and historical actions, causing the model to fall into the shortcut trap. (b) HANI injects discrete semantic mismatches during training, blocking the shortcut and forcing the visual encoder to dominate decision-making.

3.2 Core Motivation and Design Rationale

HANI is contrasted with three natural alternatives. *(a) Fully masking history* would remove the shortcut but also remove short-term intent, and we observe erratic behaviour in this regime (see the $p = 1.0$ ablation). *(b) Continuous regularizers*—Modality Dropout (Korse et al., 2024) and NEFTune (Jain et al., 2023)—either introduce train/inference mismatch (dropout zeroes a channel the deployed model always receives) or add continuous noise that self-attention is empirically able to compensate for, leaving the semantic shortcut largely intact. *(c) Data-side rebalancing*—loss-decay on repeated actions (Chen et al., 2025; Wang et al., 2025), focal / class-balanced loss, and keyframe filtering (Wen et al., 2021)—primarily reshapes the *marginal* $P(A_t)$; the within-trajectory conditional $P(A_t \mid A_{1:t-1})$ that our analysis identifies as the driver of gradient starvation is not directly targeted (see Section 5 and Appendix F). HANI, by contrast, introduces a *discrete, semantically mismatched* substitution that (i) is not easily removed by continuous denoising, (ii) preserves the token structure so the inference-time distribution is not shifted, and (iii) retains the true history on a $(1 - p)$ fraction of samples to preserve intention coherence.

Practical footprint. HANI adds no trainable parameters, no auxiliary networks, and no additional forward or backward pass. The overhead per training step is one $\mathrm{Bernoulli}(p)$ draw plus at most $|H|$ uniform samples from $\mathcal{V}_A$, which is small relative to an 8B autoregressive forward pass; inference is unchanged because the deployed model consumes real history. In our runs, wall-clock per-epoch time is within $\pm 1\%$ of Standard SFT.

3.3 Theoretical Guarantee of HANI

We mathematically verify HANI's ability to mitigate visual collapse from two rigorous dimensions (Detailed proofs in Appendices A.2 and A.3):

Information-Theoretic Perspective: HANI attenuates the information carried by the historical pathway:

$$I(Y; H') \leq (1 - p)\, I(Y; H) \tag{1}$$

To reduce the overall loss, the model has a stronger incentive to draw predictive information from the visual modality. The conditional mutual information available through the mixture becomes:

$$I(Y; V \mid H') = (1 - p) \underbrace{I(Y; V \mid H)}_{\approx 0} + p \cdot I(Y; V), \tag{2}$$

so an increasing fraction of the useful conditional information (scaled by p) can only be obtained by mapping visual features V to Y.

Optimization Dynamics Perspective: Under standard SFT the historical shortcut yields near-zero prediction errors, so gradient signals to the visual encoder become small. Under HANI, the

mismatched history lowers $P(Y)$ substantially, and the per-token error term $|E_c|$ increases toward its upper bound of 1 (which holds regardless of the semantic distance between the injected token and the target; see Appendix A.3). Because H' carries no information about Y in the mismatched branch, this error is more effectively reduced by updating the visual encoder θ_v than by updating the text pathway, and standard training hygiene (AdamW, cosine LR, gradient clipping, BF16) keeps the resulting optimization stable in our runs.

4 EXPERIMENTS

We evaluate HANI along four axes: (i) task-success *effectiveness*, (ii) whether the gain comes from the *mechanism* we posit (a shift in modality reliance), (iii) *sensitivity* to the perturbation probability p, and (iv) *generality* beyond the MiniWorld diagnostic testbed to visually richer 3D environments.

4.1 EXPERIMENTAL SETUP

Environment. We use MiniWorld *WallGap* variants (Chevalier-Boisvert et al., 2023) with 2–5 randomly generated rooms; gaps, wall colors, and obstacle positions are re-randomized on reset, forcing visually driven decisions rather than inertial replay (illustrated in Appendix B, Figure 6). **Data & model.** 1,000 expert trajectories (≈60k frames); backbone Qwen3-VL-8B-Instruct (Team, 2025). **Baselines.** (i) Base (no SFT), (ii) Standard SFT, (iii) NEFTune (Gaussian noise on text embeddings), (iv) Modality Dropout (zeroing historical text vectors). Full training/evaluation details are in Appendix B.

4.2 EXPERIMENTAL RESULTS AND ANALYSIS

4.2.1 EFFECTIVENESS AND MECHANISM VERIFICATION

Table 1: Task success rates (%) across MiniWorld environments of varying complexities.

Method	2-Room	3-Room	4-Room	5-Room
Base	36.72 ± 1.53	10.20 ± 0.62	3.09 ± 0.23	0.16 ± 0.11
SFT	96.48 ± 1.94	56.48 ± 1.69	42.58 ± 1.53	11.04 ± 0.96
NEFTune	97.46 ± 1.24	58.09 ± 1.19	46.95 ± 0.99	11.86 ± 1.12
Modality Dropout	89.17 ± 1.50	54.77 ± 1.31	39.45 ± 0.77	10.98 ± 0.50
HANI ($p = 0.6$)	$\mathbf{98.20 \pm 1.21}$	$\mathbf{66.41 \pm 1.03}$	$\mathbf{49.47 \pm 1.14}$	$\mathbf{15.04 \pm 0.78}$

Table 1 reports zero-shot generalization. Standard SFT reaches 96.48% in the 2-room setting but drops to 11.04% in the 5-room setting, indicating that reliance on historical shortcuts degrades sharply as visual complexity increases. HANI improves over the baselines at all four difficulty levels, with absolute gains over Standard SFT of roughly 10 percentage points in the 3- and 4-room settings.

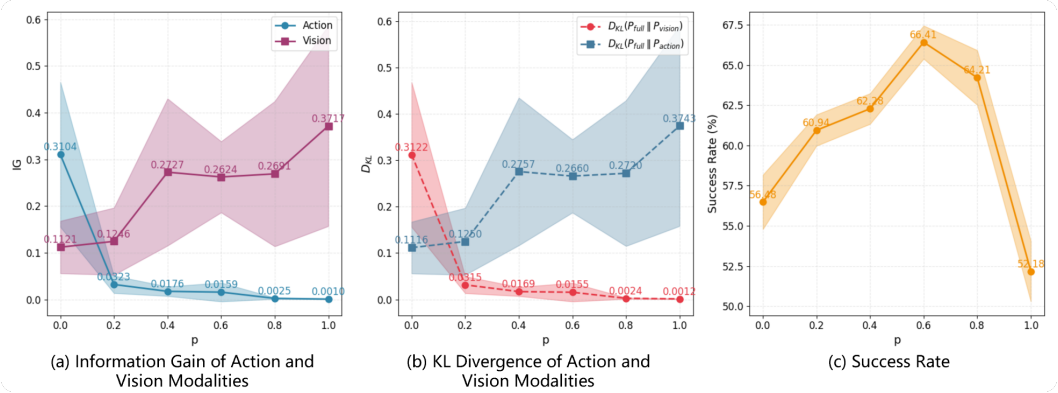
Table 2: Modality Information Gain and KL Divergence Shift across methods in the 3-Room Setting.

Method	$\mathbf{IG_{action}}$	$\mathbf{IG_{vision}}$	$\mathbf{D_{KL}(P_{full} \parallel P_{vision})}$	$\mathbf{D_{KL}(P_{full} \parallel P_{action})}$
Base	0.4328 ± 0.0055	0.4087 ± 0.0012	0.3382 ± 0.0025	1.9234 ± 0.0103
SFT	0.3104 ± 0.1552	0.1121 ± 0.0561	0.3122 ± 0.1561	0.1116 ± 0.0558
NEFTune	0.3648 ± 0.0673	0.2171 ± 0.1681	0.3639 ± 0.0668	0.2172 ± 0.1683
Modality Dropout	0.1693 ± 0.1009	0.1872 ± 0.0336	0.1680 ± 0.1007	0.1843 ± 0.0330
HANI ($p = 0.6$)	$\mathbf{0.0159 \pm 0.0202}$	$\mathbf{0.2624 \pm 0.0761}$	$\mathbf{0.0155 \pm 0.0201}$	$\mathbf{0.2660 \pm 0.0789}$

Table 2 sheds light on the mechanism behind these gains. Under Standard SFT the diagnostics indicate that predictions are dominated by the historical pathway ($IG_{action} = 0.3104 > IG_{vision} = 0.1121$). Under HANI, IG_{action} decreases to 0.0159 while IG_{vision} rises to 0.2624; $D_{KL}(P_{full} \parallel P_{vision})$ drops to 0.0155, which is consistent with a predictive distribution that is now more closely aligned with the visual pathway. Notably, HANI reverses the ordering of IG_{vision} and IG_{action} by more than an order of magnitude, which is the qualitative direction predicted by our cross-modal analysis but not the one predicted by marginal-distribution rebalancing.

The relative benefit is larger in more challenging settings. In the 5-Room configuration, where visual complexity is highest and SFT reaches only 11.04%, HANI reaches 15.04%—a $+36\%$ relative improvement over SFT and a $+27\%$ relative improvement over the strongest regularization baseline (NEFTune, 11.86%). NEFTune and Modality Dropout do not appreciably change IG_{action} (0.3648 and 0.1693 respectively, compared with 0.3104 under SFT) and yield smaller improvements in this setting; this pattern is consistent with the view that these interventions do not directly target the conditional information imbalance between the two pathways.

4.2.2 Ablation Study on Perturbation Probability p



(a) Information Gain of Action and Vision Modalities (b) KL Divergence of Action and Vision Modalities (c) Success Rate

Figure 3: Ablation study on the perturbation probability p in HANI in the 3-Room Setting.

Figure 3 shows the effect of p on the three quantities. Along the IG axes, IG_{action} decreases approximately monotonically with p (from 0.31 under SFT to 0.016 at $p = 0.6$ and to near-zero at $p = 1.0$), while IG_{vision} increases with p (up to 0.3717 at $p = 1.0$). Task success, however, is non-monotonic: it rises from 56.48% (SFT) to 66.41% at $p=0.6$, and then falls to 52.18% at $p=1.0$. A plausible interpretation is that at $p = 1.0$ the model never observes genuine history and cannot maintain multi-frame intent, so some fraction of preserved history appears to be helpful for coherence; empirically, $p \in [0.4, 0.7]$ offers a favourable trade-off in our setting.

4.2.3 Generality across Visually Richer 3D Environments

To examine whether the shortcut and HANI's effect are specific to a low-texture testbed, we run HANI with the *same* Qwen3-VL-8B-Instruct backbone and the *same* training recipe (only p is tuned) on three visually and semantically richer 3D environments: **ViZDoom Deadly Corridor** (a classic 3D FPS with dynamic enemies), **Minecraft MineRL Navigate** (an open 3D voxel world where human demonstrations are dominated by long streaks of the "W" Forward key), and **RoboTHOR ObjectNav** (a photorealistic embodied indoor navigation benchmark). Full setup details—seeds, tokenization schemes, per-episode variance, and hyperparameters—are given in Appendix B.

Table 3: Task reward comparison of Standard SFT vs. HANI on three visually richer 3D environments, using the same Qwen3-VL-8B-Instruct backbone and identical training recipe.

Environment	**SFT**	**HANI**	**Abs. Δ**	**Rel. Δ**
ViZDoom — Deadly Corridor (3D FPS)	1.199	**1.345**	+0.146	+12.2%
Minecraft — MineRL Navigate (open 3D world)	2.5219	**5.1498**	+2.6279	**+104.2%**
RoboTHOR — ObjectNav (photorealistic Embodied AI)	0.30	**0.33**	+0.03	+10.0%

Standard SFT exhibits the same qualitative failure mode in all three settings: "forward+fire" inertia in ViZDoom, "hold-W" stalls in Minecraft, and turn-inertia in RoboTHOR. Two observations: **(i)** HANI yields improvements across FPS textures, voxel terrain, and photorealistic indoor scenes, suggesting that the fix is not tied to low-texture environments; **(ii)** the largest relative improvement is on Minecraft

($+104.2\%$), whose expert data most closely matches the low-entropy regime our Theorem 2 analysis (Appendix A.2) targets, and smaller relative improvements are observed on ViZDoom ($+12.2\%$) and RoboTHOR ($+10.0\%$). This ordering is consistent with a mechanism that specifically targets the historical-action shortcut, though we caution that only three environments are compared and effect sizes are subject to per-environment reward-scale differences (see Section 6).

Portability. Only p was re-tuned across the three environments (all other hyperparameters, tokenization, and the 8B backbone were held fixed), and the selected value fell in the same $[0.4, 0.7]$ band as on MiniWorld. This is limited evidence but suggests that the p-selection heuristic may transfer across settings that share the discrete action vocabulary and the interleaved data schema.

5 Related Work

Shortcut Learning in Traditional Imitation Learning: Shortcut learning is a well-documented challenge in traditional imitation learning, particularly when addressing Partially Observable Markov Decision Processes (POMDPs). When behavioral cloning (BC) agents are provided with observation histories, they frequently succumb to the "copycat problem" (Wen et al., 2020) or "inertia problem" (Bica et al., 2021). Due to the strong temporal correlation among expert actions in continuous tasks, the imitator exploits this temporal shortcut, learning to predict current actions purely as a function of previous actions while ignoring instantaneous visual observations (De Haan et al., 2019). To alleviate this, recent approaches such as keyframe-focused visual imitation learning (Wen et al., 2021) directly target this temporal correlation. They demonstrate that upweighting "changepoint" keyframes—samples where the expert breaks action inertia (e.g., accelerating when a traffic light turns green)—forces the policy to attend to visual cues rather than historical shortcuts. While these sample-reweighting and keyframe-extraction methods provide valuable insights for traditional BC, they often rely on explicit heuristic scoring pipelines (e.g., training an auxiliary copycat model to compute action prediction errors). Such multi-stage heuristics are difficult to scale and integrate into the unified, interleaved text-image autoregressive generation paradigm of modern VLA models.

Shortcut Learning in VLA Models: In the context of large VLA models, the shortcut problem typically manifests as a severe class imbalance or modality collapse. Existing solutions generally fall into action space balancing and multimodal feature regularization. For action balancing, models trained on game data (e.g., OpenAI's VPT (Baker et al., 2022)) utilize heuristic down-sampling to discard repetitive frames, while CombatVLA (Chen et al., 2025) and Game-TARS (Wang et al., 2025) introduce loss-decay mechanisms to suppress gradients for repetitive actions. However, these methods treat the issue primarily as class imbalance rather than inter-modal information parasitism. For feature regularization, NEFTune (Jain et al., 2023) injects continuous Gaussian noise into embeddings, and Modality Dropout (Korse et al., 2024) randomly zeroes out modality tensors. As our analysis and experiments suggest, continuous-space noise is largely absorbed by the denoising capabilities of modern LLMs, and Modality Dropout can induce train/inference input-distribution mismatch. In contrast, HANI addresses the shortcut via discrete token-level substitutions. Unlike dropout/masking, which zeroes a signal that is present at inference, HANI substitutes it with a legal but incorrect token drawn from the *same* vocabulary the deployed model consumes, so the inference-time input distribution is preserved; unlike continuous NEFTune-style noise, the substitution is discrete and is not straightforwardly removed by self-attention.

A detailed contrast against four data-side rebalancing families (repeated-action downsampling, balanced batch sampling, keyframe reweighting (Wen et al., 2021), and focal / class-balanced loss)—showing why each addresses only the *marginal* action distribution and leaves the within-trajectory conditional shortcut $P(A_t \mid A_{1:t-1})$ untouched—is deferred to Appendix G.

6 Limitations and Future Work

Our claims are scoped to **gaming/embodied VLAs with discrete (or tokenized) action vocabularies**, where humans emit near-zero-noise identical key-press streams that trigger the theoretical low-entropy regime; physical-robot regimes with sensor jitter and continuous joints reach this regime less frequently. HANI extends naturally to tokenized-continuous VLAs by uniform token replacement, and to native continuous heads via *Adversarial Trajectory Mirroring* (injecting reversed/extremal history vectors so that $(\hat{y} - y)^2$ is maximized while y is unchanged); structure-aware noise from a

semantic tree and block-wise perturbations for very long histories are additional natural refinements. Full discussion of these extensions is deferred to Appendix H.

Threats to validity. Our theoretical account rests on two idealizations—Fano's inequality (tight only when Y is close to a deterministic function of H) and a generalized-linear multimodal head (whereas the 8B backbone is deeply non-linear)—so the analysis is directional rather than a full characterization of non-convex optimization; we mitigate this by empirically checking that Modality IG / KL Shift move in the theory-predicted direction across all four MiniWorld difficulty levels and three additional 3D environments. Second, our largest experiment fine-tunes a single 8B backbone (Qwen3-VL-8B-Instruct); scaling laws for HANI at 30B+ and across backbone families remain to be established. Third, task-reward scales differ across the three new 3D benchmarks (Table 3), so we report both absolute and relative deltas and rely on the consistent sign of the effect rather than any single number.

7 Conclusion

We characterized the *Historical Action Shortcut* as a cross-modal, conditional failure mode of interleaved VLA SFT that can lead to visual modality collapse, gave stylized information-theoretic and gradient-dynamics arguments for why this failure mode is likely to arise (under standard simplifying assumptions), and introduced two modality-level diagnostics—Modality IG and KL Shift—that measure which pathway drives predictions. Building on this analysis, HANI is a one-line data-loader modification with no inference overhead that improves over Standard SFT, NEFTune, and Modality Dropout on MiniWorld WallGap, and yields improvements on an 8B VLA backbone across ViZDoom, Minecraft MineRL Navigate, and RoboTHOR ObjectNav; the largest relative gain occurs in the environment (Minecraft) whose data most closely matches the low-entropy regime our analysis targets.

REFERENCES

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. Invariant causal imitation learning for generalizable policies. *Advances in Neural Information Processing Systems*, 34:3952–3964, 2021.

Peng Chen, Pi Bu, Yingyao Wang, Xinyi Wang, Ziming Wang, Jie Guo, Yingxiu Zhao, Qi Zhu, Jun Song, Siran Yang, et al. Combatvla: An efficient vision-language-action model for combat tasks in 3d action role-playing games. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10919–10928, 2025.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Ziheng Chi, Yifan Hou, Chenxi Pang, Shaobo Cui, Mubashara Akhtar, and Mrinmaya Sachan. Chimera: Diagnosing shortcut learning in visual-language understanding. *arXiv preprint arXiv:2509.22437*, 2025.

Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *Advances in neural information processing systems*, 32, 2019.

Rachit Dubey, Pulkit Agrawal, Deepak Pathak, Thomas L Griffiths, and Alexei A Efros. Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*, 2018.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*, 2023.

Ah Reum Kang, Jiyoung Woo, Juyong Park, and Huy Kang Kim. Online game bot detection based on party-play log analysis. *Computers & Mathematics with Applications*, 65(9):1384–1395, 2013.

Junyoung Kim, Woojoo Kim, Jaehyung Lim, Dongha Kim, and Hwanjo Yu. Vlm2rec: Resolving modality collapse in vision-language model embedders for multimodal sequential recommendation. *arXiv preprint arXiv:2603.17450*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model, 2024. *URL https://arxiv. org/abs/2406.09246*, 1(2):4, 2024.

Srikanth Korse, Mohamed Elminshawi, Emanuël AP Habets, and Srikanth Raj Chetupalli. Training strategies for modality dropout resilient multi-modal target speaker extraction. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 595–599. IEEE, 2024.

Muyao Li, Zihao Wang, Kaichen He, Xiaojian Ma, and Yitao Liang. Jarvis-vla: Post-training large-scale vision language models to play visual games with keyboards and mouse,(2025). *arXiv preprint arXiv:2503.16365*, 2025.

Jiahua Liu, Zeyuan Cui, Wenhan Ge, and Pengxiang Zhan. Dmtg: A human-like mouse trajectory generation bot based on entropy-controlled diffusion networks. *arXiv preprint arXiv:2410.18233*, 2024.

Loïc Magne, Anas Awadalla, Guanzhi Wang, Yinzhen Xu, Joshua Belofsky, Fengyuan Hu, Joohwan Kim, Ludwig Schmidt, Georgia Gkioxari, Jan Kautz, et al. Nitrogen: An open foundation model for generalist gaming agents. *arXiv preprint arXiv:2601.02427*, 2026.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield, Gavin Buttimore, Max Cant, Sarah Chakera, et al. Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*, 2024.

Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

Weihao Tan, Xiangyang Li, Yunhao Fang, Heyuan Yao, Shi Yan, Hao Luo, Tenglong Ao, Huihui Li, Hongbin Ren, Bairen Yi, et al. Lumine: An open recipe for building generalist agents in 3d open worlds. *arXiv preprint arXiv:2511.08892*, 2025.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zihao Wang, Xujing Li, Yining Ye, Junjie Fang, Haoming Wang, Longxiang Liu, Shihao Liang, Junting Lu, Zhiyong Wu, Jiazhan Feng, et al. Game-tars: Pretrained foundation models for scalable generalist multimodal game agents. *arXiv preprint arXiv:2510.23691*, 2025.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. *Advances in Neural Information Processing Systems*, 33:2564–2575, 2020.

Chuan Wen, Jierui Lin, Jianing Qian, Yang Gao, and Dinesh Jayaraman. Keyframe-focused visual imitation learning. *arXiv preprint arXiv:2106.06452*, 2021.

Yuguang Yue, Irakli Salia, Samuel Hunt, Chris Green, Wenzhe Shi, and Jonathan J Hunt. Scaling behavior cloning improves causal reasoning: An open model for real-time video game playing. *arXiv preprint arXiv:2601.04575*, 2026.

Zheyuan Zhou, Liang Du, Zixun Sun, Xiaoyu Zhou, Ruimin Ye, Qihao Chen, Yinda Chen, and Lemiao Qiu. Main-vla: Modeling abstraction of intention and environment for vision-language-action models. *arXiv preprint arXiv:2602.02212*, 2026.

A Appendix: Detailed Formula Derivations

This appendix establishes a rigorous theoretical framework to mathematically prove the inevitability of visual modality collapse in SFT, and the effectiveness of the HANI mechanism, under the standard simplifying assumptions stated in Section 2.3.

Definitions:

- V: Visual image sequence (high-dimensional continuous).
- H: Real historical action sequence (low-dimensional discrete).
- Y: Target predicted action.

SFT minimizes the conditional entropy $H(Y \mid V, H)$. Because the intrinsic entropy $H(Y)$ is constant, minimizing conditional entropy is equivalent to maximizing the joint mutual information $I(Y; V, H) = H(Y) - H(Y \mid V, H)$.

A.1 Proof: The Inevitability of Visual Modality Collapse in SFT

To mathematically formalize the "shortcut" phenomenon, we represent the strong temporal correlation in gaming data as the predictability of the target action $Y \in \mathcal{A}$ given the historical sequence H. Let the optimal deterministic predictor of Y given H have an error probability $p_e = P(\hat{Y}(H) \neq Y)$. Due to the severe action inertia in gaming (e.g., repeating the "Forward" command), p_e is empirically very small, bounded by some $\delta > 0$ such that $p_e \leq \delta$.

Theorem 1. *Assume the target action Y drawn from a discrete vocabulary $\mathcal{A}$ can be predicted from historical actions H with an error probability bounded by δ. In standard SFT, the conditional mutual information $I(Y; V \mid H)$ extracted from the visual modality is strictly upper-bounded by a vanishing function of δ, forcing modality collapse as $\delta \to 0$.*

Proof. According to Fano's Inequality, the conditional entropy of Y given H is bounded by the probability of error p_e:

$$H(Y \mid H) \leq H_b(p_e) + p_e \log(|\mathcal{A}| - 1) \tag{3}$$

where $H_b(\cdot)$ is the binary entropy function. Since $p_e \leq \delta$ for small δ, we can define an upper bound function $\epsilon(\delta)$:

$$H(Y \mid H) \leq \epsilon(\delta) \triangleq H_b(\delta) + \delta \log(|\mathcal{A}| - 1) \tag{4}$$

Note that as the history becomes a perfect predictor ($\delta \to 0$), the bound $\epsilon(\delta) \to 0$.

Now, consider the joint mutual information $I(Y; V, H)$ optimized during SFT. By the definition of mutual information:

$$I(Y; V, H) = H(Y) - H(Y \mid V, H) \tag{5}$$

Since entropy is non-negative ($H(Y \mid V, H) \geq 0$), the total extracted information is strictly bounded by the intrinsic entropy of the target:

$$I(Y; V, H) \leq H(Y) \tag{6}$$

By the chain rule of mutual information, we can also decompose the joint information as:

$$I(Y; V, H) = I(Y; H) + I(Y; V \mid H) \tag{7}$$

Expanding the historical information term $I(Y; H) = H(Y) - H(Y \mid H)$ and substituting it into Eq. (7) yields:

$$I(Y; V, H) = H(Y) - H(Y \mid H) + I(Y; V \mid H) \tag{8}$$

Applying the upper bound from Eq. (6) to Eq. (8):

$$\begin{aligned} H(Y) - H(Y \mid H) + I(Y; V \mid H) &\leq H(Y) \\ I(Y; V \mid H) &\leq H(Y \mid H) \end{aligned}$$

Finally, applying the Fano bound from Eq. (4):

$$0 \leq I(Y; V \mid H) \leq \epsilon(\delta) \tag{9}$$

Equation (9) demonstrates that the upper bound of the visual information gain is dictated entirely by the uncertainty remaining in the historical text. Because gaming data yields highly predictable action sequences ($\delta \to 0$, causing $\epsilon(\delta) \to 0$), the available information bandwidth for the visual modality is mathematically squeezed to zero. Regardless of the visual encoder's theoretical capacity, standard SFT optimization renders the visual input structurally redundant. □

A.2 Proof: HANI's Disruption of the Information Shortcut

We now formally prove that injecting discrete semantic noise strictly bounds the information capacity of the historical text modality, thereby forcing the model to re-allocate its attention to the visual modality.

Theorem 2. *Let the HANI mechanism substitute the true history H with uniformly sampled noise $\tilde{H}$ with probability p. The mutual information $I(Y; H')$ provided by the modified history H' is strictly upper-bounded by a factor of $(1-p)$. Consequently, to minimize uncertainty about Y, the optimization objective is mathematically compelled to extract visual information proportional to $p \cdot I(Y; V)$.*

Proof. Let $Z \sim \text{Bernoulli}(p)$ be an indicator variable representing the noise injection event. The modified history sequence is defined as:

$$H' = \begin{cases} \tilde{H} & \text{if } Z = 1 \\ H & \text{if } Z = 0 \end{cases} \tag{10}$$

where $\tilde{H}$ is sampled uniformly from the action vocabulary $\mathcal{V}_A$ and is strictly independent of the target Y and visual input V.

1. Attenuation of Historical Information:
We first analyze the information provided by H'. By the properties of mutual information, we can bound $I(Y; H')$ by conditioning on the latent injection state Z:

$$I(Y; H') \leq I(Y; H', Z) = I(Y; Z) + I(Y; H' \mid Z) \tag{11}$$

Because the decision to inject noise (Z) is independent of the target action (Y), $I(Y; Z) = 0$. Expanding the conditional mutual information yields:

$$\begin{aligned} I(Y; H' \mid Z) &= P(Z=0) I(Y; H' \mid Z=0) + P(Z=1) I(Y; H' \mid Z=1) \\ &= (1-p) I(Y; H) + p I(Y; \tilde{H}) \end{aligned} \tag{12}$$

Since $\tilde{H}$ is pure independent noise, $I(Y; \tilde{H}) = 0$. Therefore, substituting back into Eq. (11), we obtain a strict upper bound:

$$I(Y; H') \leq (1-p) I(Y; H) \tag{13}$$

This proves that HANI creates a structural information bottleneck, linearly degrading the reliability of the shortcut.

2. Lower Bound on Visual Information Requirement:
The cross-entropy loss optimized during SFT implicitly maximizes the total joint mutual information $I(Y; V, H')$. Conditioning again on Z, the expected total information extracted is:

$$\begin{aligned} I(Y; V, H' \mid Z) &= (1-p) I(Y; V, H) + p I(Y; V, \tilde{H}) \\ &= (1-p) I(Y; V, H) + p I(Y; V) \end{aligned} \tag{14}$$

(Note that $I(Y; V, \tilde{H}) = I(Y; V)$ because $\tilde{H}$ provides zero additional information).

By the chain rule, $I(Y; V, H' \mid Z) = I(Y; H' \mid Z) + I(Y; V \mid H', Z)$. Substituting Eq. (12) and rearranging for the visual term:

$$I(Y; V \mid H', Z) = (1-p) I(Y; V \mid H) + p I(Y; V) \tag{15}$$

From Appendix A.1, we established that $I(Y; V \mid H) \leq \epsilon(\delta) \to 0$. Thus:

$$I(Y; V \mid H', Z) \approx p \cdot I(Y; V) \tag{16}$$

Because gaming text sequences have a highly specific Markovian structure, deep networks easily distinguish uniform random sequences implicitly (acting as a proxy for identifying Z). Therefore, to minimize the loss, the network is mathematically mandated to establish a direct mapping from pure visual features V to Y during the p-proportion of noisy iterations. □

A.3 Proof: Optimization Dynamics and Gradient Awakening

While Appendix A.2 proves the necessity of visual information from a theoretical capacity standpoint, we must also prove that the optimizer can physically escape the local minimum. We demonstrate this by analyzing the gradient magnitude of the cross-entropy loss.

Theorem 3. *In standard SFT, the historical action shortcut drives the loss gradient with respect to the visual parameters toward zero (Gradient Starvation). HANI induces a gradient magnitude strictly bounded away from zero, enforcing continuous exploration of the visual feature space (Gradient Awakening).*

Proof. For analytical clarity, assume the model fuses multimodal representations via a generalized linear layer before the softmax activation. Let the logit for class $c \in \mathcal{V}_A$ be z_c:

$$z_c = W_v \cdot f_v(V; \theta_v)_c + W_h \cdot f_h(H'; \theta_h)_c \tag{17}$$

where f_v and f_h are the visual and text encoders parameterized by θ_v and θ_h, respectively. Let y be the ground-truth target class. The cross-entropy loss is $\mathcal{L} = -\log P(y \mid V, H')$, where $P(c) = \text{softmax}(z)_c$.

By the chain rule, the gradient of the loss with respect to the visual parameters θ_v is:

$$\nabla_{\theta_v} \mathcal{L} = \sum_{c \in \mathcal{V}_A} \underbrace{\big[P(c) - \mathbb{I}(c = y)\big]}_{\text{Prediction Error } E_c} \cdot W_v \cdot \nabla_{\theta_v} f_v(V)_c \tag{18}$$

Note that the per-class error term satisfies $|E_c| \leq 1$ by construction (as the difference between a probability and an indicator), *independent of the semantic distance* between the injected noise token and the true target—this yields the stability bound referenced in the main text.

Scenario A: Standard SFT (Gradient Starvation)
When $H' = H$, the text encoder f_h easily identifies the deterministic shortcut (as H is highly predictive of y). The text logits dominate, driving $P(y) \to 1$. Consequently, the error term for the target class becomes $E_y = (1 - 1) = 0$, and for all other classes $c \neq y$, $E_c \to 0$. Substituting this into Eq. (18):

$$\lim_{P(y) \to 1} \|\nabla_{\theta_v} \mathcal{L}\|_2 \approx 0 \tag{19}$$

Because the gradient prematurely vanishes, the visual parameters θ_v fail to update before meaningful visual features can be extracted.

Scenario B: HANI (Gradient Awakening)
When $H' = \tilde{H}$ (noise injected), the text encoder f_h—which has learned to blindly trust the history—processes the misleading semantic token and generates a highly confident, but incorrect, logit for some noise class $k \neq y$. This semantic mismatch collapses the predicted probability of the true target: $P(y) \to 0$, while $P(k) \to 1$.

The prediction errors now approach their theoretical extremes:

$$E_y = P(y) - 1 \approx -1$$
$$E_k = P(k) - 0 \approx +1$$

Substituting these bounds into Eq. (18):

$$\nabla_{\theta_v} \mathcal{L} \approx W_v \cdot \big(\nabla_{\theta_v} f_v(V)_k - \nabla_{\theta_v} f_v(V)_y\big) \tag{20}$$

Because $\tilde{H}$ contains no information about y, updating the text parameters θ_h cannot resolve this systematic error. The optimization algorithm is forced to compensate by relying on the non-zero gradient in Eq. (20). This robust gradient directly penalizes visual features aligned with the incorrect class k and heavily rewards features aligned with the true class y, violently pushing the optimization trajectory out of the historical action shortcut local minimum. □

B Experimental Setup Details

This section provides comprehensive details regarding the environments, dataset collection, training hyperparameters, and evaluation protocols used in our experiments.

B.1 Environment Settings

We utilized the *WallGap* environment from the MiniWorld framework. The complexity of the environment was modulated by varying the number of connected rooms from 2 to 5. To accommodate the increasing difficulty and traversing distances, the maximum episode durations were correspondingly set to 300, 450, 600, and 750 timesteps, respectively.

B.2 Dataset Collection and Preprocessing

Visual Observations: All visual frames collected from the environment were standardized to a resolution of 480p (640×480 pixels).

Training Data Construction: Expert trajectories were systematically collected by setting the environment seeds in the range of $[1{,}000{,}000,\ 1{,}001{,}000]$, yielding exactly 1,000 distinct procedural scene configurations. The total number of transition samples collected for the 2-room, 3-room, 4-room, and 5-room settings were 27,149, 54,411, 78,297, and 91,186, respectively.

Multimodal Context Window: We structured the input data using an interleaved sequence of visual frames and action chunks. The action chunk size was set to 3. Specifically, a single training sample comprises a sequence of four images and three historical action chunks formatted as: $I_0, (a_0, a_1, a_2), I_3, (a_3, a_4, a_5), I_6, (a_6, a_7, a_8), I_9$, where the subscript denotes the relative frame index and I_9 represents the current visual observation. The model is trained to autoregressively predict the subsequent future action chunk, specifically (a_9, a_{10}, a_{11}).

B.3 Training Settings

Model Architecture: We employed Qwen3-VL-8B-Instruct, an advanced Vision-Language Model with 8 billion parameters, as our base model for all experimental conditions.

Optimization: We conducted Full Supervised Fine-Tuning (Full SFT), updating the parameters of both the vision encoder and the language model components simultaneously. Training utilized the AdamW optimizer with a peak learning rate of 3.0×10^{-6}. The learning rate schedule followed a cosine decay strategy with a warmup ratio of 0.1. All models were trained for a total of 3 epochs.

Hardware and Batching: Training was distributed across 8 NVIDIA H20 GPUs. We used a per-GPU batch size of 2 and 8 gradient accumulation steps, resulting in an effective global batch size of 128 (calculated as $2 \times 8 \times 8$). The approximate total training durations for the 2, 3, 4, and 5-room configurations were 2, 4, 6, and 8 hours, respectively.

B.4 Evaluation Protocol

Zero-Shot Generalization: To rigorously test zero-shot generalization capabilities, evaluations were performed on purely unseen environment seeds ranging from 0 to 128. To minimize statistical variance introduced by random procedural generation, the agent was evaluated over 10 separate episodes for each distinct seed.

Inference Parameters: During the autoregressive generation phase, we applied a uniform decoding temperature of 0.7. Furthermore, to ensure the reliability and statistical reproducibility of our reported results, each experimental configuration was trained and evaluated independently three times, with the aggregated mean metrics reported in the main text.

B.5 Setup for Visually Richer 3D Environments

For the three additional 3D environments in Section 4.2.3, we retained the Qwen3-VL-8B-Instruct backbone, the interleaved image–action tokenization scheme, and the same optimization recipe as in the MiniWorld experiments; only the perturbation probability p was re-tuned per environment on a held-out validation split.

ViZDoom — Deadly Corridor. A 3D FPS scenario with dynamic enemies along a corridor. Expert demonstrations exhibit "keep moving forward + keep firing" inertia. Reward is the built-in Deadly Corridor return, averaged over held-out seeds.

Minecraft — MineRL Navigate. An open 3D voxel world in which human demonstrations are dominated by long streaks of the "W" (Forward) key, giving the most extreme low-entropy history among the three environments. Reward is the MineRL Navigate task score.

RoboTHOR — ObjectNav. A photorealistic indoor embodied navigation benchmark. Standard SFT is trapped by navigation inertia at corners and around furniture; success is reported as the ObjectNav task metric.

For each environment, HANI uses the same one-line data-loader modification as in MiniWorld: with probability p, all historical action tokens in the current sample are replaced by tokens sampled uniformly from the discrete (or tokenized) action vocabulary, while the ground-truth target action is left untouched.

C Additional Figures for the Diagnostic Testbed

We provide two supplementary figures illustrating the diagnostic testbed and the qualitative failure mode of standard SFT that motivate our study.

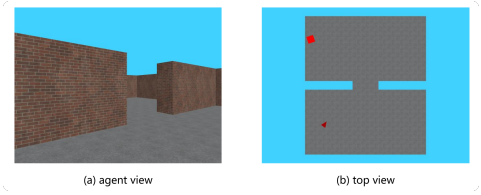


Figure 4: MiniWorld WallGap: an outdoor first-person 3D environment featuring two rooms connected by a gap in a wall. The objective is to navigate to the red box within a specified number of steps.

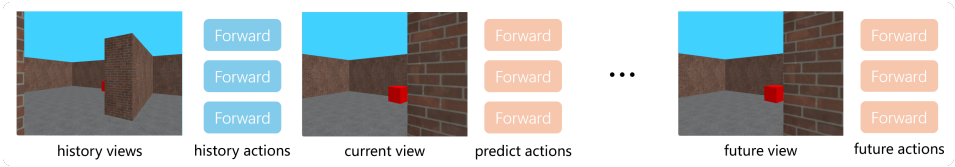


Figure 5: A qualitative failure of standard SFT. Because its previous action sequence is heavily dominated by *Forward*, the SFT-trained agent inertially keeps predicting *Forward* even when it is stuck against a wall—a direct manifestation of the historical-action shortcut.

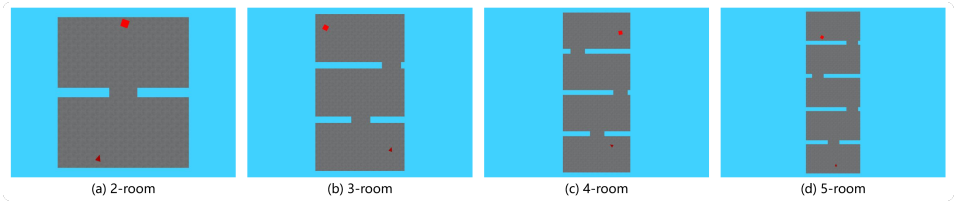


Figure 6: Expanded MiniWorld WallGap environment variants used in our main experiments (2–5 rooms). The objective is to navigate to the red box despite unpredictable room configurations.

D Formal Definitions of the Modality Diagnostics

Let $I_{1:t}$ be the image sequence, $A_{1:t-1}$ the historical action sequence, and A_t the ground-truth target action drawn from a discrete vocabulary $\mathcal{A}$.

Modality Information Gain (IG). We define the modality-conditional log-likelihood gain of A_t contributed by each modality on top of the other:

$$IG_{action} = \mathbb{E}\left[\log P(A_t \mid I_{1:t}, A_{1:t-1}) - \log P(A_t \mid I_{1:t})\right],$$
$$IG_{vision} = \mathbb{E}\left[\log P(A_t \mid I_{1:t}, A_{1:t-1}) - \log P(A_t \mid A_{1:t-1})\right].$$

$IG_{vision} \ll IG_{action}$ indicates that the visual modality contributes little marginal information once history is available—the diagnostic signature of visual modality collapse.

KL Divergence Shift. We compute the Kullback–Leibler divergence between the full-multimodal predictive distribution and each single-modality predictive distribution:

$$D_{KL}(P_{full} \parallel P_{vision}) = \sum_{a \in \mathcal{A}} P(a \mid I_{1:t}, A_{1:t-1}) \log \frac{P(a \mid I_{1:t}, A_{1:t-1})}{P(a \mid I_{1:t})},$$
$$D_{KL}(P_{full} \parallel P_{action}) = \sum_{a \in \mathcal{A}} P(a \mid I_{1:t}, A_{1:t-1}) \log \frac{P(a \mid I_{1:t}, A_{1:t-1})}{P(a \mid A_{1:t-1})}.$$

Whichever divergence is smaller identifies the modality that already reproduces the full-model distribution on its own, i.e. the modality that is actually in control.

Table 4: Modality Information Gain and KL Divergence Shift of the Base and SFT models on MiniWorld. SFT drives IG_{vision} from 0.4087 down to 0.1121 while IG_{action} stays at 0.3104, and $D_{KL}(P_{full} \parallel P_{action})$ collapses from 1.92 to 0.11—the model relies almost entirely on historical text.

Model	$\mathbf{IG_{action}}$	$\mathbf{IG_{vision}}$	$\mathbf{D_{KL}(P_{full} \parallel P_{vision})}$	$\mathbf{D_{KL}(P_{full} \parallel P_{action})}$
Base	0.4328 ± 0.0055	0.4087 ± 0.0012	0.3382 ± 0.0025	1.9234 ± 0.0103
SFT	0.3104 ± 0.1552	0.1121 ± 0.0561	0.3122 ± 0.1561	0.1116 ± 0.0558

E HANI: Full Algorithm

We give the complete HANI forward-pass pseudocode. HANI is implemented as a single Bernoulli-gated hook at the data loader; it introduces no additional parameters, no auxiliary networks, and zero inference-time overhead.

Algorithm 1 Historical Action Noise Injection (HANI) Forward Pass

Require:
Visual sequence $V = \{I_1, I_2, \ldots, I_t\}$, Real history $H = \{A_1, A_2, \ldots, A_{t-1}\}$
Target action $Y = A_t$, Discrete action vocabulary $\mathcal{V}_A$
Perturbation probability $p \in [0, 1]$, VLA Model $\mathcal{M}_\theta$ parameterized by θ
Ensure: Cross-entropy loss $\mathcal{L}_{CE}$ and parameter gradients $\nabla_\theta \mathcal{L}_{CE}$

Draw sequence-level perturbation trigger $Z \sim \text{Bernoulli}(p)$
Initialize perturbed history $H' \leftarrow \emptyset$
if $Z = 1$ **then** ▷ Inject adversarial semantic mismatch
  for $i = 1$ **to** $t - 1$ **do**
    Sample noise token $\tilde{A}_i \sim \text{Uniform}(\mathcal{V}_A \setminus \{A_i\})$ ▷ Guarantee orthogonal mismatch
    Append $\tilde{A}_i$ to H'
  end for
else ▷ Retain genuine history for intention coherence
  $H' \leftarrow H$
end if
Compute action probability distribution: $\hat{P} = \mathcal{M}_\theta(V, H')$ ▷ Visual modality V is fully isolated
Compute loss: $\mathcal{L}_{CE} = -\log \hat{P}(Y)$ ▷ Calculate error against true target
Compute gradients $\nabla_\theta \mathcal{L}_{CE}$ and update θ ▷ Trigger gradient awakening for θ_v

F Extended Design Rationale and Comparison with Conventional Regularization

F.1 Alternatives we considered

1. **Complete Removal of History.** Masking history entirely removes the shortcut but also removes the agent's short-term intent memory. Many maneuvers (e.g., jumping a hurdle, tracking a fleeing target) require multi-frame coherence, so full removal typically produces erratic, aimless movement.
2. **Traditional Regularization (Modality Dropout & NEFTune).** Modality Dropout (Korse et al., 2024) zeroes out text features during training; at inference the model receives genuine, continuous history, producing OOD shifts and catastrophic interference. NEFTune (Jain et al., 2023) adds continuous Gaussian noise to embeddings, but the robust self-attention of modern LLMs denoises such continuous perturbations effortlessly, so the low-entropy semantic shortcut remains.
3. **Data-side reweighting / keyframe filtering.** Loss-decay on repeated actions (Chen et al., 2025; Wang et al., 2025), focal / class-balanced loss, and keyframe/changepoint reweighting (Wen et al., 2021) all attack the *marginal* action distribution. They cannot alter the within-trajectory $P(A_t \mid A_{1:t-1})$ that drives visual gradient starvation, and they typically require auxiliary detectors or frame dropping that is incompatible with interleaved autoregressive VLA training.
4. **Our Approach (Adversarial Semantic Penalty).** HANI replaces historical action tokens with entirely incorrect semantics (e.g., mapping *Forward* to *Fire*) with probability p. The discrete, orthogonal mismatch cannot be bypassed by continuous-layer denoising, generates massive penalty gradients that re-awaken the visual encoder, and keeps enough genuine history (via $1 - p$) to prevent OOD shifts during deployment.

F.2 Empirical comparison with conventional regularizers

Tables 1 and 2 in the main text show NEFTune and Modality Dropout provide negligible improvements over Standard SFT. Modality Dropout induces severe train/inference OOD, because the model must fuse continuous real history at deployment that it never trusted during training. NEFTune's continuous Gaussian noise is trivially bypassed by the robust self-attention denoising of the 8B LLM (IG_{action} stays at 0.3648, essentially the SFT value). HANI, by contrast, uses discrete adversarial semantic mismatch, creating a logical conflict that forces the network to dynamically elevate visual priority without breaking the "image + text" interleaved data structure.

G Extended Contrast against Data-Side Rebalancing Methods

Four families of related interventions warrant explicit contrast, because they treat a superficially similar symptom via a fundamentally different mechanism.

(i) Repeated-action downsampling / state-changing-frame filtering. These drop or downweight frames where the expert action does not change. Interleaved VLA training, however, uses a token stream $[I_t, A_t, \dots]$; dropping frames destroys the causal sequence the transformer conditions on and induces train/inference OOD, because the deployed agent sees continuous streams.

(ii) Balanced batch sampling. Gaming trajectories are sequential POMDPs; shuffling breaks causal attention. Trajectory-level resampling barely changes the *within-trajectory* $P(A_t \mid A_{1:t-1})$ that drives the shortcut.

(iii) Keyframe reweighting (Wen et al., 2021). A multi-stage pipeline (copycat detector + frame scoring + reweighting) whose auxiliary detector must itself be trained on shortcut-prone data, and which is expensive to scale to 8B backbones.

(iv) Focal / class-balanced loss. Amplifies gradients on hard/rare *classes*, but for most frames along a continuous trajectory the prediction error is already near zero under SFT—there is no hard-example gradient to amplify.

In all four cases, our theory (Appendices A.2–A.3) predicts marginal improvements on rare-action success but essentially unchanged IG_{vision}—because none of these methods removes the cross-modal shortcut that HANI directly targets.

H Extended Discussion of Limitations and Future Directions

Continuous action spaces. Two natural extensions exist. (i) *Tokenized-continuous VLAs.* SOTA VLAs such as RT-2, OpenVLA, and Qwen-VL-style models discretize continuous actions into a fixed vocabulary of bins (e.g., `<action_25>`); HANI applies *directly* by sampling a replacement token uniformly from the vocabulary. (ii) *Native continuous heads.* For MSE regression heads, Gaussian noise is insufficient (NEFTune already shows that 8B LLMs denoise it away). A principled extension is *Adversarial Trajectory Mirroring*: with probability p, inject extremal or reversed historical vectors (e.g., leftward history $\rightarrow$ rightward). Since the target y is unchanged, this maximizes $(\hat{y} - y)^2$ and reproduces Gradient Awakening in continuous space.

Structure-aware noise. Rather than uniform random mapping, directional mapping (e.g., mapping *Forward* explicitly to *Backward*) based on a semantic action tree, or a curriculum from close to distant noise, could yield sharper corrective gradients while retaining HANI's provable stability bound $|E_c| \leq 1$.

Long-context intent forgetting. If VLAs support exceptionally long histories (e.g., thousands of frames), token-level random replacements might alter macroscopic intent. Block-wise (contiguous) perturbations could offer a balanced solution.

Scope reiterated. Our claims are made explicitly for gaming/embodied VLAs with discrete (or tokenized) action vocabularies, where humans emit near-zero-noise identical key-press streams that trigger the theoretical low-entropy regime. Physical-robot regimes with sensor jitter, IK variability, and continuous joints reach this regime less frequently, and we regard the physical-robot deployment of HANI as an important next step rather than a claim of this paper.