INTERVENTION, NOT SHARED LATENTS: BLOCKING VISUAL SHORTCUTS IN AUDIO–VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint audio–video (AV) generators are trained on data in which *what an event looks like* and *what it sounds like* are spuriously correlated. We present a *controlled causal study* of the resulting failure mode. In an AV structural causal model where the audio is, by construction, independent of the video's nuisance appearance, models that let audio read video directly—through cross-attention or a shared latent—learn a *visual shortcut*: they predict sound from appearance rather than the causal event and, when the appearance–event correlation is broken at test time, synthesize the wrong event's sound. Crucially, the popular remedy of routing both modalities through a *shared common-cause latent* does *not* fix this—a bottleneck, an unsupervised shared/private factorization, and a faithful shared-prior model all grab the appearance proxy and fail like the direct model. Blocking the shortcut instead requires an *intervention on the nuisance*: under the stated assumptions we prove that counterfactual invariance is necessary and sufficient to identify the causal predictor, and we verify the mechanism from feature-vector SCMs to procedural pixel video, real images with spectrogram audio, moving real digits, and a conditional generator. On a *real, pretrained* V2A generator (MMAudio), an input-intervention test shows the model is far from invariant to sound-irrelevant edits, though a generic-noise control reveals it is broadly input-brittle rather than specifically colour-shortcutting—clean isolation of the shortcut needs the controlled confounds our synthetic studies provide. We characterize *when* the shortcut occurs, compare the objective against supervised counterfactual augmentation, and isolate the *unknown-nuisance* regime—where the intervention cannot be applied—as the central open problem.

1 INTRODUCTION

A generative model of paired audio and video must decide *where the sound comes from*. In natural data the honest answer is an *event*: a cup hits the floor, and that single cause produces both a visual impact and an acoustic one. But the same datasets also contain strong *nuisance* regularities—a given instrument almost always looks a certain way, a given material almost always has a certain texture—so that appearance and event are tightly, and partly spuriously, coupled during training. A model is then free to take a shortcut (Geirhos et al., 2020): rather than reading the causal event content out of the video, it can read the nuisance appearance, use it as a proxy for the event, and generate the associated sound. In-distribution this is indistinguishable from understanding; under any shift that decouples appearance from event—a new material, an occluded source, a changed viewpoint—it is wrong.

Recent joint AV diffusion models couple the two streams either by direct cross-attention (Ruan et al., 2023; HaCohen et al., 2026) or, more recently, by a *shared* spatio-temporal prior that both branches align to (Liu et al., 2026). The latter is often motivated precisely as a way to avoid spurious cross-modal coupling. Independent physical-commonsense benchmarks, however, report that current video-to-audio and joint models remain sensitive to exactly the visual factors they should ignore (Cui et al., 2026; Li et al., 2026), suggesting the shortcut survives these architectural choices. This paper asks a sharp version of the question: *is sharing a latent enough to remove the visual shortcut, and if not, what is?*

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.

Our answer is negative for architecture and positive for intervention. We first formalize the shortcut as a confounded structural causal model (SCM) in which audio is, by construction, independent of the video's nuisance appearance given the event and scene. We then show, in controlled settings where ground truth is known exactly, three things:

- **The shortcut is real and catastrophic.** A direct model attains near-zero in-distribution audio error but fails by up to two orders of magnitude when appearance is decoupled from event, and the failure is *slice-local*: it appears only on the event classes that were confounded in training.
- **Sharing a latent does not fix it.** A common-cause bottleneck, *and* an unsupervised shared/private factorization that generates audio from the shared code alone, both fail as badly as the direct model: when appearance perfectly predicts the event, the "shared" code simply encodes appearance. This is consistent with known impossibility results for unsupervised disentanglement (Locatello et al., 2019).
- **Counterfactual invariance is necessary and sufficient under the stated SCM assumptions.** We enforce it through a counterfactual-consistency (do-operator) objective: forcing generated audio to be invariant to interventions on the nuisance restores near-oracle robustness at negligible in-distribution cost and without harming unconfounded data. When the nuisance can be augmented (e.g. recoloring, restyling), a single feasible augmentation-based intervention suffices—and, importantly, the mechanism survives the move from abstract feature vectors to rendered pixel video.

What is and is not new here and the study's scope are discussed in Appendix A; limitations and the unknown-nuisance problem in Appendix B.

2 Related Work

Joint audio–video generation. Joint AV diffusion models synchronize modalities through cross-modal attention or shared backbones (Ruan et al., 2023; Xing et al., 2024; HaCohen et al., 2026), and more recent work introduces an explicit shared spatio-temporal prior that both branches align to (Liu et al., 2026). These designs target *synchrony*; we instead target *causal correctness* of the sound given the event, and we show that a shared latent alone does not deliver it. Physical reasoning benchmarks for AV generation (Cui et al., 2026; Li et al., 2026) document that generated audio tracks spurious visual factors, providing external evidence that the shortcut we isolate occurs in practice.

Shortcut learning, causality, and invariance. Shortcut learning (Geirhos et al., 2020) and its remedies—invariant risk minimization (Arjovsky et al., 2019), counterfactual invariance (Veitch et al., 2021), and neural structural causal models (Yang et al., 2021)—provide the conceptual backbone for our objective. We contribute the specific instantiation for *generative* audio–video models: audio generation invariant to interventions on visual nuisance, together with controlled evidence that the intervention, not the shared-latent architecture, is what removes the shortcut.

3 The Visual Shortcut in Audio–Video Generation

Structural causal model. Let an *event* e and a *scene* s (e.g. source distance) be the common causes of a video x_{v} and an audio signal x_{a}, and let n be a *nuisance* appearance factor (material, texture, color) that affects the video only:

$$x_{\mathrm{v}} = g_{\mathrm{v}}(\mathrm{e}, \mathrm{n}) + \epsilon_{\mathrm{v}}, \qquad x_{\mathrm{a}} = g_{\mathrm{a}}(\mathrm{e}, s) + \epsilon_{\mathrm{a}}, \qquad x_{\mathrm{a}} \perp \mathrm{n} \mid (\mathrm{e}, s). \tag{1}$$

By construction the sound depends on the event and scene but *not* on the video's appearance. Training data, however, is confounded: the nuisance is correlated with the event, $\mathrm{n} = \mathrm{e}$ (in appropriate coordinates) with probability ρ_{spur}, and independent otherwise. A predictor or conditional generator $f(x_{\mathrm{v}}, s) \approx x_{\mathrm{a}}$ may therefore route through either the *causal* path (event content in x_{v}) or the *shortcut* (read n off x_{v}, use it as a proxy for e). The two are indistinguishable in-distribution; we evaluate under the counterfactual shift $\mathrm{do}(\mathrm{n} \perp \mathrm{e})$ that decouples appearance from event.

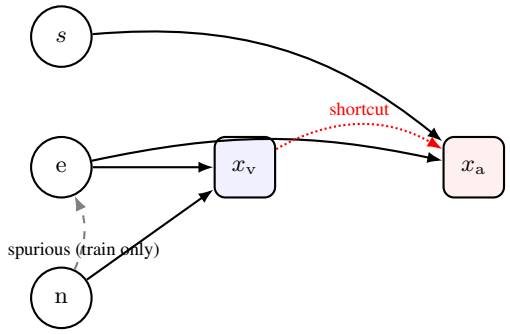

Figure 1: The audio–video SCM. The event e (and scene s) cause both video and audio; the nuisance appearance n affects only the video, and $x_a \perp \mathrm{n} \mid (\mathrm{e}, s)$. Training confounds n with e (dashed), so a model can generate audio via the causal path (e read out of x_v) or via the shortcut (n read off x_v). Our objective forbids the shortcut by enforcing invariance to interventions on n.

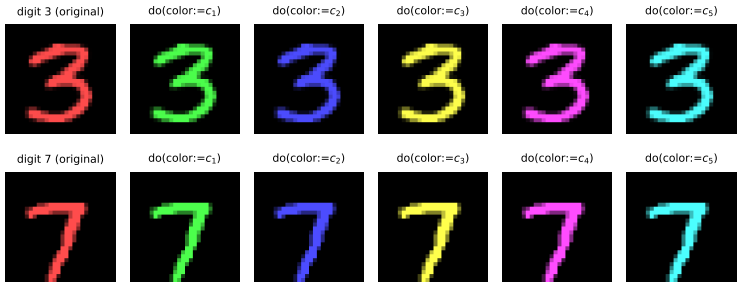

Figure 2: The counterfactual intervention $\mathrm{do}(\mathrm{n} := \mathrm{n}')$ made concrete (Colored-MNIST): the event—the digit—is fixed, the nuisance—its colour—is resampled. $\mathcal{L}_{\mathrm{cf}}$ penalizes any change in the generated audio along such a row.

Why a shared latent is not enough. A common-cause model infers a shared code $z_s = E(x_v)$ and generates audio from z_s alone. If n perfectly predicts e in training, then n is a *sufficient statistic* for the audio-relevant information, and nothing in a reconstruction- or prediction-driven objective prevents z_s from encoding appearance instead of event. A shared/private split $x_v \mapsto (z_s, z_v)$ with audio from z_s does not resolve this either, in line with the impossibility of unsupervised disentanglement without inductive bias or supervision (Locatello et al., 2019). We confirm both failures empirically in Section 6.

4 Method: Counterfactual Consistency

We block the shortcut by requiring the generated audio to be invariant to interventions on the nuisance. Let $\mathrm{do}(\mathrm{n} := \mathrm{n}')$ denote replacing the appearance while holding the event and scene fixed, producing a counterfactual video x_v'. We add to the generation/prediction loss $\mathcal{L}_{\mathrm{gen}} = \|f(x_v, s) - x_a\|^2$ the *counterfactual-consistency* term

$$\mathcal{L}_{\mathrm{cf}} = \mathbb{E}_{\mathrm{n}'} \left\| f(x_v, s) - f\left(x_v', s\right) \right\|^2, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \lambda \mathcal{L}_{\mathrm{cf}}. \tag{2}$$

This is a do-operator objective: it enforces $f(\mathrm{do}(\mathrm{n}))$ invariance, which under the SCM of Section 3 is exactly the condition that f use only the event/scene path. Figure 2 shows what the intervention looks like concretely in one of our real-image settings: the event (the digit) is held fixed while the nuisance (its colour) is resampled, and the objective demands that the generated audio not change across the row.

Two regimes arise. When the nuisance is *identifiable and augmentable* (color, texture, style), the counterfactual x_v' is produced by an appearance augmentation and $\mathcal{L}_{\mathrm{cf}}$ is directly computable—the practical recipe we validate. When the nuisance is *unknown*, no augmentation is available, and one

must fall back on representation learning; we show this fallback currently fails, and we flag it as the open problem (Appendix B).

5 THEORY

We make the claims of Sections 3–4 precise. Let the latent factors $(\mathrm{e}, s, \mathrm{n})$ have joint law P, let $x_\mathrm{v} = g_\mathrm{v}(\mathrm{e}, \mathrm{n})$, and write $\mu(\mathrm{e}, s) := \mathbb{E}[x_\mathrm{a} \mid \mathrm{e}, s, \mathrm{n}]$. All proofs are deferred to Appendix D.

Assumption 1 (Causal audio)**.** $\mathbb{E}[x_\mathrm{a} \mid \mathrm{e}, s, \mathrm{n}] = \mu(\mathrm{e}, s)$ *does not depend on* n*; equivalently* $x_\mathrm{a} \perp \mathrm{n} \mid (\mathrm{e}, s)$.

Assumption 2 (Recoverability)**.** g_v *is injective in* (e, n)*, so any measurable predictor factors as* $f(x_\mathrm{v}, s) = \tilde{f}(\mathrm{e}, \mathrm{n}, s)$.

Write $R_D(f) = \mathbb{E}_D \|f(x_\mathrm{v}, s) - x_\mathrm{a}\|^2$. Since $R_D(f) = \mathbb{E}_D \|\tilde{f} - \mu\|^2 + \mathrm{const}$, minimizing risk is equivalent to matching μ. The *causal predictor* is $f^\star(x_\mathrm{v}, s) := \mu(\mathrm{e}, s)$. Let P_tr be the (confounded) training law and P_te the counterfactual test law with $\mathrm{n} \perp (\mathrm{e}, s)$ but the same (e, s) marginal.

Proposition 1 (Confounding makes ERM shortcut-blind)**.** *Suppose* $\mathrm{n} = \pi(\mathrm{e})$ P_tr*-a.s. for an injective* π *(perfect confounding), and define the shortcut predictor* $f_\mathrm{sc}(x_\mathrm{v}, s) := \mu(\pi^{-1}(\mathrm{n}), s)$*. Then (i)* $R_\mathrm{tr}(f_\mathrm{sc}) = R_\mathrm{tr}(f^\star)$*, so both minimize the training risk; and (ii)* $R_\mathrm{te}(f_\mathrm{sc}) - R_\mathrm{te}(f^\star) = \mathbb{E}\,\|\mu(\mathrm{e}', s) - \mu(\mathrm{e}, s)\|^2 > 0$ *with* $\mathrm{e}' = \pi^{-1}(\mathrm{n}) \perp \mathrm{e}$*, whenever* $\mu(\cdot, s)$ *is non-constant.*

Thus training risk does not identify $f^\star$: an ERM learner may return the shortcut, whose OOD excess risk is exactly the event-explained audio variance. This is the phase transition of Table 1 (any decorrelated mass breaks $\mathrm{n} = \pi(\mathrm{e})$ and restores identifiability).

Proposition 2 (Shared-latent / bottleneck insufficiency)**.** *Consider generators that produce audio from a code* $z = E(x_\mathrm{v})$ *chosen to minimize* R_tr *under an information (or dimension) constraint, and assume (as used in the proof) that the event is audio-identifiable on a set of scenes of positive measure—formally* $\forall\, \mathrm{e} \neq \mathrm{e}'$*,* $P_s[\mu(\mathrm{e}, s) \neq \mu(\mathrm{e}', s)] > 0$*—so that an audio-optimal code must determine the event. Under the hypothesis of Proposition 1,* $E(x_\mathrm{v}) = \mathrm{n}$ *attains the optimal training risk at rate* $I(z; x_\mathrm{v}) = H(\mathrm{n}) = H(\mathrm{e})$*, i.e. it is simultaneously risk-optimal and maximally compressive; hence the bottleneck objective admits* $E(x_\mathrm{v}) = \mathrm{n}$ *among its minimizers: it* cannot guarantee *the event is preferred over the nuisance. This does not mean every shared-latent model must fail—only that sharing a latent is not* sufficient *to rule the shortcut out. The same holds for a shared/private split with audio decoded from the shared code.*

This matches the collapse of the bottleneck and unsupervised shared/private models in Table 8, and is consistent with the impossibility of unsupervised disentanglement without inductive bias (Locatello et al., 2019).

Theorem 1 (Counterfactual invariance identifies the causal predictor, under known and faithful nuisance interventions (Assumptions 1–2))**.** *Let* $\mathcal{F}_\mathrm{inv} = \{f : \tilde{f}(\mathrm{e}, \mathrm{n}, s)$ *is independent of* $\mathrm{n}\}$*. Then:*

- *(i)* ***Identification.*** $\arg\min_{f \in \mathcal{F}_\mathrm{inv}} R_\mathrm{tr}(f) = f^\star$*, and this predictor also minimizes* R_te *over all measurable* f*.*
- *(ii)* ***Gap closure.*** *For every* $f \in \mathcal{F}_\mathrm{inv}$*,* $R_\mathrm{tr}(f) = R_\mathrm{te}(f)$ *(the two laws share the* (e, s) *marginal), so constrained ERM transfers without a nuisance-induced gap.*
- *(iii)* ***Necessity.*** *The global* R_te*-minimizer equals* $f^\star \in \mathcal{F}_\mathrm{inv}$ P_te*-a.s.; restricting to* $\mathcal{F}_\mathrm{inv}$ *loses no optimality.*

Moreover the counterfactual penalty $\mathcal{L}_\mathrm{cf}(f) = \mathbb{E}_{\mathrm{n}'} \|f(g_\mathrm{v}(\mathrm{e}, \mathrm{n}), s) - f(g_\mathrm{v}(\mathrm{e}, \mathrm{n}'), s)\|^2$ *satisfies* $\mathcal{L}_\mathrm{cf}(f) = 0 \iff f \in \mathcal{F}_\mathrm{inv}$*, so minimizing* $R_\mathrm{tr}(f) + \lambda\, \mathcal{L}_\mathrm{cf}(f)$ *recovers* $f^\star$ *as the constraint becomes active.*

Theorem 1 is the formal counterpart of Tables 1–8: the intervention restricts the hypothesis class to the causal path, which is simultaneously the train- and test-optimal solution.

What the theorem does *not* say. The statement is necessary-and-sufficient *under the stated SCM and intervention assumptions*, and those assumptions are strong: the nuisance is known and can be

Table 1: Audio MSE vs. training confounding strength ρ_{spur} (feature-vector SCM). OOD-cf decouples appearance from event. The direct and bottleneck models fail catastrophically as confounding approaches perfect in this setting; counterfactual consistency is robust throughout.

	Direct		CC-bottleneck		CC+CF (ours)	
ρ_{spur}	IID	OOD-cf	IID	OOD-cf	IID	OOD-cf
0.50	0.0012	0.0013	0.0023	0.0024	0.0019	0.0020
0.70	0.0012	0.0015	0.0023	0.0026	0.0020	0.0020
0.80	0.0012	0.0017	0.0022	0.0028	0.0020	0.0020
0.90	0.0012	0.0021	0.0021	0.0035	0.0020	0.0021
0.95	0.0011	0.0026	0.0020	0.0049	0.0020	0.0021
1.00	0.0010	**0.1816**	0.0016	**0.1637**	0.0019	**0.0020**

intervened on, the intervention is faithful (it changes only the nuisance and leaves the event intact), *every* confounded nuisance is covered (Proposition 3), the audio is genuinely conditionally independent of the nuisance (Assumption 1), and the hypothesis class and optimizer are rich enough to realize $f^\star$. Where any of these fails—most importantly when the nuisance is unknown—the guarantee does not transfer, and Appendix B is explicit about this. Under these assumptions the result is admittedly close to a definitional consequence—constraining a predictor to be n-invariant leaves only the n-independent causal mean—and it is a *population*, exact-realizability statement: we do not give finite-λ, finite-sample, or non-convex-optimization identification or generalization guarantees, and we make no such claims. The value is in tying the failure and its cure to a single, checkable condition, not in the depth of the proof. The theory is a statement about the idealized causal structure, not a claim about deployed AV generators. Theorem 1 also concerns conditional-*mean* prediction under squared error; its extension to conditional generative *distributions* $p(x_{\mathrm{a}} \mid \mathrm{e}, s)$—the setting of our flow-matching generator (Table 5)—is supported *empirically* rather than covered by the theorem, although the same nuisance-invariance argument goes through for any strictly proper scoring rule (Appendix D, Remark 1). Finally, the partial-augmentation behavior of Section 6 is explained by:

Proposition 3 (Partial augmentation leaves a residual shortcut)**.** *Let* $\mathrm{n} = (\mathrm{n}_1, \ldots, \mathrm{n}_L)$ *be conditionally independent given* e *and each perfectly confounded,* $\mathrm{n}_l = \pi_l(\mathrm{e})$*. Enforcing CF-invariance only on* $S \subseteq \{1, \ldots, L\}$ *yields predictors that may still depend on* n_{S^c}*; the resulting OOD excess risk equals the event-explained audio variance recoverable from* n_{S^c}*, and is zero iff* S *contains every confounded nuisance.*

6 Experiments

We isolate the statistical question in settings where the causal structure, and hence the ground-truth robustness, is known exactly. All models predict audio x_{a} from video x_{v} and scene s; we report mean-squared audio error in-distribution (IID, with the train-time confounding intact) and under the counterfactual shift (OOD-cf, appearance decoupled from event). We compare a **direct** model (audio reads the full video; the cross-attention analogue), a **common-cause bottleneck**, an unsupervised **shared/private disentanglement** (audio from the shared code only), and **counterfactual consistency (CC+CF)**.

6.1 The shortcut is real, and (in this setting) sharpest near perfect confounding

Table 1 sweeps the training confounding strength ρ_{spur} in a feature-vector SCM (K=8 events, 8 nuisance values). The direct model is essentially perfect IID at every ρ_{spur}, but its OOD-cf error rises and then jumps catastrophically as confounding approaches 1: a few decorrelated training examples are enough for it to learn the causal path, but under *perfect* confounding it has no incentive to, and fails by $\sim 140\times$ (Figure 3). The bottleneck tracks the direct model and also collapses. Counterfactual consistency stays flat and low across the entire sweep, at the cost of a small constant IID overhead.

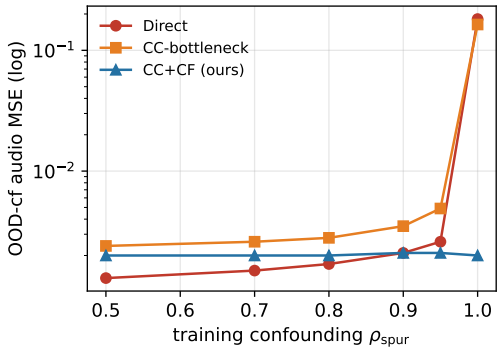

Figure 3: Out-of-distribution audio error vs. training confounding strength (log scale; same data as Table 1). The shortcut failure of the direct and bottleneck models is a sharp phase transition at $\rho_{\text{spur}} \to 1$; counterfactual consistency is flat and robust.

Table 2: Slice-local confounding: half the event classes are perfectly confounded, half are clean. Audio MSE under the OOD-cf shift, split by slice.

Model	OOD-cf (confounded slice)	OOD-cf (clean slice)
Direct	0.0488	0.0012
CC+CF (ours)	**0.0015**	0.0016

6.2 The failure is slice-local, and the fix has no collateral cost

Real datasets confound only *some* event–appearance pairs. Table 2 confounds half of the event classes perfectly and leaves the rest decoupled. The direct model fails only on the confounded slice ($40\times$ worse) and is perfect on the clean slice; counterfactual consistency is low on both, fixing the shortcut exactly where it exists without harming unconfounded events. The objective thus costs little: it fixes the confounded slice at a small in-distribution price (a modest rise in mel-MSE that barely dents event accuracy; Table 3), and does no harm on unconfounded events.

On genuine pixels, and in temporal video. The mechanism is not an artifact of abstract feature vectors. On procedurally rendered pixel video (event = shape arrangement, nuisance = colour) the direct model's OOD-cf error explodes by $100\times$ and a bottleneck and an unsupervised shared/private factorization fail just as hard, while a recolor-consistency intervention stays robust—architecture alone does not block the shortcut. The same holds when the causal cue is genuinely *temporal*: with the event a cross-frame *motion* and colour a per-frame shortcut, the direct model's motion-accuracy collapses from 1.00 to 0.52 while counterfactual consistency stays at 1.00. Full tables and the video illustration are in Appendix F (Tables 8 and 9, Figure 8).

6.3 A faithful shared-prior model fails too

Our claim that "sharing a latent is not enough" should be tested against the mechanism actually used by recent joint AV models, which align both modalities to a *shared prior* (Liu et al., 2026), rather than against a generic bottleneck. We therefore add a faithful stand-in: a single shared code $z_{\text{s}} = E(x_{\text{v}})$ from which the audio is generated *and* the video is reconstructed, so that both modalities align to it. Because z_{s} must carry appearance in order to reconstruct the video, the audio branch inherits the shortcut, and the model collapses under the counterfactual shift exactly like the direct model (OOD-cf MSE 0.152 vs. direct 0.182 vs. ours 0.002; full table in Appendix G). This stand-in is a small encoder–decoder, not a full-scale joint-AV diffusion architecture, so we do not claim shared-prior models *necessarily* fail; the point is the one our theory makes—aligning modalities to a shared prior is *not sufficient* to remove the shortcut (consistent with Proposition 2)—so the intervention is still needed.

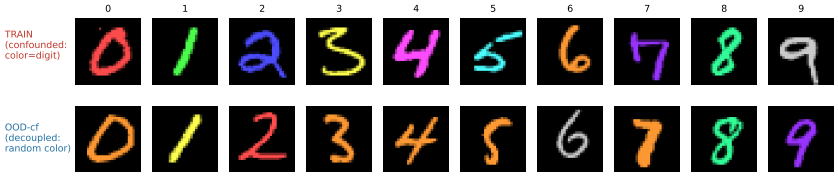

Figure 4: The real-modality setting. In training every digit always wears its own colour (perfect confounding); at test the colour is resampled independently of the digit. A model that has learned "colour ⇒ sound" will now render the wrong tone.

Table 3: Real modalities: Colored-MNIST (real digits, real spectrogram audio, frozen pretrained ResNet-18), ρ_{spur}=1. Mel-spectrogram MSE and event-accuracy (chance = 0.1; mean over 3 seeds, std $\leq$ 0.03 on MSE and $\leq$ 0.008 on accuracy). The color shortcut is fully exploited by the direct model and blocked by counterfactual consistency.

	IID		OOD-cf	
Model	mel-MSE	event-acc	mel-MSE	event-acc
Direct	0.001	1.00	2.359	0.160
CC+CF (ours)	0.037	0.98	**0.105**	**0.930**

6.4 Real image pixels and a real audio representation

We finally test whether the effect survives real image pixels, a real audio representation, and a real pretrained model. We stress that this is *not* a real audio–video generation system: the audio is synthesized and the confound constructed, and it is the closest offline proxy we can build, not a validation on a deployed AV generator. We use the Colored-MNIST protocol (Arjovsky et al., 2019): real MNIST digits (the event) are tinted with a color (the nuisance) that is perfectly confounded with the digit class in training and decoupled at test (Figure 4); the audio target is the log-mel spectrogram of a synthesized instrument tone whose pitch is set by the digit. The image encoder is a *frozen, ImageNet-pretrained ResNet-18* (He et al., 2016), and only a small audio head is trained.

Table 3 shows the same picture on real image pixels: the direct model is perfect in-distribution but its event-accuracy collapses to near chance (0.160, chance = 0.1) under the counterfactual color shift—it renders the wrong digit's tone—while counterfactual consistency stays robust (0.930), a $\sim 6\times$ improvement, at a modest in-distribution cost. Figure 9 makes this concrete: the direct model's generated spectrogram matches the tone of the digit whose *colour* it was shown, not the digit itself. The phenomenon and its fix thus survive real image statistics, a real audio representation, and a real pretrained backbone—though, we reiterate, not yet a real AV generator.

6.5 Is the consistency term doing anything a simple augmentation cannot?

The sharpest test of our objective is supervised counterfactual *augmentation*: if one can synthesize $x'_{\text{v}} = \text{do}(\text{n} := \text{n}')$ and knows it carries the same audio, one can simply train on it. We report this honestly. **With full labels the two are statistically indistinguishable** (0.0012 vs. 0.0011 OOD-cf MSE, 5 seeds), so we do *not* claim the consistency loss as the contribution. What the comparison establishes is the paper's thesis: every method that *intervenes on the nuisance* closes the gap ($\sim$ 0.001), while every method that merely reshapes the representation—the bottleneck, unsupervised shared/private factorization, the shared-prior model, and representation-level consistency (0.0140, $12\times$ worse)—does not. The two do diverge once labels are scarce: the consistency term is label-free and keeps improving from unlabeled video, a consistent $\sim 1.5\times$ advantage as the labeled fraction falls to 0.02. Full table, label-scarcity curve and discussion are in Appendix F (Table 6, Figure 5).

When is the shortcut taken? A shortcut is only attractive when it is *easier* than the causal path, so the phenomenon is conditional. Sweeping the causal-cue strength shows the direct model's counterfactual penalty collapsing smoothly from $263\times$ to $2.7\times$ as the cue becomes easy to read, while

Table 4: Sensitivity of a real V2A generator (MMAudio) to input edits, over 24 real videos $\times 3$ seeds (95% bootstrap CIs): sound-class flip rate and mean CLAP distance of the generated audio. Appearance edits change the sound far more than the minimal edit, but a *magnitude-matched* Gaussian-noise control (same per-video PSNR) moves it at least as much: the effect is broad input-brittleness, not an isolated colour shortcut.

Input edit	sound-class flip rate	mean CLAP distance
tiny (+2% brightness), robust ref.	22% [13,32]	0.060 [0.04,0.08]
time reversal (causal)	44% [33,57]	0.276 [0.24,0.32]
colour rotation (nuisance)	54% [43,65]	0.357 [0.32,0.40]
grayscale (nuisance)	62% [51,74]	0.342 [0.30,0.38]
Gaussian noise, generic (σ=0.15)	60% [47,71]	0.452 [0.42,0.49]
Gaussian noise, magnitude-matched to grayscale	60% [49,71]	0.400 [0.36,0.44]
Gaussian noise, magnitude-matched to colour	**65%** [54,76]	**0.472** [0.43,0.51]
different video (ceiling)	—	0.673

the counterfactual objective stays robust and nearly free where it is unnecessary. A Moving-MNIST test with *real* digit sprites confirms the falsifiable prediction: the shortcut appears only in the hard-motion regime (direct OOD accuracy 0.21 vs. ours 0.995) and there is nothing to fix when the motion is easy (both $\approx$1.0). Full sweep and table in Appendix F (Figure 6, Table 7).

6.6 A REAL VIDEO-TO-AUDIO GENERATOR IS NOT INVARIANT TO SOUND-IRRELEVANT EDITS

Our controlled settings construct the confound to isolate the shortcut. On a *real, deployed* model we cannot construct one, so instead of isolating a shortcut we test a necessary condition: is the model invariant to input edits that do not change what an event sounds like? We run MMAudio (Cheng et al., 2025), a pretrained video-to-audio generator, on 24 real videos (UCF101 (Soomro et al., 2012) actions with characteristic sounds, plus the model's own example clips). For each video and 3 seeds we generate audio from the original and from edited versions at a fixed generation seed (so any change is attributable to the edit) and measure the shift of the generated audio in CLAP space (Wu et al., 2023), as a distance and as a zero-shot top-1 *sound-class flip*. Two edits change only *appearance* while preserving motion (grayscale, colour-channel rotation); a +2% brightness edit is the robustness reference and time reversal is a causal (motion-changing) reference. Crucially we add a *generic out-of-distribution control*: additive Gaussian pixel noise, which carries no event information and is not a colour change.

The model is clearly *not* invariant to the appearance edits (Figure 10, Table 4): grayscale and colour rotation flip the top-1 sound class 62% and 54% of the time versus 22% for the minimal edit (non-overlapping CIs), shifting the CLAP embedding $\sim$ 5–6$\times$ as far. Recolouring or graying a video substantially changes the sound this real model generates. **But the appearance edits are not special once perturbation magnitude is controlled.** Calibrating a semantically-empty Gaussian-noise control to the same per-video PSNR as each appearance edit (grayscale $\approx$ 19.9 dB, σ=0.10; colour $\approx$ 15.6 dB, σ=0.17) and re-measuring, the matched noise flips the sound class as often or more: grayscale 62% vs. 60% (overlapping CIs), colour 54% vs. **65**%. The model's sensitivity tracks perturbation size, not appearance content. We therefore read the real-model result as *non-invariance* to sound-irrelevant edits (a necessary symptom consistent with shortcutting) and *do not* claim it isolates a colour/appearance shortcut; that isolation needs the controlled confound available only in the synthetic settings above.

In short, the real-model experiment confirms a necessary symptom—a deployed V2A generator is far from invariant to edits that leave the event unchanged—but, honestly, cannot attribute it specifically to a colour/appearance shortcut, because a magnitude-matched, semantically-empty noise perturbation is just as disruptive. The clean isolation of the shortcut, and of its fix, lives in the controlled studies; the real model shows the symptom is present at scale.

Table 5: Conditional *generation* (flow matching), $\rho_{\text{spur}}=1$. We report generated-sample audio MSE and the event-accuracy of the sampled audio (mean±std over 3 seeds). The direct generator synthesizes the shortcut-implied wrong event's sound OOD.

	IID		OOD-cf	
Model	MSE	event-acc	MSE	event-acc
Direct generator	0.0016	1.00	0.171 ± 0.002	0.130 ± 0.000
CC+CF generator (ours)	0.0016	1.00	$\mathbf{0.0016\pm0.000}$	$\mathbf{1.00\pm0.00}$

Does the intervention transfer to the real model? Finally, we ask whether the counterfactual objective can *reduce* the observed nuisance sensitivity in this real model. Without any real audio, we fine-tune MMAudio's velocity network for 1500 steps on 18 held-out videos with a recolor-consistency term ($\|v_\theta(x_t, t \mid \text{recolor}) - v_{\text{frozen}}(x_t, t \mid \text{video})\|^2$) anchored by distillation to the frozen model on the original video, with x_t sampled along the frozen model's own generation path. To gauge stability we fine-tune 3 times with independent seeds and evaluate every model, and the frozen baseline, on a common held-out set (12 videos $\times 2$ seeds) so "before" and "after" are strictly matched. The intervention *does* reduce the grayscale sensitivity it targets, reproducibly and without collapsing the model: the grayscale flip rate falls from 70.8% (frozen) to $54.2\% \pm 3.4$, while the different-video ceiling is preserved (0.67 to 0.69), so the model still produces distinct, video-dependent audio. **But the effect is not the surgical appearance-only correction one might hope for.** On the same set the colour-rotation flip is essentially unchanged (62.5% to $63.9\%\pm2.0$), while the generic Gaussian-noise sensitivity (83.3% to $70.8\%\pm0.0$) and even the *causal* time-reversal sensitivity (58.3% to $43.1\%\pm1.9$) fall by a similar margin: the lightweight fine-tune broadly *dampens* sensitivity to input perturbations rather than selectively erasing an appearance shortcut, as expected if the deployed model's non-invariance is broad rather than colour-specific. We therefore claim only what the data support: the objective transfers far enough to *measurably and stably reduce* the targeted appearance sensitivity while keeping the model functional, but it neither eliminates that sensitivity nor isolates it from the model's general input-brittleness. Cleanly de-biasing a large pretrained generator (separating appearance from causal sensitivity, via real-audio supervision or training from scratch) is left to future work.

6.7 The shortcut, and its fix, appear in a generative model

To confirm the phenomenon is not an artifact of audio *regression*, we repeat it with a genuine conditional *generator*: a flow-matching model (Lipman et al., 2023) that samples audio x_{a} from video features, trained directly or with the counterfactual-consistency term. Alongside sample error, Table 5 reports the *event-accuracy* of the sampled audio (does it match the correct event's spectrum). The direct generator is perfect in-distribution but its event-accuracy collapses to chance under the counterfactual shift (0.13, chance 0.125), generating the wrong event's sound (the one implied by the shortcut appearance), whereas counterfactual consistency keeps it at 1.0. A linear probe explains the gap: the nuisance material is fully decodable from the direct and bottleneck representations (accuracy 1.0) but close to chance (0.24) under counterfactual consistency, with event-decodability still 1.0, so the representation is reorganized around the cause. Finally, with several individually-sufficient nuisances robustness is all-or-nothing: any single un-augmented confounder preserves a complete shortcut (Proposition 3). The probe figure and the partial-augmentation curve are in Appendix F (Figure 7).

7 Conclusion

We showed that joint audio–video generators learn a visual shortcut, that sharing a latent does not remove it, and that—under the stated SCM assumptions— *counterfactual invariance* is necessary and sufficient to identify the causal predictor, an intervention on the nuisance rather than a change of architecture. This is a population, exact-realizability characterization; empirically the resulting consistency objective blocks the shortcut in controlled settings, including on pixel video and in a conditional generator. The contribution reframes robust AV generation as a question of intervention, and isolates the unknown-nuisance regime as the key remaining challenge.

Reproducibility Statement

All theoretical claims are stated with their assumptions in Section 5 and proved in full in Appendix D. The structural causal models, architectures, training objectives, and hyperparameters for every experiment are specified in Appendix E, which also documents the seed protocol and how means and standard deviations are reported (multi-seed tables and the matched-eval protocol for the real-model fine-tune). All datasets used are either procedurally generated from the described SCMs or public (MNIST; UCF101 (Soomro et al., 2012)); the real video-to-audio experiments use the publicly released MMAudio (Cheng et al., 2025) and CLAP (Wu et al., 2023) checkpoints. Anonymized code to reproduce all controlled experiments and figures is provided with the submission.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28901–28911, 2025.

Zijun Cui, Xiulong Liu, Hao Fang, Mingwei Xu, Jiageng Liu, Zexin Xu, Weiguo Pian, Shijian Deng, Feiyu Du, Chenming Ge, et al. Do joint audio-video generation models understand physics? *arXiv preprint arXiv:2605.07061*, 2026.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning (ICML)*, pp. 5815–5826, 2021.

Tingle Li, Siddharth Gururani, Kevin J Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking single-factor physical video-to-audio generation. *arXiv preprint arXiv:2605.30339*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations (ICLR)*, 2026.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Debiasing classifier from biased classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10219–10228, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208, 2021.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7161, 2024.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9593–9602, 2021.

A Positioning and Scope

What is and is not new here. Consistency regularization under a known transformation is a standard tool, and we do not claim it as a novel mechanism; the individual ingredients—counterfactual invariance (Veitch et al., 2021), invariant risk minimization (Arjovsky et al., 2019), causal latent-variable models (Yang et al., 2021)—all predate us. Our contribution is (i) to identify and formalize the visual shortcut as the failure mode of joint AV generation, (ii) to show, in controlled settings where the ground truth is known exactly, that the field's current remedy—aligning modalities to a shared latent—provably and empirically does *not* address it, (iii) to establish precisely when the intervention is required (and, equally, when it is not), and (iv) to isolate the unknown-nuisance regime as the concrete obstacle to deploying it. The shift is one of *burden*: from representation design to intervention.

Scope. This is a controlled causal study. Our most realistic settings use real image pixels, a real audio representation, and a pretrained backbone, but the audio is synthesized and the confounding is constructed—because measuring a shortcut counterfactually *requires* knowing the confound. We therefore establish a mechanism and its conditions, not a validated fix for any deployed AV system; transferring the objective to recorded in-the-wild AV corpora and full-scale joint AV generators is the natural next step, discussed in Section B.

B Limitations and the Unknown-Nuisance Problem

The main limitation concerns breadth of real-model evidence. On a real, pretrained video-to-audio generator we can only establish a *necessary symptom*, not an isolated shortcut: the model is far from invariant to sound-irrelevant input edits (Section 6.6), which is consistent with shortcutting but—because a generic Gaussian-noise control moves the generated sound at least as much as the appearance edits— cannot be attributed specifically to a colour/appearance shortcut. Isolating a specific shortcut requires the constructed confound we have only in the synthetic settings, where the *controlled* evidence necessarily uses synthesized audio (measuring a shortcut counterfactually requires knowing the confound). Our recolor-consistency fine-tune of the real model is likewise honest about what it shows: across 3 seeds on a matched evaluation set it *stably reduces the targeted grayscale sensitivity* ($70.8\% \rightarrow 54.2\%$) without collapsing the model (the different-video semantic ceiling is preserved), but it does not isolate appearance—colour flip is essentially unchanged and the generic-noise and causal time-reversal sensitivities fall by a similar margin, so the intervention broadly *dampens* input sensitivity rather than surgically removing an appearance shortcut. It thus neither demonstrates nor removes a real-model shortcut; that, and a survey beyond one model and global colour edits, remain open. Closing this gap—fully removing the shortcut inside a deployed joint-AV generator (with real-audio supervision or training from scratch), extending the intervention across models and to material/background edits, and building an appearance-preserving event-counterfactual benchmark on recorded AV keyed to physical-commonsense benchmarks (Cui et al., 2026; Li et al., 2026)—is the most important direction this work leaves open. Three scope conditions are intrinsic rather than incidental. First, the intervention is decisive mainly under near-perfect (possibly slice-local) confounding; establishing that real AV corpora contain such slices is an empirical prerequisite, for which existing benchmarks provide encouraging but indirect evidence. Second, the shortcut is taken only when the causal cue is hard to read relative to the nuisance (Section F.2): where the event is visually obvious—as with a large, unambiguous motion—a direct model is already robust and our objective neither helps nor hurts. Which real AV events are visually subtle enough to be shortcut is an open empirical question; the near-zero cost of the objective, however, makes leaving it on the safe default. Third, and most important, the practical recipe assumes *every* confounding nuisance is *identifiable and augmentable*: Proposition 3 and its experiment show robustness is all-or-nothing, so missing even one confounder leaves a complete shortcut, and our unsupervised shared/private baseline shows that representation learning alone does not recover robustness. Discovering unknown nuisances and intervening on all of them—rather than assuming they are given—is, in our view, the central open problem for causally robust audio–video generation, and the most promising direction this work points to.

C Extended Related Work

Our contribution sits at the intersection of several literatures; we expand on each here and mark where we differ.

Joint audio–video and video-to-audio generation. Recent systems couple the two modalities either by direct cross-modal attention (Ruan et al., 2023; Xing et al., 2024) or, more recently, by aligning both to a shared spatio-temporal prior (Liu et al., 2026; HaCohen et al., 2026); video-to-audio models such as Diff-Foley (Luo et al., 2023) and MMAudio (Cheng et al., 2025) map video (and text) to audio through a latent diffusion / flow-matching backbone. Physical-commonsense benchmarks report that current joint and V2A models are sensitive to exactly the visual factors they should ignore (Cui et al., 2026; Li et al., 2026), providing external evidence that the shortcut we isolate occurs in practice. Unlike these works, which propose architectures or benchmarks, we give a causal account of *why* the coupling produces a shortcut and what provably removes it.

Shortcut learning and spurious correlation. Shortcut learning is a general failure of deep models (Geirhos et al., 2020); remedies include reweighting worst-case groups (Sagawa et al., 2020) and de-biasing from a deliberately biased reference model (Nam et al., 2020). These operate on classification with known or inferred group labels; we instead exploit the generative AV structure, where a *feasible* nuisance intervention (recolouring, restyling) replaces the need for group annotations.

Invariance and domain generalization. IRM (Arjovsky et al., 2019), V-REx (Krueger et al., 2021) and GroupDRO (Sagawa et al., 2020) seek predictors invariant across training environments; DomainBed (Gulrajani & Lopez-Paz, 2021) shows their gains over ERM are often fragile, and adversarial feature alignment (Ganin et al., 2016) is unstable. As our Appendix H comparison makes explicit, these methods trade one requirement—many decorrelating environments—for the one we use—a single known, faithful counterfactual—and the latter is available precisely in the AV setting we study.

Counterfactual invariance and causal representation learning. Counterfactual invariance (Veitch et al., 2021) and neural structural causal models (Yang et al., 2021) provide the conceptual backbone; the impossibility of unsupervised disentanglement without inductive bias (Locatello et al., 2019) explains why a shared/private factorization cannot, on its own, separate event from appearance. Our contribution is to instantiate counterfactual invariance for the specific audio–video common-cause SCM, to prove that the shared-latent remedy is insufficient there, and to identify the unknown-nuisance regime as the barrier to deployment.

D Full Proofs

Throughout, $(\mathrm{e}, s, \mathrm{n})$ are the latent factors, $x_{\mathrm{v}} = g_{\mathrm{v}}(\mathrm{e}, \mathrm{n})$, $\mu(\mathrm{e}, s) = \mathbb{E}[x_{\mathrm{a}} \mid \mathrm{e}, s, \mathrm{n}]$ (Assumption 1), and by Assumption 2 every measurable predictor factors as $f(x_{\mathrm{v}}, s) = \tilde{f}(\mathrm{e}, \mathrm{n}, s)$. We write $R_D(f) = \mathbb{E}_D \|f(x_{\mathrm{v}}, s) - x_{\mathrm{a}}\|^2$. P_{tr} denotes the confounded training law and P_{te} the counterfactual test law, in which $\mathrm{n} \perp (\mathrm{e}, s)$ and the (e, s) marginal is unchanged. e is discrete with finite support.

Lemma 1 (Risk decomposition)**.** *For any measurable f and any law D over $(\mathrm{e}, s, \mathrm{n})$,*

$$R_D(f) = \mathbb{E}_D \big\| \tilde{f}(\mathrm{e}, \mathrm{n}, s) - \mu(\mathrm{e}, s) \big\|^2 + \mathbb{E}_D \big[\operatorname{tr} \operatorname{Cov}(x_{\mathrm{a}} \mid \mathrm{e}, s) \big],$$

and the second term does not depend on f.

Proof. Write $\|\tilde{f} - x_{\mathrm{a}}\|^2 = \|\tilde{f} - \mu\|^2 + \|\mu - x_{\mathrm{a}}\|^2 + 2\langle \tilde{f} - \mu,\ \mu - x_{\mathrm{a}} \rangle$. Conditioning the cross term on $(\mathrm{e}, \mathrm{n}, s)$ and using $\mathbb{E}[x_{\mathrm{a}} \mid \mathrm{e}, \mathrm{n}, s] = \mu(\mathrm{e}, s)$ (Assumption 1) together with the $(\mathrm{e}, \mathrm{n}, s)$-measurability of $\tilde{f} - \mu$ gives $\mathbb{E}\langle \tilde{f} - \mu, \mu - x_{\mathrm{a}} \rangle = \mathbb{E}\langle \tilde{f} - \mu,\ \mu - \mathbb{E}[x_{\mathrm{a}} \mid \mathrm{e}, \mathrm{n}, s] \rangle = 0$. The remaining term $\mathbb{E}\|\mu - x_{\mathrm{a}}\|^2$ equals $\mathbb{E}[\operatorname{tr} \operatorname{Cov}(x_{\mathrm{a}} \mid \mathrm{e}, s)]$ by Assumption 1 and is free of f. □

Lemma 1 means minimizing risk is exactly matching μ, and that the causal predictor $f^\star = \mu(\mathrm{e}, s)$ is the Bayes predictor under *any* law.

Proof of Proposition 1. **(i)** Under perfect confounding $\mathrm{n} = \pi(\mathrm{e})$ P_{tr}-a.s. with π injective, we have $\pi^{-1}(\mathrm{n}) = \mathrm{e}$ P_{tr}-a.s., hence $f_{\mathrm{sc}} = \mu(\pi^{-1}(\mathrm{n}), s) = \mu(\mathrm{e}, s) = f^\star$ P_{tr}-a.s. Two predictors that agree a.s. have equal risk, so $R_{\mathrm{tr}}(f_{\mathrm{sc}}) = R_{\mathrm{tr}}(f^\star)$; by Lemma 1 this is the Bayes risk, so both are training-risk minimizers.

(ii) Under P_{te} set $\mathrm{e}' := \pi^{-1}(\mathrm{n})$. Since $\mathrm{n} \perp (\mathrm{e}, s)$ and π is a bijection onto its range, $\mathrm{e}' \perp (\mathrm{e}, s)$ and e' has the marginal law of e. Applying Lemma 1 to both predictors under P_{te}, the f-free terms cancel and

$$R_{\mathrm{te}}(f_{\mathrm{sc}}) - R_{\mathrm{te}}(f^\star) = \mathbb{E} \big\| \mu(\mathrm{e}', s) - \mu(\mathrm{e}, s) \big\|^2 - 0 = \mathbb{E} \big\| \mu(\mathrm{e}', s) - \mu(\mathrm{e}, s) \big\|^2 \ \geq 0.$$

If $\mu(\cdot, s)$ is non-constant on a set of positive measure, then since $\mathrm{e}' \perp \mathrm{e}$ with the same marginal, $\Pr[\mu(\mathrm{e}', s) \neq \mu(\mathrm{e}, s)] > 0$ and the gap is strictly positive. □

Proof of Proposition 2. Assume perfect confounding, so $\sigma(\mathrm{n}) = \sigma(\mathrm{e})$ under P_{tr}.

$z = \mathrm{n}$ is risk-optimal. Because n determines e, the predictor $(\mathrm{n}, s) \mapsto \mu(\pi^{-1}(\mathrm{n}), s)$ equals $\mu(\mathrm{e}, s)$ P_{tr}-a.s., so by Lemma 1 an encoder $E(x_{\mathrm{v}}) = \mathrm{n}$ with an appropriate audio decoder attains the Bayes training risk.

$z = \mathrm{n}$ is rate-minimal among risk-optimal encoders. Suppose an encoder $z = E(x_{\mathrm{v}})$ admits a decoder attaining the Bayes training risk; by Lemma 1 its decoder output equals $\mu(\mathrm{e}, s)$ a.s. If the event

is audio-identifiable on a positive-measure set of scenes—$\forall\, \mathrm{e} \neq \mathrm{e}'$, $P_s[\mu(\mathrm{e}, s) \neq \mu(\mathrm{e}', s)] > 0$—then the map $s \mapsto \mu(\mathrm{e}, s)$ pins down e, so z (which reproduces $\mu(\cdot, s)$ across scenes) determines e a.s., giving $I(z; x_\mathrm{v}) \geq H(\mathrm{e})$. Meanwhile $I(\mathrm{n}; x_\mathrm{v}) = H(\mathrm{n}) = H(\mathrm{e})$ because π is a bijection. Hence $z = \mathrm{n}$ attains the minimum achievable rate among risk-optimal encoders, i.e. the pair $(\text{risk}, \text{rate})$ at $z = \mathrm{n}$ is Pareto-optimal.

Conclusion. For any budget $\beta \geq H(\mathrm{e})$, the program $\min_E R_\mathrm{tr}$ s.t. $I(z; x_\mathrm{v}) \leq \beta$ admits $E(x_\mathrm{v}) = \mathrm{n}$ as a minimizer; no such objective can prefer $z = \mathrm{e}$ over $z = \mathrm{n}$, since the two are indistinguishable in both risk and rate. The predictor induced by $z = \mathrm{n}$ is exactly f_sc, whose OOD excess risk is strictly positive by Proposition 1. Finally, for a shared/private split $x_\mathrm{v} \mapsto (z_\mathrm{s}, z_\mathrm{v})$ with audio decoded from z_s only, the audio term of the objective is exactly the program above, and the video reconstruction term can be satisfied by z_v; hence $z_\mathrm{s} = \mathrm{n}$ again lies among the minimizers. □

Proof of Theorem 1. **(i) Identification.** Let $f \in \mathcal{F}_\mathrm{inv}$. By definition $\tilde{f}(\mathrm{e}, \mathrm{n}, s)$ does not depend on n, so $f = h(\mathrm{e}, s)$ for some measurable h. By Lemma 1, $R_\mathrm{tr}(f) = \mathbb{E}_\mathrm{tr}\|h(\mathrm{e}, s) - \mu(\mathrm{e}, s)\|^2 + \mathrm{const}$, which is minimized if and only if $h = \mu$ P_tr-a.s.; hence $\arg\min_{\mathcal{F}_\mathrm{inv}} R_\mathrm{tr} = f^\star$. For the second claim, Lemma 1 applied under P_te gives, for *any* measurable g, $R_\mathrm{te}(g) \geq \mathrm{const}$ with equality iff $\tilde{g} = \mu$ a.s.; $f^\star$ attains it, so $f^\star$ is the global R_te-minimizer.

(ii) Gap closure. For $f = h(\mathrm{e}, s) \in \mathcal{F}_\mathrm{inv}$, Lemma 1 gives $R_D(f) = \mathbb{E}_{D(\mathrm{e},s)}\|h - \mu\|^2 + \mathbb{E}_D[\mathrm{tr}\,\mathrm{Cov}(x_\mathrm{a} \mid \mathrm{e}, s)]$, and both terms depend on D only through the (e, s) marginal, which P_tr and P_te share. Hence $R_\mathrm{tr}(f) = R_\mathrm{te}(f)$.

(iii) Necessity. By (i), any global R_te-minimizer g satisfies $\tilde{g} = \mu$ P_te-a.s. Since under P_te we have $\mathrm{n} \perp (\mathrm{e}, s)$ with n ranging over its full support, $\tilde{g}(\mathrm{e}, \mathrm{n}, s) = \mu(\mathrm{e}, s)$ for a.e. $(\mathrm{e}, \mathrm{n}, s)$, i.e. $\tilde{g}$ is a.s. constant in n; thus $g \in \mathcal{F}_\mathrm{inv}$ up to a null set, and $f^\star \in \mathcal{F}_\mathrm{inv}$. Restricting to $\mathcal{F}_\mathrm{inv}$ therefore excludes no optimal predictor.

Penalty equivalence. $\mathcal{L}_\mathrm{cf}(f) = \mathbb{E}_{\mathrm{n}'}\|\tilde{f}(\mathrm{e}, \mathrm{n}, s) - \tilde{f}(\mathrm{e}, \mathrm{n}', s)\|^2 = 0$ iff $\tilde{f}(\mathrm{e}, \cdot, s)$ is a.s. constant on the support of n, i.e. iff $f \in \mathcal{F}_\mathrm{inv}$. Consequently $\min_f R_\mathrm{tr}(f) + \lambda\mathcal{L}_\mathrm{cf}(f)$ is, in the limit of an active constraint, the program of (i), whose solution is $f^\star$ whenever $f^\star$ is realizable. □

Proof of Proposition 3. Let $S \subseteq \{1, \ldots, L\}$ and let f be invariant to n_S. By Assumption 2, $f = h(\mathrm{e}, \mathrm{n}_{S^c}, s)$. If $S^c = \emptyset$ then $f = h(\mathrm{e}, s) \in \mathcal{F}_\mathrm{inv}$ and Theorem 1(i) gives zero excess OOD risk at the optimum. Otherwise pick any $l \in S^c$; by hypothesis $\mathrm{n}_l = \pi_l(\mathrm{e})$ P_tr-a.s. with π_l injective, so the predictor $h_\mathrm{sc}(\mathrm{e}, \mathrm{n}_{S^c}, s) := \mu(\pi_l^{-1}(\mathrm{n}_l), s)$ is invariant to n_S, agrees with $f^\star$ P_tr-a.s., and therefore attains the Bayes training risk—so it lies in the constrained argmin. Applying Proposition 1 with the reduced nuisance n_l gives OOD excess risk $\mathbb{E}\|\mu(\mathrm{e}', s) - \mu(\mathrm{e}, s)\|^2 > 0$ with $\mathrm{e}' = \pi_l^{-1}(\mathrm{n}_l) \perp \mathrm{e}$. Hence the constrained program admits a shortcut solution with strictly positive OOD excess risk whenever some confounded nuisance is left un-augmented, and zero excess risk exactly when S contains every confounded nuisance. □

Remark 1 (Extension to conditional generative distributions)**.** *Theorem 1 is stated for conditional-*mean *prediction under squared error, but the identification argument does not depend on that choice. Let the model output a conditional density $q(\cdot \mid x_\mathrm{v}, s)$ and be scored by a strictly proper scoring rule $S(q, x_\mathrm{a})$, e.g. the log score. By strict propriety the population risk $\mathbb{E}[S(q(\cdot \mid x_\mathrm{v}, s), x_\mathrm{a})]$ is uniquely minimized at the true conditional $q^\star(\cdot \mid x_\mathrm{v}, s) = p(x_\mathrm{a} \mid \mathrm{e}, \mathrm{n}, s)$, which by Assumption 1 equals $p(x_\mathrm{a} \mid \mathrm{e}, s)$ and is therefore* n*-invariant. Restricting to the counterfactually invariant class $\mathcal{F}_\mathrm{inv}$ (now defined on the conditional law $q(\cdot \mid g_\mathrm{v}(\mathrm{e}, \mathrm{n}), s)$) thus loses no optimality and closes the train/test gap, exactly as in Theorem 1(i)–(iii); the counterfactual-consistency penalty becomes a divergence between $q(\cdot \mid x_\mathrm{v}, s)$ and $q(\cdot \mid x'_\mathrm{v}, s)$. We use this only to justify the framing; the generative results (Table 5) are reported empirically.*

E Experimental Details

All experiments run on a single NVIDIA RTX A6000; every run completes in minutes. Optimizer is Adam throughout. Unless stated otherwise the counterfactual weight is $\lambda = 1$ and the training confounding is $\rho_\mathrm{spur} = 1$.

E.1 Feature-vector SCM (Tables 1, 2, 5, 10; Figures 3, 7).

K=8 events, M=8 nuisance values, video dimension d_{v}=32, audio dimension d_{a}=16. The video is $x_{\mathrm{v}} = [\,\gamma \cdot \mathrm{onehot(e)}\ ;\ \mathrm{onehot(n)}\,]W_{\mathrm{v}} + 0.05\,\epsilon$ with a fixed random $W_{\mathrm{v}} \in \mathbb{R}^{(K+M)\times d_{\mathrm{v}}}$ scaled by $1/\sqrt{K+M}$ and *geometry strength* γ=0.6: the causal event signal is deliberately weaker than the clean one-hot nuisance, which makes the shortcut the path of least resistance. The audio is $x_{\mathrm{a}} = (\mathrm{onehot(e)}W_{\mathrm{a}}) \cdot (1/s) + 0.03\,\epsilon$ with $W_{\mathrm{a}} \in \mathbb{R}^{K\times d_{\mathrm{a}}}$ and scene distance $s \sim \mathcal{U}[0.5, 1.5]$ (so the scene sets loudness, and audio depends on (e, s) only). Training draws $\mathrm{n} = \mathrm{e} \bmod M$ with probability ρ_{spur} and uniform otherwise; the OOD-cf test draws n uniformly, independent of e. *Direct*: MLP $[d_{\mathrm{v}}{+}1 \to 256 \to 256 \to d_{\mathrm{a}}]$ with SiLU. *CC-bottleneck / CC+CF*: encoder $[d_{\mathrm{v}} \to 256 \to z]$ with z=6, decoder $[z{+}1 \to 256 \to d_{\mathrm{a}}]$. Training: lr 2×10^{-3}, batch 256, 3000 steps. Evaluation: 4000 fresh samples. The counterfactual batch resamples n while holding (e, s) fixed.

E.2 Slice-local confounding (Table 2).

Same SCM; the first 4 of 8 event classes are perfectly confounded ($\mathrm{n} = \mathrm{e}$), the remaining classes always receive a uniform nuisance. The test decouples all classes; we report the median per-sample error separately over the confounded and clean event classes (8000 evaluation samples).

E.3 Procedural pixel video (Table 8).

32×32 RGB. Each of K=8 events is a fixed template of 2–4 Gaussian blobs (σ=0.16) at event-specific offsets, rendered with random global translation ($\pm$0.3) and scale (0.8–1.2); the nuisance is one of M=8 RGB colors multiplying the grayscale render. Audio is a per-event 16-d spectrum scaled by $1/s$. Encoder: three stride-2 convolutions (3→32→64→64, kernel 4) followed by a linear map. *Direct* uses a 128-d code, *bottleneck* a 6-d code, *disentangle* z_{s}=6 shared and z_{v}=32 private with a mirrored transposed-convolution image decoder trained on reconstruction plus the audio loss from z_{s} alone. *CFaug* re-renders the same grayscale with a random palette color and penalizes the audio discrepancy. Training: lr 10^{-3}, batch 128, 3000 steps; evaluation on 3000 samples.

E.4 Temporal video (Table 9).

K=8 motion directions (angles $2\pi k/8$), T=8 frames, 32×32 RGB. A Gaussian blob (σ=0.13) starts at a uniformly random position in $[-0.35, 0.35]^2$ and is displaced by $1.1 \cdot (t/T)$ along its direction, so the *start position carries no information about the event* and only the cross-frame displacement identifies it; the color (one of 8) is constant within a clip and confounded with the direction. Audio is a per-direction 16-d spectrum. The model stacks time on channels ($3T$=24) and applies three stride-2 convolutions (24→32→64→64) plus a linear head. Training: lr 10^{-3}, batch 128, 4000 steps. The counterfactual re-renders the same clip (same motion and start) in a random color. Motion-accuracy is nearest-neighbour classification of the predicted audio against the 8 direction spectra (2000 evaluation samples).

E.5 Real modalities: Colored-MNIST with spectrogram audio (Table 3).

Images: the 60,000 MNIST training digits, tinted by one of 10 palette colors, bilinearly resized to 64×64 and ImageNet-normalized. Audio: for digit d the tone has fundamental $f_0 = 220 \cdot 2^{d/12}$ Hz with five harmonics of amplitude $1/h$ and envelope $e^{-3t/T}$, T=0.4 s at 16 kHz; the target is $\log(1 + \mathrm{Mel})$ with n_{fft}=512, hop 256, 32 mel bands, flattened to 832 dimensions. Encoder: torchvision ResNet-18 with ImageNet weights, final layer removed, **frozen** (only the head is trained), giving 512-d features. Head: MLP $[512 \to 512 \to 512 \to 832]$ with SiLU, lr 10^{-3}, batch 128, 4000 steps. The counterfactual re-tints the same digit with a random palette color. Event-accuracy is nearest-neighbour classification of the predicted spectrogram against the 10 digit tones (3000 evaluation samples; chance = 0.1).

E.6 Conditional generation (Table 5).

Conditional flow matching (Lipman et al., 2023) on the SCM of Appendix E.1. The velocity field takes $(x_{\mathrm{a}}, t, E(x_{\mathrm{v}}), s)$ with a video encoder $[d_{\mathrm{v}} \to 256 \to 64]$ and a trunk $[d_{\mathrm{a}}{+}1{+}64{+}1 \to 256 \to 256 \to d_{\mathrm{a}}]$. Training samples $x_0 \sim \mathcal{N}(0, I)$, $x_1 = x_{\mathrm{a}}$, $t \sim \mathcal{U}[0,1]$, $x_t = (1-t)x_0 + tx_1$, and regresses the target velocity $x_1 - x_0$; lr 2×10^{-3}, batch 256, 4000 steps. The counterfactual term matches the velocity fields conditioned on the original and on a nuisance-resampled video. Sampling integrates the ODE with 20 Euler steps. Event-accuracy compares the *sampled* audio to the K event spectra scaled by the sample's $1/s$.

E.7 Mechanistic probe (Figure 7, left).

For each trained model we capture the input to its final linear layer—the audio-producing representation—via a forward hook, on 4000 samples drawn with n *decoupled* from e (so that event and nuisance are separable). We then fit a linear classifier (Adam, lr 10^{-2}, 800 steps) on one half of the samples to predict the event, and, separately, the nuisance, and report accuracy on the held-out half. Chance is $1/8 = 0.125$.

E.8 Multiple nuisances (Figure 7, right).

$L{=}4$ nuisances, each with $M{=}8$ values and each perfectly confounded with the event in training; video dimension 48 with $W_{\mathrm{v}} \in \mathbb{R}^{(K+LM)\times d_{\mathrm{v}}}$. The intervention resamples the first k nuisances ($k = 0, \ldots, 4$) while holding the event and the remaining nuisances fixed. Everything else follows Appendix E.1.

E.9 Continuous nuisance (Appendix G).

The nuisance is a continuous vector $\mathrm{n} \in \mathbb{R}^8$ drawn as $\mathrm{emb}(\mathrm{e}) + 0.25\,\epsilon$ in training, with a fixed random $\mathrm{emb} \in \mathbb{R}^{K\times 8}$ scaled by 1.5; at test the embedding index is drawn independently of the event. The counterfactual resamples the continuous nuisance for a fresh, independently drawn event index.

E.10 Shared-prior baseline and capacity sweep (Appendix G).

The shared-prior model encodes $z_{\mathrm{s}} = E(x_{\mathrm{v}}) \in \mathbb{R}^8$ and trains $\|\mathrm{aud}(z_{\mathrm{s}}, s) - x_{\mathrm{a}}\|^2 + \beta\|\mathrm{vid}(z_{\mathrm{s}}) - x_{\mathrm{v}}\|^2$ with $\beta{=}1$, so that both modalities are aligned to the single shared code. The capacity sweep varies the hidden width of the *direct* model over $\{32, 64, 128, 256, 512, 1024\}$ (2.7K–1.1M parameters) with all else fixed.

E.11 Baseline study and label scarcity (Table 6, Figure 5).

All five variants share one architecture with an explicit encoder/head split ($[d_{\mathrm{v}}{+}1 \to 256 \to 256]$ encoder, linear head to d_{a}) so that representation-level consistency is well defined; only the loss differs. In the label-scarce runs a Bernoulli(p) mask marks which samples carry an audio label: the supervised terms (including augmentation's counterfactual term) are averaged over labeled samples only, while the consistency term—which never references x_{a}—is applied to the whole batch. This is exactly the asymmetry the experiment measures.

Reproducibility and error bars. All worlds, model initializations, and training batches are generated from fixed seeds. We report mean $\pm$ std over independent training seeds wherever a comparison could plausibly be seed-sensitive: the baseline study (Table 6, 5 seeds), the cue and label-scarcity sweeps (Figures 6, 5, 5/3 seeds), the emergence and DG baselines, the representation-depth and partially-informative-nuisance ablations (Appendix H, 3 seeds), and the two headline real-modality tables (Colored-MNIST Table 3 and pixel Table 8, 3 seeds). We give per-quantity std where a table has few columns and a summary bound ("std $\leq x$") where a table is wide, purely for readability; the underlying runs are the same. The conditional-generation (Table 5), temporal-video (Table 9) and moving-digit (Table 7) tables are likewise reported over 3 seeds; their between-method gaps are one-to-three orders of magnitude (e.g. OOD accuracy 0.13–0.21 vs. 1.00) and the measured std

is far below those gaps, so the qualitative conclusion is seed-independent. Finally, the MMAudio fine-tune (Section 6.6) is now run with 3 independent optimization seeds and evaluated, together with the frozen baseline, on a common held-out set (12 videos $\times 2$ seeds) so that the before/after comparison is strictly matched; we report mean$\pm$std and, honestly, find that the intervention stably reduces the targeted grayscale sensitivity but broadly dampens input sensitivity rather than isolating an appearance shortcut. Other reported numbers are medians of per-sample errors or means over the evaluation set, as indicated in each caption.

F Experiments Moved from the Main Text for Space

These experiments support the main-text narrative but are placed here for space. They use the same settings as Section 6.

F.1 Consistency vs. supervised augmentation (full results)

If one can synthesize the counterfactual $x'_{\mathrm{v}} = \mathrm{do}(\mathrm{n} := \mathrm{n}')$ and one knows it carries the same audio x_{a}, the obvious alternative is to simply *train on it with supervision*: $\mathcal{L}_{\mathrm{aug}} = \|f(x_{\mathrm{v}}, s) - x_{\mathrm{a}}\|^2 + \|f(x'_{\mathrm{v}}, s) - x_{\mathrm{a}}\|^2$. This baseline is the sharpest test of our objective, and we report it honestly. Table 6 compares five variants over five seeds.

Table 6: Is the counterfactual *consistency* objective necessary, or does supervised counterfactual *augmentation* suffice? Feature-vector SCM, ρ_{spur}=1, mean $\pm$ std over 5 seeds. With full labels the two are statistically indistinguishable—the essential ingredient is the intervention, not the loss form. Representation-level consistency, by contrast, is *not* sufficient.

Method	IID MSE	OOD-cf MSE
Direct (no intervention)	0.0010 ± 0.0000	0.1827 ± 0.0029
Supervised CF augmentation	0.0011 ± 0.0000	0.0012 ± 0.0000
CF consistency, prediction-level (ours)	0.0010 ± 0.0000	$\mathbf{0.0011 \pm 0.0000}$
CF consistency, representation-level	0.0012 ± 0.0000	0.0140 ± 0.0007
Augmentation + consistency	0.0010 ± 0.0000	$\mathbf{0.0011 \pm 0.0000}$

Two findings follow, and the first is a concession. **When audio labels are available for every sample, supervised counterfactual augmentation matches our consistency objective exactly** (0.0012 vs. 0.0011); the two are interchangeable. We therefore do not claim the consistency *loss* as the contribution. What the comparison does establish is the paper's actual thesis: every method that *intervenes on the nuisance* closes the gap (~ 0.001), while every method that merely reshapes the representation—the bottleneck, the unsupervised shared/private factorization, the faithful shared-prior model, and, as the table shows, representation-level consistency (0.0140, $12\times$ worse)—does not. The dividing line runs between intervention and architecture, not between two losses.

Second, the two are *not* interchangeable once labels are scarce (Figure 5). The consistency term is *label-free*: it constrains $f(x_{\mathrm{v}})$ against $f(x'_{\mathrm{v}})$ without ever needing x_{a}, so it can be applied to unlabeled video, whereas augmentation can only use counterfactuals of *labeled* samples. As the labeled fraction falls from 1.0 to 0.02, augmentation degrades by $4.6\times$ ($0.0012 \to 0.0055$) while consistency degrades by $3.3\times$ ($0.0011 \to 0.0036$), a consistent $\sim 1.5\times$ advantage at every scarce setting with non-overlapping error bars. This is the regime real AV data actually lives in— abundant unlabeled video, few clean paired examples—and it is the one place where the choice of loss, rather than the presence of the intervention, matters.

F.2 When is the shortcut taken? Cue difficulty decides

A shortcut is only attractive if it is *easier* than the causal path, so the phenomenon should not be unconditional—and an honest account must say when it occurs. We sweep the strength γ of the causal event signal in the video while holding the nuisance a clean one-hot (Figure 6; 5 seeds, dense sampling through the transition). The direct model's in-distribution error is flat and near zero throughout (0.0010 ± 0.0000), but its counterfactual error falls monotonically and smoothly

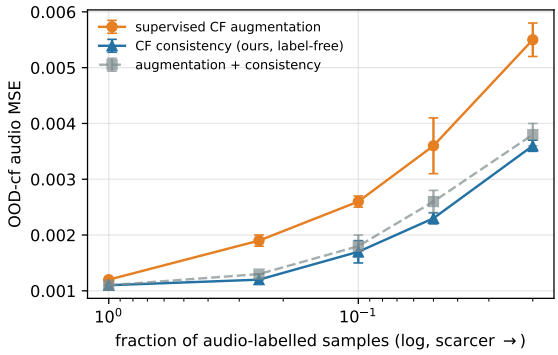

Figure 5: Supervised counterfactual augmentation and our consistency objective are interchangeable when every sample is labeled, but the consistency term needs no audio label for the counterfactual and therefore keeps improving from unlabeled video as labels become scarce (3 seeds, ±std).

from 0.2632 ± 0.0006 at γ=0.2 to 0.0027 ± 0.0001 at γ=4— a 263× penalty shrinking to 2.7×: once the event is easy to read, the model simply reads it and there is no shortcut to block. The transition ($\gamma \in [0.8, 2.0]$) is smooth and stable across seeds (std $\approx 1\%$ of the mean), not a knife-edge. Counterfactual consistency is flat and robust across the whole sweep and—the practically important point—costs almost nothing where it is unnecessary (0.0013 vs. 0.0010 at γ=4). It is therefore cheap insurance that can be left on by default.

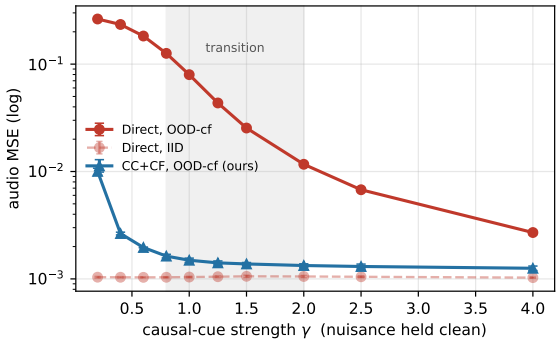

Figure 6: The shortcut is conditional. As the causal cue becomes easier to read ($\gamma \uparrow$) the direct model's counterfactual penalty collapses from 263× to 2.7×, smoothly and stably across seeds; the counterfactual objective is robust throughout and nearly free when unnecessary. Shaded: the transition region. (5 seeds, ±std; error bars are smaller than the markers.)

F.3 Real digits with a temporal cause: both regimes, as predicted

Section F.2 makes a falsifiable prediction, which we test in the most realistic setting we can build: a *real* MNIST digit sprite translating in one of eight directions (the event), tinted with a confounded colour (the nuisance), with the digit's identity irrelevant and its start position random—so only the motion identifies the event. Table 7 confirms both regimes. When the motion is easy (a large 0.85 travel, no noise), the direct model reads it and is *already robust* (1.00 OOD accuracy): there is no shortcut and nothing to fix. When the motion is made hard (travel 0.30, frame noise 0.35), the direct model switches to the colour and collapses to near chance (0.21, chance = 0.125), while counterfactual consistency stays at 0.995. The shortcut is thus neither universal nor contrived: it appears exactly where the theory says it should, on real pixels with a genuinely temporal cause.

Table 7: Moving-MNIST (real digit sprites, event = motion direction, nuisance = colour, chance = 0.125). The shortcut appears only when the causal motion cue is hard, exactly as the cue-difficulty sweep predicts; the counterfactual objective is robust in both regimes. Mean±std over 3 seeds (OOD columns).

		IID		OOD-cf	
Motion cue	Model	MSE	acc	MSE	acc
easy (travel 0.85)	Direct	0.0014	1.00	0.006±0.002	1.00±0.00
	CC+CF (ours)	0.0009	1.00	0.001±0.000	1.00±0.00
hard (travel 0.30, noise 0.35)	Direct	0.0015	1.00	1.180±0.141	**0.209±0.033**
	CC+CF (ours)	0.0049	1.00	0.009±0.002	**0.995±0.002**

F.4 Mechanistic probe: the fix suppresses the nuisance in the audio code

Why do the models differ? We linearly probe the audio-producing representation (the input to the final audio layer) for event vs. material information, on decoupled data where the two are separable (Figure 7). In the direct and bottleneck models the material is fully decodable from the audio representation (accuracy 1.0)—the sound is being read off the nuisance. Counterfactual consistency drives material-decodability down sharply, close to chance (0.24 vs. chance 0.125; not to zero) while keeping event-decodability at 1.0: the representation feeding audio is reorganized around the cause, though a residual trace of the nuisance remains. This is the mechanism behind the OOD gap. This reorganization holds specifically at the *output* representation; applying the same consistency constraint at earlier layers removes the shortcut only partially (Appendix H.4).

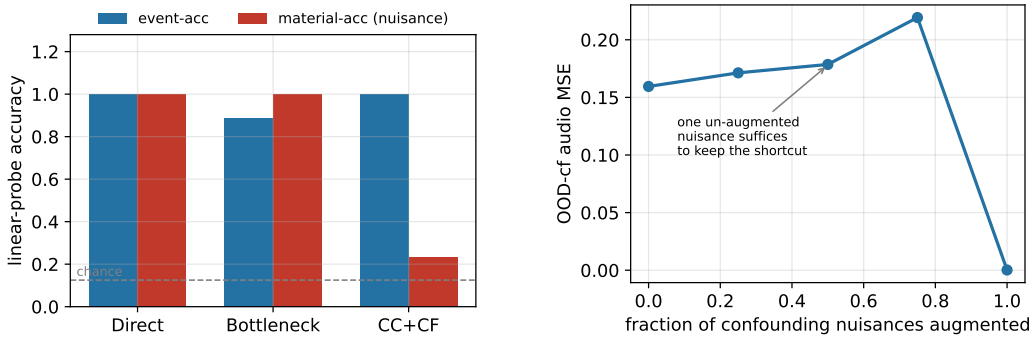

Figure 7: **Left:** linear-probe accuracy of the audio-producing representation. Direct/bottleneck leak the nuisance (high material-accuracy); CC+CF suppresses it (to near, not exactly, chance). **Right:** with multiple perfectly-confounded nuisances, robustness is all-or-nothing—any single un-augmented nuisance keeps a complete shortcut, so OOD error only collapses once *every* confounder is intervened on (experiment of Section 6, matching Proposition 3).

F.5 Multiple nuisances: robustness is all-or-nothing

Real data confounds through several appearance factors at once, and one may be able to augment only some of them. With $L{=}4$ nuisances, each perfectly confounded with the event, we vary the fraction on which the intervention is applied (Figure 7, right). Because each nuisance individually determines the event, any single *un*-augmented nuisance is a complete shortcut: OOD error stays high (indeed slightly worsens, as the model concentrates on the remaining shortcuts) until *all* confounders are covered, at which point it collapses to near zero. This is exactly Proposition 3: partial augmentation buys nothing when nuisances are individually sufficient. This all-or-nothing behaviour is specific to that regime; when each nuisance is only partially predictive of the event, partial augmentation already helps and the response is graded (Appendix H.5). Either way it sharp-

ens the open problem—to be safe one must identify and intervene on the confounding factors, and knowing which they are is exactly what one lacks in the wild.

F.6 The mechanism survives real pixels; architecture alone does not

Table 8 moves to procedurally rendered pixel video: the event is a shape arrangement (the causal signal), the nuisance is color, color is perfectly confounded with the shape in training and decoupled at test, and audio is the event-determined target. On genuine pixels the direct model exploits color and its OOD-cf error explodes by $100\times$; the bottleneck and the unsupervised shared/private disentanglement fail just as hard, confirming that architecture alone does not block the shortcut. Only the counterfactual intervention—here a random-recolor augmentation, a feasible real-world counterfactual—remains robust (OOD-cf $\approx$ IID), while fitting the IID data as well as the direct model.

Table 8: Procedural pixel audio–video (32×32 RGB, color = nuisance, perfectly confounded). Audio MSE, mean over 3 seeds (std ≤ 0.001 IID, ≤ 0.055 OOD). Only the counterfactual intervention (recolor augmentation) blocks the shortcut; a bottleneck and an unsupervised shared/private factorization do not.

Model	IID	OOD-cf	OOD/IID
Direct	0.010	1.042	105×
Bottleneck	0.011	1.155	107×
Disentangle (shared/priv)	0.010	1.119	109×
CC+CF, recolor (ours)	0.010	**0.010**	**1.0**×

F.7 Temporal video: when the causal cue is motion

All settings so far treat a frame; but in video the causal event is often *temporal*. We therefore make the event a *motion*: a colored blob travels in one of $K=8$ directions over $T=8$ frames, and the audio is determined by that direction. The blob's start position is randomized, so no single frame reveals the direction—only the cross-frame displacement does—while the color is visible in *every* frame and is confounded with the direction in training (Figure 8). The shortcut is therefore "read the color from any frame" and the causal path requires integrating motion. Table 9 shows the direct model's motion-accuracy falling from 1.00 to 0.52 and its error exploding by $2000\times$ under the counterfactual color shift, whereas counterfactual consistency—which forces the model onto the temporal cue—is exactly robust (1.00 accuracy, error unchanged).

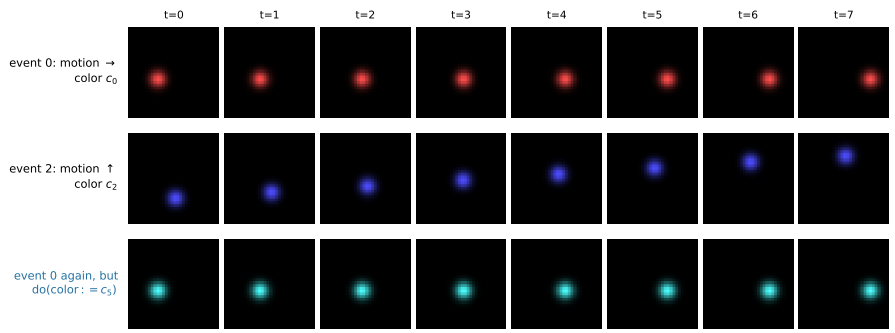

Figure 8: The temporal setting. The event is the *motion*, which is identifiable only across frames because the start position is random; the nuisance colour is visible in every frame and is therefore the tempting shortcut. The third row is the counterfactual of the first: identical motion, resampled colour.

Table 9: Temporal video: event = motion direction, nuisance = color (ρ_{spur}=1, chance accuracy = 0.125). Only the cross-frame motion identifies the event; color is a per-frame shortcut. Counterfactual consistency forces the model onto the temporal cue. Mean±std over 3 seeds.

	IID		OOD-cf	
Model	MSE	motion-acc	MSE	motion-acc
Direct	0.0003	1.00	0.664 ± 0.044	0.517 ± 0.025
CC+CF (ours)	0.0002	1.00	$\mathbf{0.0002 \pm 0.0001}$	$\mathbf{1.00 \pm 0.00}$

F.8 Summary of the experimental findings

Across a feature-vector SCM, slice-local confounding, procedural pixel video, real images with spectrogram audio, moving real digits, a conditional generator, and a mechanistic probe, one line separates what works from what does not, and it is not the line we might have hoped for. *Intervening* on the nuisance blocks the shortcut— whether the intervention is delivered as supervised counterfactual augmentation or as a prediction-level consistency penalty, which are interchangeable when labels are plentiful (Section 6.5). *Reshaping the representation* does not: the bottleneck, unsupervised shared/private disentanglement, a faithful shared-prior model, and representation-level consistency all leave the shortcut intact. Two conditions delimit when it matters, and both are empirical, not assumed: the training confounding must be near-perfect (possibly slice-local), or at least sufficiently strong—Table 11 shows a clear shortcut already at ρ=0.99 under a limited training budget—and the causal cue must be hard to read relative to the nuisance. Where those conditions fail, the direct model is already fine—and the objective costs almost nothing there, so it can be left on by default. Where they hold, it is the difference between synthesizing the right event's sound and the wrong one.

Qualitative and real-model figures. Figures 9 (the failure made audible on Colored-MNIST) and 10 (the real MMAudio generator's non-invariance) accompany the main-text discussion.

G Additional Robustness Experiments

All experiments here use the feature-vector SCM of Section 6 at ρ_{spur}=1.

Faithful shared-prior baseline. Table 10 gives the full numbers for the shared-prior stand-in discussed in Section 6: a single code z_s that generates audio and reconstructs video (both modalities aligned to it). It collapses under the counterfactual shift like the direct model, while the counterfactual objective is $\sim 76\times$ better.

Table 10: A faithful shared-prior model (audio generated from, and video reconstructed from, one shared code) does not block the shortcut.

Model	IID MSE	OOD-cf MSE
Direct	0.0010	0.1816
Shared-prior (JavisDiT-style)	0.0016	0.1515
CC+CF (ours)	0.0019	**0.0020**

Weight of the counterfactual term. Figure 11 (left) sweeps λ. Robustness appears as soon as the term is switched on and holds over two orders of magnitude ($\lambda \in [0.1, 10]$: OOD-cf MSE $0.0032 \to 0.0020$), while the in-distribution cost is negligible throughout ($0.0016 \to 0.0022$). The method requires no tuning.

Continuous nuisance. The nuisance need not be a discrete palette. With a *continuous* nuisance vector (an event-dependent embedding plus noise in training, independent at test), the direct model still takes the shortcut (OOD-cf MSE 0.241) and the counterfactual objective, whose intervention resamples the continuous nuisance, remains robust (0.0002).

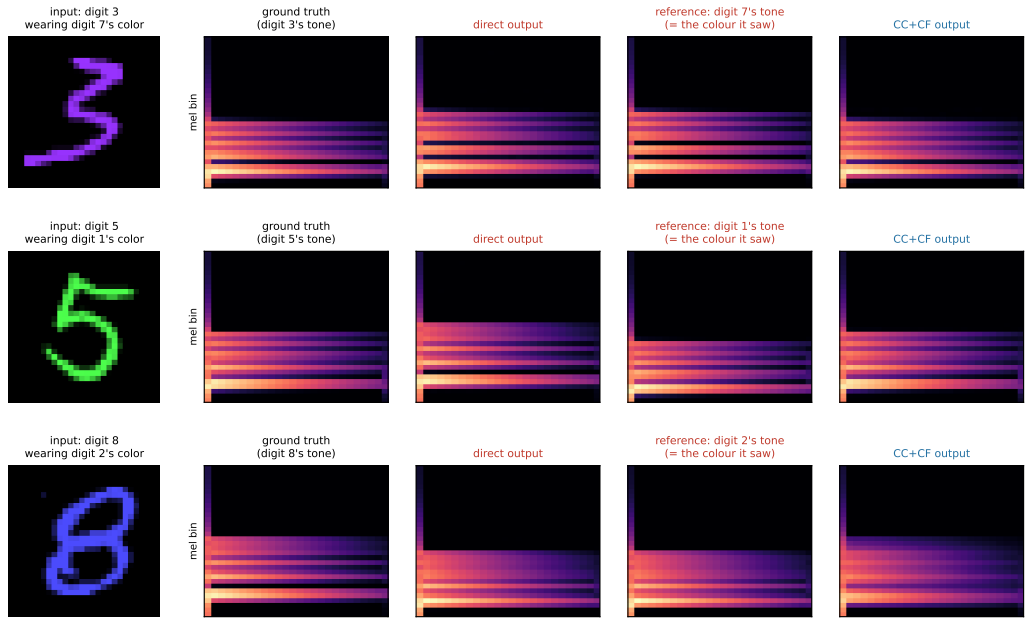

Figure 9: **The failure, made audible.** Each row shows a digit wearing another digit's colour. The direct model's generated spectrogram (col. 3) matches the tone of the digit whose *colour* it saw (col. 4)—not the ground truth of the digit it was actually shown (col. 2). Counterfactual consistency (col. 5) recovers the correct tone. The shortcut is not a small quantitative degradation: the model synthesizes the wrong event's sound.

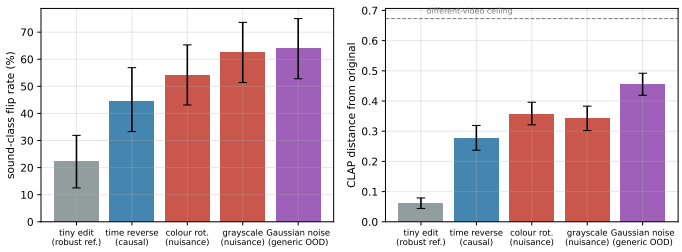

Figure 10: A real pretrained video-to-audio generator (MMAudio) is far from invariant to sound-irrelevant edits, but not specifically to appearance. Sound-class flip rate (left) and CLAP distance (right) of the generated audio under each input edit (24 real videos, 3 seeds; 95% bootstrap CIs). Appearance edits (grayscale, colour rotation; red) move the sound far above the minimal-edit reference—but a generic Gaussian-noise OOD control (purple), which carries no event information, moves it at least as much. The real model is broadly input-brittle, and—because the noise control is a larger, unmatched perturbation—this experiment does *not* establish colour-specific shortcutting. Isolating a colour-specific shortcut requires the controlled confounds of the synthetic studies.

Capacity does not fix the shortcut. Figure 11 (right) scales the direct model from 2.7K to 1.1M parameters ($400\times$). In-distribution error is flat and near zero throughout, and *the OOD-cf error does not improve at all* ($0.173 \to 0.184$). The shortcut is therefore not an underfitting artifact and cannot be scaled away—it is a property of what the training distribution identifies (Proposition 1), not of model capacity.

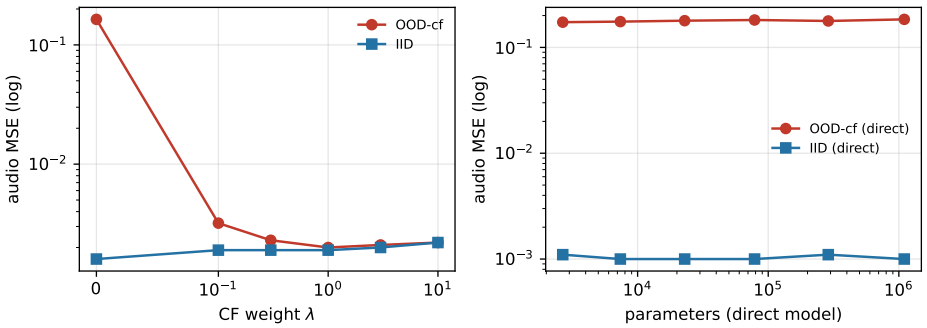

Figure 11: **Left:** the counterfactual term needs no tuning—robustness holds across two orders of magnitude of λ at negligible IID cost. **Right:** scaling the direct model $400\times$ leaves the shortcut entirely intact (OOD-cf flat) while IID error stays near zero.

H When the Shortcut Emerges, and Comparisons to Invariance Baselines

This appendix maps the conditions under which the shortcut becomes catastrophic and compares the counterfactual intervention against domain-generalization and invariance methods. All experiments use the feature-vector SCM; error bars, where shown, are over 3 seeds.

H.1 Emergence: confounding strength, training budget, and nuisance richness.

Proposition 1 proves non-identifiability at *perfect* confounding but does not, by itself, say the shortcut is absent below it. Table 11 and Figure 12 therefore vary the confounding ρ_{spur} against the training budget. The failure is *graded*: at $\rho_{\text{spur}} \leq 0.95$ more training drives the direct model toward the causal solution (the decorrelated examples eventually dominate), whereas at $\rho_{\text{spur}}{=}1$ more training pushes it *further* into the shortcut. "Near-perfect confounding" is thus where the failure becomes catastrophic *in this regime*, and we state it that way rather than as an unconditional law. Richer nuisances make the shortcut easier still: OOD-cf error rises from 0.003 at $M{=}2$ nuisance values to 0.13 at $M{=}8$.

Table 11: Emergence of the shortcut: direct-model OOD-cf MSE as a function of confounding ρ_{spur} and training budget. Below $\rho{=}1$ more training *removes* the shortcut; at $\rho{=}1$ it entrenches it.

ρ_{spur}	400 steps	1500 steps	6000 steps
0.80	0.0033	0.0018	0.0010
0.90	0.0049	0.0021	0.0011
0.95	0.0099	0.0028	0.0012
0.99	0.0519	0.0216	0.0015
1.00	0.1046	0.1131	0.1320

H.2 Long-tail events and partial coverage.

Two further slices confirm the mechanism is targeted, not diffuse. With Zipf-distributed event frequencies, rare (tail) events suffer at least as much as frequent ones (OOD-cf 0.127 vs. 0.117):

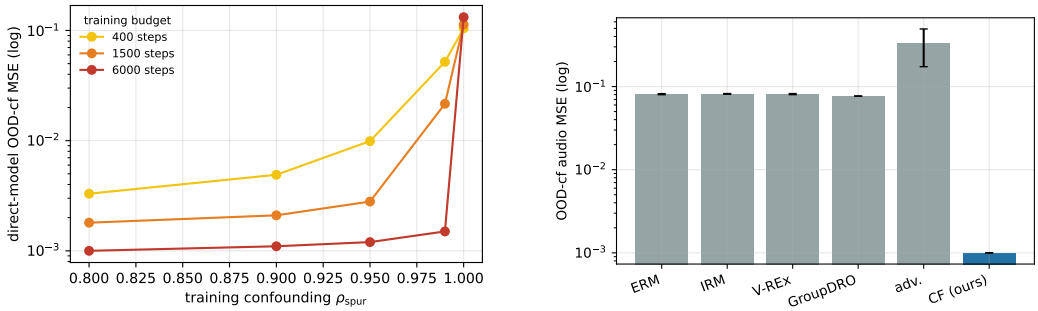

Figure 12: **Left:** the shortcut is graded in confounding and training budget—only at $\rho_{\text{spur}} \to 1$ does more training entrench it. **Right:** domain-generalization / invariance baselines (given two training environments) versus the counterfactual intervention (given none); log scale, $\pm$std over 3 seeds.

they have the fewest decorrelated examples to learn the causal path from. And when only *some* event classes are ever seen decorrelated in training, the shortcut is taken *exactly* on the classes that were never decorrelated (OOD-cf 0.049 on confounded-only classes vs. 0.0012 on classes seen decorrelated)—a class-level version of the slice-locality of Section 6.

H.3 Comparison to invariance / domain-generalization baselines.

A natural question is whether standard invariance methods already solve this. We compare IRM (Arjovsky et al., 2019), V-REx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2020), and adversarial nuisance removal (Ganin et al., 2016) against the counterfactual intervention (Table 12, Figure 12 right). Because these methods require *multiple environments* with varying spurious correlation, we grant them two training environments with different colour–event maps. (This multi-environment setup is a harder learning problem than the single confounded distribution used elsewhere, which is why the absolute IID errors here, ~ 0.08, sit above the ~ 0.001 of the other tables; the comparison of interest is within this table, across methods.) The counterfactual intervention is given no environment labels. With only two environments whose spurious correlations still overlap, IRM, V-REx, and GroupDRO barely improve on ERM (OOD-cf ≈ 0.08), and adversarial removal is unstable (0.33 ± 0.16), while the intervention reaches 0.0010. The lesson is not that these methods are bad—given enough diverse environments they can work—but that they trade one requirement (a known, faithful nuisance intervention) for another (many environments that decorrelate the nuisance), and the former is available precisely when one can construct the counterfactual.

Table 12: Invariance / domain-generalization baselines (two training environments) vs. the counterfactual intervention (no environments). Mean $\pm$ std over 3 seeds.

Method	IID MSE	OOD-cf MSE
ERM (direct)	0.0813 ± 0.0021	0.0810 ± 0.0012
IRM	0.0820 ± 0.0019	0.0816 ± 0.0008
V-REx	0.0813 ± 0.0021	0.0811 ± 0.0012
GroupDRO	0.0769 ± 0.0020	0.0768 ± 0.0004
Adversarial nuisance removal	0.3309 ± 0.1630	0.3329 ± 0.1596
CF intervention (ours)	0.0010 ± 0.0000	$\mathbf{0.0010 \pm 0.0000}$

Baseline implementation details. For reproducibility we specify the baselines exactly. *Environments.* All DG baselines receive two training environments that share the (e, s) structure but differ in the colour–event map: environment 1 uses $\mathrm{n} = \mathrm{e} \bmod M$ and environment 2 a fixed derangement $\mathrm{n} = \sigma(\mathrm{e})$ with $\sigma(\mathrm{e}) \neq \mathrm{e}$, so the spurious correlation is present in both but not identical; the event marginal is matched across environments. All methods share the same backbone (the encoder of Appendix E) and are trained for the same number of steps with Adam at the same base learning rate as the direct model. *IRM* (Arjovsky et al., 2019): penalty weight $\lambda_{\text{IRM}}{=}10^4$ after a

500-step ERM warm-up (swept $\{10^2, 10^3, 10^4, 10^5\}$; the reported value is the best OOD-cf). *V-REx* (Krueger et al., 2021): variance penalty β=10 on the per-environment risks, same warm-up (swept $\{1, 10, 100\}$). *GroupDRO* (Sagawa et al., 2020): groups are the two environments, group-weight step size η_q=0.01. *Adversarial nuisance removal* (Ganin et al., 2016): a gradient-reversal head predicts n from the audio representation with adversary weight α=1.0 (its high variance in Table 12 reflects the well-known instability of adversarial training at this scale, not a tuning artefact—it was the best of $\alpha \in \{0.1, 1, 10\}$). None of these methods is given the counterfactual pairing; the CF intervention is given no environment labels.

H.4 Where to apply the consistency constraint.

Our claim that reshaping the representation is insufficient should not rest on a single layer, so we apply the consistency penalty at every depth (Table 13). The effect generally improves toward later layers: an input-level penalty does nothing, penalties at intermediate layers help only partially (and not strictly monotonically— layer 2 is slightly worse than layer 1), and only an *output*-level (prediction) penalty fully removes the shortcut. Consistency deep in the network still leaves a usable nuisance trace in the later layers.

Table 13: Consistency applied at increasing depth (mean $\pm$ std, 3 seeds). Only output-level (prediction) consistency fully blocks the shortcut.

Consistency at	IID MSE	OOD-cf MSE
none / input	0.0010 ± 0.0000	0.1836 ± 0.0018
layer 1	0.0015 ± 0.0000	0.0827 ± 0.0030
layer 2	0.0016 ± 0.0001	0.0979 ± 0.0021
layer 3	0.0011 ± 0.0000	0.0018 ± 0.0001
output	0.0010 ± 0.0000	$\mathbf{0.0010 \pm 0.0000}$

H.5 Partially-informative nuisances: the all-or-nothing result is conditional.

Proposition 3 and its experiment assumed each nuisance *individually* determines the event, which is what makes partial augmentation useless there. When instead each of L=3 nuisances predicts the event only with probability $q < 1$, no single one is a sufficient shortcut, and augmenting a subset already helps: the response becomes graded (Table 14). The all-or-nothing behaviour is therefore a property of the individually-sufficient regime, not a universal law.

Table 14: Partially-informative nuisances: OOD-cf MSE as a function of how many of L=3 nuisances are augmented, for different per-nuisance predictiveness q. At q=1 the row is all-or-nothing; for $q < 1$ partial augmentation already helps.

q	augment 0/3	1/3	2/3	3/3
1.00	0.1597	0.1710	0.2005	0.0001
0.80	0.0040	0.0005	0.0002	0.0001
0.60	0.0006	0.0002	0.0001	0.0001

AI Use Statement

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.