BEYOND FEASIBILITY: MINIMAL-SHIFT MECHANISM SELECTION FOR COUNTERFACTUAL IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual image editing asks how a given image would change under an intervention on a semantic attribute, accounting for causal effects while preserving unaffected attributes. Observational data and a causal graph cannot generally determine this change uniquely, as observationally equivalent models may imply different counterfactual images. Existing methods characterize admissible solutions within partial identification bounds but leave their selection unresolved, and their implementations can require full generative training or image–factor invertibility. To address these limitations, we propose Minimal-Shift Mechanism Selection (MSMS), which uses a large pretrained generator as additional evidence for ranking counterfactually consistent mechanisms. MSMS favors candidates whose editing responses most closely match those of the pretrained model, subject to preserving unaffected attributes. We formulate this comparison in a shared space across heterogeneous architectures and establish sufficient conditions for L_3-consistency under realizability and reference calibration. By separating factor abduction from image rendering, MSMS reuses pretrained rendering capabilities without requiring an invertible mapping between factors and images, supporting diffusion models and GANs. Experiments across three benchmarks demonstrate MSMS's ability to perform targeted edits while preserving non-descendant attributes. We further examine how pretrained responses guide the selection of candidate causal models.

1 INTRODUCTION

Counterfactual editing aims to determine how a factual observation, such as a high-dimensional image, would appear under an intervention on a specific semantic attribute. In a structural causal model (SCM), a valid edit corresponds to a do-intervention on a target factor that correctly modifies the intervened node, preserves non-descendants, and propagates changes to descendants according to the underlying counterfactual world (Pearl, 2009; Peters et al., 2017; Pan & Bareinboim, 2024). Causal generative methods span SCM-grounded families (Pawlowski et al., 2020; Yang et al., 2021; Sauer & Geiger, 2021; De Sousa Ribeiro et al., 2023), including diffusion-based mechanisms (Sanchez & Tsaftaris, 2022; Chao et al., 2024; Komanduri et al., 2024), and pretrained-backbone adaptations (Pan & Bareinboim, 2025). Their evaluation considers counterfactual effectiveness alongside image quality (Melistas et al., 2024). A related line of work emphasizes minimal changes to the source image through diffusion guidance and abductive inference (Jeanneret et al., 2022; Song et al., 2024).

A fundamental challenge, however, is determining which proxy causal model should be used for editing. In practice, the true data-generating SCM is inaccessible, so one must rely on observational data. By the Causal Hierarchy Theorem, observational agreement alone is insufficient to identify counterfactual behavior (Bareinboim et al., 2020). Neural causal models (NCMs) provide expressive structural proxy classes (Xia et al., 2021; 2023), but this expressiveness does not resolve the underlying non-identifiability at the counterfactual layer. Even given a known causal topology and a perfectly matched observational distribution, multiple graph-compatible proxy models can yield qualitatively different counterfactual distributions (Shpitser & Pearl, 2008). For nonidentifiable image queries, Pan & Bareinboim (2024) introduce *counterfactual consistency*, characterizing admissible counterfactual distributions through optimal bounds induced by the graph and observational data, following the partial-identification perspective (Manski, 2003). Pan & Bareinboim (2025) extend this guarantee to editing with a pretrained generator. This consistency criterion establishes admissibility while leaving the choice among observationally feasible proxies unresolved.

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

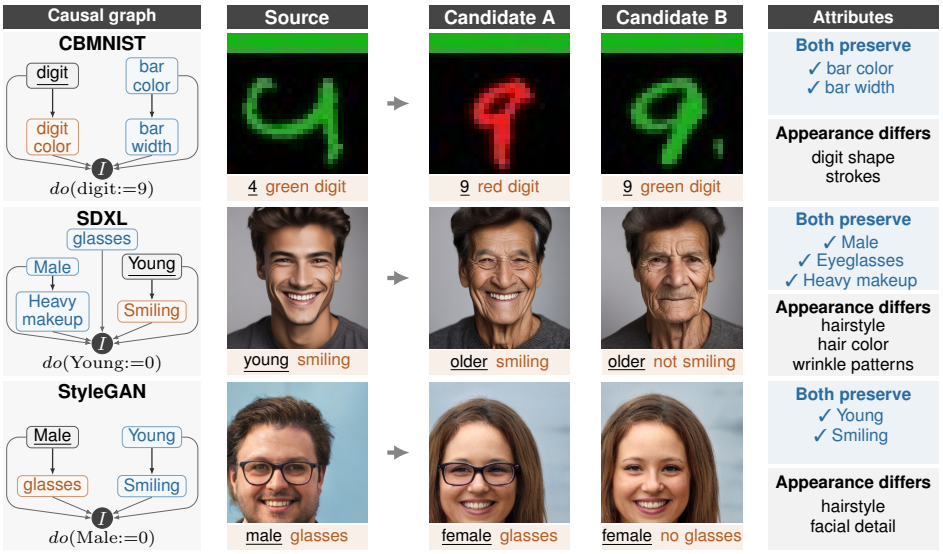


Figure 1: **The same causal requirements admit different image edits.** Both candidates attain the target and preserve measured non-descendants under the same intervention. If both descendant responses are admissible, these checks do not determine the resulting appearance.

Consider changing the green handwritten 4 in Fig. 1 to a 9. In our ColoredBar-MNIST data-generating process, adapted from prior benchmarks (Pan & Bareinboim, 2024; 2025), digit identity affects digit color by construction, whereas the color and width of the horizontal colored bar above the digit are non-descendants of digit identity. The intervention therefore requires a 9 and an unchanged bar, while allowing digit color to respond. When both red and green are admissible, candidates A and B satisfy these attribute requirements, yet differ in shape and strokes. Likewise, the aging example shows how edits satisfying the same attribute requirements can differ in hairstyle, hair color, and wrinkles. Counterfactual consistency alone leaves these aspects of the edited appearance undetermined.

Then, which candidate model should guide editing when several satisfy these causal requirements? A large pretrained generator offers additional information through the image-generation mechanism it has learned from visual data. Its editing responses reveal how semantic changes affect image appearance, making the generator an *anchor* for model selection. Our approach, *Minimal-Shift Mechanism Selection* (MSMS), therefore selects the candidate whose responses most closely agree with the anchor, subject to observational fit and non-descendant preservation relative to the factual instance. To compare models with different architectures, we evaluate them on common editing queries and represent their response distributions in a *shared mechanism space*. The selected NCM then determines counterfactual factors through abduction and intervention, and the pretrained generator renders them using the factual image. Separating these operations supports diffusion models and GANs without image–factor invertibility or full generative retraining.

Our contributions are threefold. **(1)** We introduce MSMS, combining selection in a shared mechanism space with an editing procedure that reuses pretrained diffusion and GAN generators. **(2)** We establish finite-query score-recovery bounds and a conditional counterfactual-risk guarantee, including sufficient conditions for L_3-consistency (Thm. 2). **(3)** Across three benchmarks, we demonstrate MSMS's ability to perform targeted edits while preserving non-descendants with pretrained diffusion and GAN generators, and examine the role of pretrained responses in selecting causal models.

2 Preliminaries

Uppercase letters X and lowercase letters x denote random variables and their realizations, while bold letters $\boldsymbol{X}$ and $\boldsymbol{x}$ denote the corresponding tuples, with domains $\mathcal{D}_X$ and $\mathcal{D}_{\boldsymbol{X}}$. Calligraphic letters (e.g., $\mathcal{M}$) denote SCMs, and fraktur letters (e.g., $\mathfrak{M}$) denote collections of models. Sans-serif (e.g., M) denotes models that need not be SCMs, including the *anchor* M_0, a pretrained generator that edits a source image toward requested attributes. A structural causal model (SCM) (Pearl,

2009) is a tuple $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$. Here $\boldsymbol{U}$ and $\boldsymbol{V}$ are exogenous and endogenous variables, respectively, $\boldsymbol{F} = \{f_{V_j}\}$ is the set of structural assignments $V_j = f_{V_j}(\mathrm{Pa}_j, \boldsymbol{U}_j)$, and $P(\boldsymbol{U})$ is the exogenous background distribution. These assignments induce a causal diagram $\mathcal{G}_{\mathcal{M}}$, whose directed edges represent endogenous parents and bidirected edges represent latent confounding. We write $\mathrm{Pa}_{\mathcal{G}}(V_j)$ for the parents of V_j, $\mathrm{De}_{\mathcal{G}}(\boldsymbol{A}) = \bigcup_{A \in \boldsymbol{A}} \mathrm{De}_{\mathcal{G}}(A)$ for the descendants of $\boldsymbol{A}$ (including $\boldsymbol{A}$ itself), and $\mathrm{ND}_{\mathcal{G}}(\boldsymbol{A}) = \boldsymbol{V} \setminus \mathrm{De}_{\mathcal{G}}(\boldsymbol{A})$. An intervention $do(\boldsymbol{X} = \boldsymbol{x})$ induces the mutilated SCM $\mathcal{M}_{\boldsymbol{x}} = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}_{\boldsymbol{x}}, P(\boldsymbol{U}) \rangle$, in which $\boldsymbol{F}_{\boldsymbol{x}}$ replaces the assignments of $\boldsymbol{X}$ with the constants $\boldsymbol{x}$ and retains all others. For an exogenous state $\boldsymbol{u} \in \mathcal{D}_{\boldsymbol{U}}$, the potential response $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$ is the value that $\boldsymbol{Y} \subseteq \boldsymbol{V}$ takes in $\mathcal{M}_{\boldsymbol{x}}$, and $\boldsymbol{Y}(\boldsymbol{u})$ is its value in $\mathcal{M}$. Integrating these responses against $P(\boldsymbol{U})$ yields the observational (L_1), interventional (L_2), and counterfactual (L_3) distributions $P^{\mathcal{M}}(\boldsymbol{y})$, $P^{\mathcal{M}}(\boldsymbol{y} \mid do(\boldsymbol{x}))$, and $P^{\mathcal{M}}(\boldsymbol{y}_{\boldsymbol{x}} \mid \boldsymbol{z})$ (Bareinboim et al., 2020).

Formally, counterfactual editing constitutes an L_3 query of an unknown SCM $\mathcal{M}^\star$. Let $I \in \boldsymbol{V}$ denote the image, $\boldsymbol{A} \subseteq \boldsymbol{V} \setminus \{I\}$ the editable attributes, and $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{I\}$ a set of observed context variables. Given factual values $i \in \mathcal{D}_I$ and $\boldsymbol{z} \in \mathcal{D}_{\boldsymbol{Z}}$, the target distribution under $do(\boldsymbol{A} = \boldsymbol{a})$ is

$$P^{\mathcal{M}^\star}(I_{\boldsymbol{a}} \mid I = i, \boldsymbol{Z} = \boldsymbol{z}). \tag{1}$$

Given observational images and factor annotations together with the causal diagram $\mathcal{G} = \mathcal{G}_{\mathcal{M}^\star}$, we represent the unknown structural mechanisms by graph-constrained Neural Causal Models (NCMs) (Xia et al., 2021; 2023). An NCM $\hat{\mathcal{M}}(\boldsymbol{\theta})$ parameterizes the structural assignments as $V_j \leftarrow \hat{f}_{V_j,\theta_j}(\mathrm{Pa}_{\mathcal{G}}(V_j), \hat{\boldsymbol{U}}_j)$, where $\hat{\boldsymbol{U}}_j$ denotes its exogenous parents and $\boldsymbol{\theta} = (\theta_j)_j$ collects the neural parameters. Requiring the induced graph to coincide with $\mathcal{G}$ defines the proxy family $\Omega_{\boldsymbol{\theta}}(\mathcal{G}) = \{\hat{\mathcal{M}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}},\ \mathcal{G}_{\hat{\mathcal{M}}(\boldsymbol{\theta})} = \mathcal{G}\}$, detailed in Appendix B.

3 Minimal-shift mechanism selection

In this section, we formulate MSMS as a proxy-selection problem by answering the questions that open each subsection. We first specify the target-generative SCM and the counterfactual image distribution to be estimated, and then represent NCM proxies and the anchor as queryable models in a shared mechanism space (Fig. 2), on which the population selector minimizes disagreement with the anchor under observational-fit and non-descendant preservation constraints. We then estimate this selector from finitely many queries, establish conditions under which it recovers the counterfactual distribution, and describe the training and selection procedure, which Fig. 7 in the appendix summarizes.

3.1 Problem setup

What SCM underlies the editing task, and which distribution do we seek? Consider an unknown SCM $\mathcal{M}^\star$ with known diagram $\mathcal{G}$ that generates an image I from its factors $\boldsymbol{V}$, observed through factor-annotated images and modeled by the NCM family $\Omega_{\boldsymbol{\theta}}(\mathcal{G})$, together with the anchor M_0 of §2. To separate the counterfactual image distribution from the factor responses that generate it, we follow Pan & Bareinboim (2024; 2025) and define SCMs in which the image is rendered from its factors through a rendering substructure.

Definition 1 (Target-generative SCM)**.** A target-generative SCM is an SCM $\mathcal{M}$ with a designated image variable $I \in \boldsymbol{V}$ satisfying $\mathrm{De}_{\mathcal{G}_{\mathcal{M}}}(I) = \{I\}$. Its *rendering substructure* is $(\mathrm{Pa}_I, f_I, U_I)$, where $\mathrm{Pa}_I = \mathrm{Pa}_{\mathcal{G}_{\mathcal{M}}}(I)$, U_I is an exogenous variable on $\mathcal{U}_I$ entering only the assignment of I, and $f_I : \mathcal{D}_{\mathrm{Pa}_I} \times \mathcal{U}_I \to \mathcal{D}_I$ is a measurable function such that $I = f_I(\mathrm{Pa}_I, U_I)$.

We assume that the true model $\mathcal{M}^\star$ and proxies in $\Omega_{\boldsymbol{\theta}}(\mathcal{G})$ belong to this class and share the rendering specification $(\mathrm{Pa}_I, \mathcal{U}_I, P(U_I))$, while specifying their own factor mechanisms and renderers. Fixing the rendering substructure separates the target counterfactual image distribution, the counterfactual factors that a proxy obtains by abduction, intervention, and prediction, and the rendering distribution with which M_0 turns requested factors into an edited image. The proxy and the anchor can each err on a different object, as a proxy may have correct factor mechanisms while rendering through a misaligned anchor, or the reverse, so a guarantee must relate these objects through the queries that each model answers. We therefore abstract every model into a query interface.

Definition 2 (Causal queries and the editing family)**.** A causal query is a tuple $q = (\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{z})$ with $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z} \subseteq \boldsymbol{V}$ and $\boldsymbol{Y} \cap \boldsymbol{X} = \varnothing$ that requests $P^{\mathcal{M}}(\boldsymbol{Y}_{\boldsymbol{x}} \mid \boldsymbol{Z} = \boldsymbol{z})$. Here $\boldsymbol{Z}$ is read in the factual world and $\boldsymbol{Y}$ in the counterfactual world after $do(\boldsymbol{X} = \boldsymbol{x})$, the two copies of a twin network that share the same exogenous variables (Pearl, 2009). The query domain $\mathcal{D}_{\mathcal{Q}}$ collects all such tuples, and the editing family $\mathcal{D}_{\mathcal{Q},\text{edit}} \subseteq \mathcal{D}_{\mathcal{Q}}$ collects the queries $(\{I\}, \boldsymbol{A}, \boldsymbol{a}, \{I\} \cup \boldsymbol{Z}, (i, \boldsymbol{z}))$ with $I \notin \boldsymbol{A}$.

Under such construction, the counterfactual distribution of interest is represented by a response function mapping factual evidence and a factor intervention to a distribution over images. We write an editing query as $q = (i, \boldsymbol{z}, \boldsymbol{A}, \boldsymbol{a})$ and $\mathsf{T}^I_{\mathcal{M}}(q) := P^{\mathcal{M}}(I_{\boldsymbol{a}} \mid I = i, \boldsymbol{Z} = \boldsymbol{z})$ for its image response. Our estimand is $\mathsf{T}^I_{\mathcal{M}^\star}(q)$, whose factorization separates factor inference from rendering.

Proposition 1 (Structural factorization of the editing response)**.** *For a target-generative SCM $\mathcal{M}$ with exogenous state $\boldsymbol{u}_{\text{fac}}$ of the factor mechanisms, a version of its counterfactual response kernel satisfies, for almost every factual observation,*

$$\mathsf{T}^I_{\mathcal{M}}(q)(di') = \iint \delta_{f_I(\mathrm{Pa}^{\boldsymbol{a}}_I(\boldsymbol{u}_{\text{fac}}), u_I)}(di')\, dP^{\mathcal{M}}(u_I \mid i, \mathrm{Pa}_I(\boldsymbol{u}_{\text{fac}}))\, dP^{\mathcal{M}}(\boldsymbol{u}_{\text{fac}} \mid i, \boldsymbol{z}), \tag{2}$$

where δ_x is the point mass at x, and $\mathrm{Pa}_I(\boldsymbol{u}_{\text{fac}})$ and $\mathrm{Pa}^{\boldsymbol{a}}_I(\boldsymbol{u}_{\text{fac}})$ denote the rendering-parent values evaluated in $\mathcal{M}$ and $\mathcal{M}_{\boldsymbol{a}}$, respectively.

The outer integral is factor-side abduction, after which the mutilated SCM computes the counterfactual rendering parents $\mathrm{Pa}^{\boldsymbol{a}}_I$, the inner integral is image-side abduction of U_I from the source, and the point mass is deterministic rendering. In the StyleGAN row of Fig. 1, the parents keep `Young` and `Smiling` under `Male` $:= 0$ while `Eyeglasses` follows its structural assignment, and U_I carries pose and identity. Prop. 1 therefore licenses a separation in which the proxy's factor mechanisms carry the factor side, where the data leave the mechanisms ambiguous, and a renderer absorbs the image side. MSMS assigns this renderer role to M_0, which renders the proxy's counterfactual parents from the source image without inverting the image into factors. Every proxy shares this rendering, so the edit depends on the proxy only through the parents it computes, and editing reduces to choosing a proxy.

3.2 Shared mechanism space

How can the anchor guide selection across different model architectures? The anchor and a proxy share the query interface of Def. 2, so we treat them uniformly as queryable models and compare them through their responses. The anchor answers a query by editing the source image, so each query also fixes how its response is measured on an edited image, such as an attribute predictor that reads `Male`, and $\mathsf{T}_{\mathsf{M}}(q)$ denotes the distribution of this measurement over a model's edits.

Definition 3 (Queryable models and shared mechanism space)**.** A queryable model M is a pair $(\mathcal{D}_{\mathcal{Q}}(\mathsf{M}), \mathsf{T}_{\mathsf{M}})$, consisting of a supported query domain and a measurable response kernel $\mathsf{T}_{\mathsf{M}} : \mathcal{D}_{\mathcal{Q}}(\mathsf{M}) \to \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{S})$ is the space of probability distributions on $\mathcal{S}$. For a collection $\mathfrak{M}$ of such models, choose a common domain $\mathcal{Q} \subseteq \bigcap_{\mathsf{M} \in \mathfrak{M}} \mathcal{D}_{\mathcal{Q}}(\mathsf{M})$ and a query distribution μ on it. With standard Borel query and response spaces, define $\mathsf{M} \sim_\mu \mathsf{M}'$ when their response distributions agree for μ-almost every query. The shared mechanism space is $\mathfrak{T}_\mu := \mathfrak{M}/\sim_\mu$.

Models with identical measured responses occupy the same point in $\mathfrak{T}_\mu$ (Fig. 2a). What the space can distinguish depends on the common domain $\mathcal{Q}$, and here the two models differ in causal level. An NCM answers every query at the counterfactual layer (L_3), whereas the anchor, learned from observational images and attributes, answers editing requests only by conditioning on the requested attributes (L_1).

Together with the graph, the anchor provides two signals for selection among observationally equivalent proxies, and we partition $\mathcal{Q}$ accordingly into *stable* and *correction* families. An intervention cannot change its non-descendants, so stable queries measure non-descendants and take as reference F the point mass at their measurement on the factual image. Correction queries measure the targets and their descendants, whose counterfactual values the factual image does not determine, and take the anchor's response as reference. This graph-induced partition is where an L_1 anchor and an L_3 proxy can be compared.

For example, under the StyleGAN intervention $do(\texttt{Male} := 0)$ in Figs. 1 and 3a, the stable query measures `Young` and `Smiling` on the edited face with fixed attribute predictors and compares them

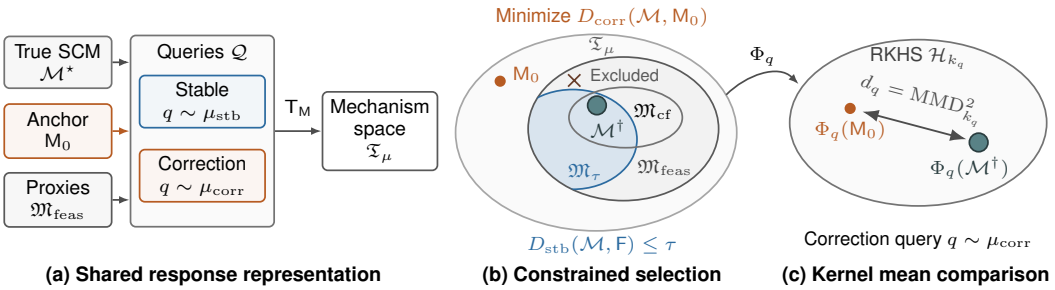


Figure 2: **Minimal-shift selection in mechanism space.** (a) Shared queries compare heterogeneous models. (b) Blue shading marks feasible models satisfying the stable constraint, $\mathfrak{M}_\tau$. The teal $\mathcal{M}^\dagger$ is selected; the cross violates the constraint. Under Theorem 2 with $\beta = 0$, $\mathcal{M}^\dagger \in \mathfrak{M}_{\text{cf}}$. (c) The same anchor and selected model are embedded for a correction query; their squared RKHS distance is d_q.

with their values on the source, the point mass F in Fig. 3a. The correction query measures `Male` and compares it with the same measurement on an anchor edit of the source. Descendants such as `Eyeglasses` are left to the proxy mechanisms, since the reported anchor edit changes only the targets (§E).

Instantiating Def. 3 on the two families, with query distributions μ_{stb} and μ_{corr} over factual cases and interventions, yields the distances used for selection. For each query, we embed a response distribution by its kernel mean Φ_q in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_{k_q}$, and $d_q(\mathsf{M}, \mathsf{M}') = \|\Phi_q(\mathsf{M}) - \Phi_q(\mathsf{M}')\|^2$ is the squared maximum mean discrepancy (MMD) (Gretton et al., 2012) between two responses (Fig. 2c). Averaging over each family gives $D_{\text{stb}}(\mathcal{M}, \mathsf{F}) = \int d_q(\mathcal{M}, \mathsf{F})\, d\mu_{\text{stb}}(q)$ and $D_{\text{corr}}(\mathcal{M}, \mathsf{M}_0) = \int d_q(\mathcal{M}, \mathsf{M}_0)\, d\mu_{\text{corr}}(q)$, on top of which we define the population selector. We use a linear kernel on bounded response features, and Appendix D.2 gives the general construction.

3.3 Minimal-shift selection

Which proxy should be selected if every query could be observed? At the population level, the graph and the observational distribution $P(\boldsymbol{V}) = P^{\mathcal{M}^\star}(\boldsymbol{V})$ restrict the proxies to

$$\mathfrak{M}_{\text{feas}} = \{\mathcal{M} \in \Omega_{\boldsymbol{\theta}}(\mathcal{G}) : P^{\mathcal{M}}(\boldsymbol{V}) = P(\boldsymbol{V})\}. \tag{3}$$

Its members satisfy the counterfactual-consistency bounds of the graph and data (Lem. 5). Within this class, the graph fixes the stable responses, whereas members may differ on correction queries through how they propagate the intervention to descendants. In the StyleGAN row of Fig. 1, candidates A and B differ only in `Eyeglasses`. Stable queries therefore serve as a constraint and correction queries as the objective.

For a stable discrepancy threshold $\tau \ge 0$, the population selection rule is

$$\mathfrak{M}_\tau = \{\mathcal{M} \in \mathfrak{M}_{\text{feas}} : D_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le \tau\}, \qquad \mathcal{M}^\dagger \in \arg\min_{\mathcal{M} \in \mathfrak{M}_\tau} D_{\text{corr}}(\mathcal{M}, \mathsf{M}_0), \tag{4}$$

assuming the minimum is attained. In Fig. 2b, the stable constraint restricts $\mathfrak{M}_{\text{feas}}$ to the blue-shaded $\mathfrak{M}_\tau$, and the selector returns the retained proxy $\mathcal{M}^\dagger$ closest to the anchor M_0, which lies outside $\mathfrak{M}_{\text{feas}}$ as it is not an SCM. The crossed candidate is closer to M_0 but violates the constraint.

3.4 Finite-query selection

What changes when only finitely many candidates and queries can be observed? In practice, training (§3.6) yields a finite pool $C \subseteq \Omega_{\boldsymbol{\theta}}(\mathcal{G})$ that samples only part of the feasible class in Eq. (4), and finitely many queries introduce estimation noise on each score.

The empirical selector averages over finite stable and correction banks B_{stb} and B_{corr} of sampled queries. Sampled responses give per-query estimates U_q of d_q, and averaging them with fixed weights w_q gives plug-in scores $s_{\mathcal{M}}$ and $c_{\mathcal{M}}$ of D_{stb} and D_{corr}. The selector then minimizes $c_{\mathcal{M}}$ over $\widehat{C}_\tau = \{\mathcal{M} \in C : s_{\mathcal{M}} \le \tau + \eta_{\text{stb}}\}$, where the slack η_{stb} absorbs estimation error near the stable boundary (Fig. 3b). For a pool and threshold fixed before scoring, let $C_\tau = \{\mathcal{M} \in C : D_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le \tau\} \neq \varnothing$

be the population counterpart of $\widehat{C}_\tau$, and let $g_{\mathrm{stb}} = \min_{\mathcal{M}\in C\setminus C_\tau} D_{\mathrm{stb}}(\mathcal{M},\mathsf{F}) - \tau$ be its stable margin.

Theorem 1 (Finite-query near-optimal recovery)**.** *Suppose that, with probability at least $1-\delta$, the scores $s_{\mathcal{M}}$ and $c_{\mathcal{M}}$ deviate from D_{stb} and D_{corr} by at most $\varepsilon_{\mathrm{stb}}$ on C and $\varepsilon_{\mathrm{corr}}$ on C_τ, and that $g_{\mathrm{stb}} > 2\varepsilon_{\mathrm{stb}}$. Then for any $\eta_{\mathrm{stb}} \in [\varepsilon_{\mathrm{stb}}, g_{\mathrm{stb}} - \varepsilon_{\mathrm{stb}})$, with probability at least $1-\delta$, $\widehat{C}_\tau = C_\tau$ and $D_{\mathrm{corr}}(\widehat{\mathcal{M}},\mathsf{M}_0) \le \min_{\mathcal{M}\in C_\tau} D_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0) + 2\varepsilon_{\mathrm{corr}}$.*

The margin condition requires that estimation error stay below half the gap between retained and excluded candidates. With bounded per-case scores, the number of selection cases this requires grows logarithmically in $|C|$ (Cor. 3).

Algorithm 1 MSMS

Input: $\mathcal{G}, \mathsf{M}_0, \mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{fit}}, B_{\mathrm{sel}}, \epsilon_{\mathrm{obs}}$
$\hat{\mathcal{M}}(\boldsymbol{\theta}_0)$, N, $\mathcal{J}$, edit tasks, Adam settings
Output: $\widehat{\mathcal{M}}$ or FAIL

Training phase (Eq. (7))
for $t = 1,\dots,N$ **do**
  $\mathcal{B}_t \sim \mathcal{D}_{\mathrm{train}}$
  $\boldsymbol{\theta}_t \leftarrow \mathrm{Adam}(\boldsymbol{\theta}_{t-1}, \nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}))$
end for
$\mathfrak{M}_{\mathrm{cand}} \leftarrow \{\hat{\mathcal{M}}(\boldsymbol{\theta}_t) : t \in \mathcal{J}\}$

Selection phase
$C \leftarrow \{\mathcal{M} \in \mathfrak{M}_{\mathrm{cand}} : \widehat{D}_{\mathrm{obs}}(\mathcal{M}) \le \epsilon_{\mathrm{obs}}\}$
if $C = \varnothing$ **then return** FAIL
For each $\mathcal{M} \in C$, compute on B_{sel}:
  $s_{\mathcal{M}} = \sum_{q\in B_{\mathrm{stb}}} w_q U_q(\mathcal{M},\mathsf{F})$
  $c_{\mathcal{M}} = \sum_{q\in B_{\mathrm{corr}}} w_q U_q(\mathcal{M},\mathsf{M}_0)$
Set τ, η_{stb} by Appendix E.3
$\widehat{C}_\tau \leftarrow \{\mathcal{M} \in C : s_{\mathcal{M}} \le \tau + \eta_{\mathrm{stb}}\}$
if $\widehat{C}_\tau = \varnothing$ **then return** FAIL
return $\widehat{\mathcal{M}} \in \arg\min_{\mathcal{M}\in\widehat{C}_\tau} c_{\mathcal{M}}$

3.5 Counterfactual recovery

When does agreement with the anchor recover the target counterfactual distribution? Thm. 1 controls the correction discrepancy of the selected model, that is, its agreement with the anchor. To connect this agreement with counterfactual correctness, we define a risk measuring image-distribution error over the editing queries of interest.

Definition 4 (Counterfactual risk)**.** For a query distribution ν_{edit} on $\mathcal{D}_{\mathcal{Q},\mathrm{edit}}$ and a measurable metric $d_{\mathcal{P}}$ on $\mathcal{P}(\mathcal{D}_I)$, the counterfactual risk of $\mathcal{M}$ is

$$\mathcal{R}_{L_3}(\mathcal{M}) := \int d_{\mathcal{P}}\big(\mathsf{T}^I_{\mathcal{M}}(q), \mathsf{T}^I_{\mathcal{M}^\star}(q)\big)\, d\nu_{\mathrm{edit}}(q). \tag{5}$$

The zero-risk subclass $\mathfrak{M}_{\mathrm{cf}} := \{\mathcal{M} \in \mathfrak{M}_{\mathrm{feas}} : \mathcal{R}_{L_3}(\mathcal{M}) = 0\}$ reproduces the target image distributions for ν_{edit}-almost every query, and the following guarantee states when $\mathcal{M}^\dagger$ falls in it.

Theorem 2 (Conditional counterfactual-risk guarantee)**.** *Suppose Eq.* (4) *attains its minimum and $\mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau \neq \varnothing$. If $\alpha > 0, \beta \ge 0$ exist such that every $\mathcal{M} \in \mathfrak{M}_\tau$ satisfies*

$$D_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0) - \inf_{\bar{\mathcal{M}}\in\mathfrak{M}_{\mathrm{cf}}\cap\mathfrak{M}_\tau} D_{\mathrm{corr}}(\bar{\mathcal{M}},\mathsf{M}_0) \ge \alpha\mathcal{R}_{L_3}(\mathcal{M}) - \beta, \tag{6}$$

then every population selector satisfies $\mathcal{R}_{L_3}(\mathcal{M}^\dagger) \le \beta/\alpha$. For $\beta = 0$, its image response distributions equal the true counterfactual distributions for ν_{edit}-almost every query.

The theorem rests on two conditions, the first of which, $\mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau \neq \varnothing$, rules out vacuity by placing a correct proxy within the stable constraint. The second, Eq. (6), ties the ranking to correctness by requiring excess correction discrepancy to bound a positive multiple of counterfactual risk. Combined with Thm. 1, the counterfactual risk of $\widehat{\mathcal{M}}$ is bounded by β/α plus terms for the pool's offset from the correct class and for $2\varepsilon_{\mathrm{corr}}$ (Cor. 4). §5.1 examines how well the second condition holds in practice by relating anchor disagreement to the editing error of candidate models.

3.6 Training and selection

How are candidates generated and selected in practice? Algorithm 1 separates training from selection, and Fig. 7 in the appendix traces the procedure from candidate generation to the final edit. Training fits the factor mechanisms on $\mathcal{D}_{\mathrm{train}}$ by minimizing

$$\mathcal{L}(\boldsymbol{\theta}) = \lambda_{\mathrm{data}}\widehat{\mathcal{L}}^{\mathrm{train}}_{\mathrm{obs}}(\boldsymbol{\theta}) + \lambda_{\mathrm{stb}}\widehat{D}^{\mathrm{train}}_{\mathrm{stb}}(\hat{\mathcal{M}}(\boldsymbol{\theta}),\mathsf{F}) + \lambda_{\mathrm{corr}}\widehat{D}^{\mathrm{train}}_{\mathrm{corr}}(\hat{\mathcal{M}}(\boldsymbol{\theta}),\mathsf{M}_0) + \gamma R(\boldsymbol{\theta}), \tag{7}$$

where the three terms fit observed factors, match training-query responses to their references, and regularize the parameters. The iterates at indices $\mathcal{J} \ni N$ whose factor-prediction discrepancy $\widehat{D}_{\mathrm{obs}}$ on $\mathcal{D}_{\mathrm{fit}}$ stays below a preset tolerance ϵ_{obs} form the pool C, a screening on factor fit only that is weaker than Eq. (3). Selection applies the selector of §3.4 to C with banks drawn from a selection

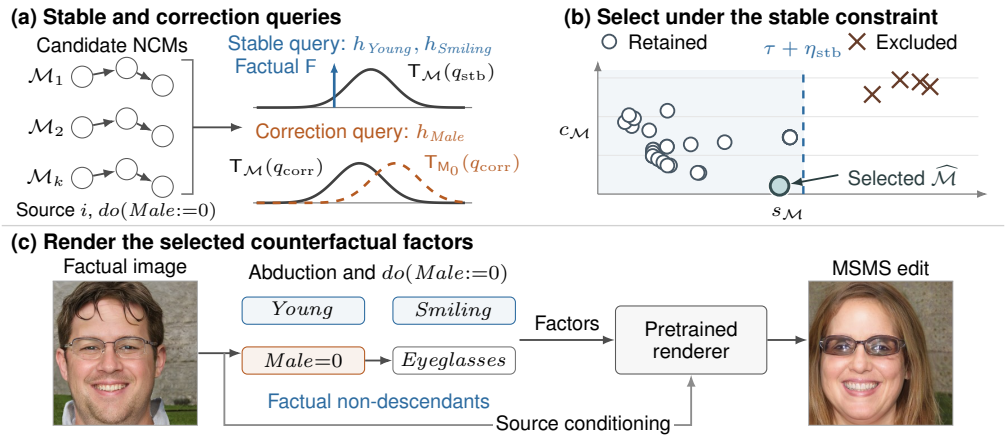


Figure 3: **MSMS procedure.** (a) Stable and correction queries use factual and pretrained references. Candidate face responses use the scoring renderer. (b) Minimize correction discrepancy under the stable constraint. (c) Render counterfactual factors with the pretrained generator.

split disjoint from $\mathcal{D}_{\text{train}}$, and the threshold is the stable-score quantile, among five, whose retained set attains the smallest $c_{\mathcal{M}}$, a data-dependent choice that Cor. 2 covers. During training and selection, face candidates are scored through a learned scoring renderer that supplies the measured image, so §5.1 examines whether these rankings transfer to the anchor's final edits. At test time, the frozen selected model supplies the requested targets, its inferred descendants, and the factual non-descendants to the anchor, which renders them from the source image. Proofs and implementation details are in Appendices D and E.

4 Related Work

Counterfactual image editing has been approached along three lines, the first of which balances the requested change against preservation of the source through diffusion-based reconstruction, guidance, and abductive inference (Jeanneret et al., 2022; Ben Alaya et al., 2024; Song et al., 2024). It also includes independent visual mechanisms (Sauer & Geiger, 2021) and causal latent representations (Yang et al., 2021; Komanduri et al., 2024), which specify how an intervention is rendered, not which causal model should drive it. The second line characterizes the counterfactual responses that the evidence permits, through partial identification (Manski, 2003) and graph-constrained NCMs (Xia et al., 2021; 2023), including counterfactual consistency for images (Pan & Bareinboim, 2024) and its extension to pretrained diffusion editors (Pan & Bareinboim, 2025). MSMS builds on this line by adding a preference among the admissible proxies. The third line relies on identifiable latent representations (Khemakhem et al., 2020; Ahuja et al., 2023) or fully specified deep SCMs (Pawlowski et al., 2020; De Sousa Ribeiro et al., 2023), whereas MSMS requires only a query-response interface to the anchor, which §F.2 relates to measure-theoretic formulations of causality (Park et al., 2023; 2026) and causal distances (Peyrard & West, 2021).

5 Experiments

This section examines three questions about model selection and the editor. §5.1 asks whether anchor agreement selects better candidates, and §5.2 asks whether holding non-descendants matters and how the editor compares with methods that train their own generator or intervene in a latent space. §§E.5 and E.6 detail the setup, and §§G and H report render-count sensitivity, the preservation ablation, FID and KID, further edits, and failure cases.

Benchmarks and metrics. ColoredBar-MNIST (Pan & Bareinboim, 2024; 2025), SDXL (Podell et al., 2024), and StyleGAN (Karras et al., 2021) cover digits and faces with diffusion and GAN renderers (Fig. 1), and SDXL also tests graph misspecification. An independent attribute predictor reads every structural metric from rendered pixels. Target success (TS) is the fraction of intervened attributes read at their requested values, and non-descendant drift (NDD) is the label mismatch with

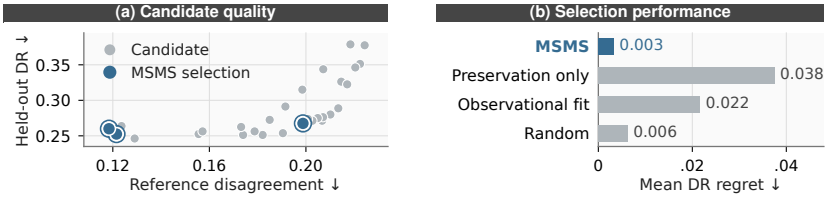


Figure 4: **Selection within original StyleGAN pools.** (a) Reference disagreement versus held-out DR. Circles mark MSMS. (b) Mean DR regret, with one random draw per run.

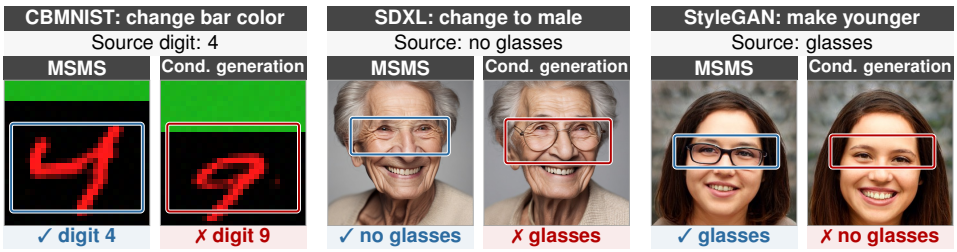


Figure 5: **Preserving non-descendants prevents unrelated changes.** With the model and rendering randomness fixed, Cond. generation re-infers non-descendants and MSMS retains factual values. Blue/red boxes mark preservation/changes. Appendix G gives sources and aggregate results.

the source's non-descendant values. Descendant response discrepancy (DR) is the label mismatch with a benchmark reference for the descendants, constructed as §E.5 describes. Target image distance (TID) averages absolute pixel differences from the same reference, and Fréchet Inception Distance (FID) compares generated and training images (Heusel et al., 2017).

Baselines. BD-CLS (Pan & Bareinboim, 2025) and CausalDiffAE (Komanduri et al., 2024) intervene in the latent space of a diffusion model and run on the two diffusion benchmarks. ANCM (Pan & Bareinboim, 2024) and CGN (Sauer & Geiger, 2021) train their own causal generators from the benchmark data and run on all three.

5.1 Model Selection

Selection is useful only when anchor agreement tracks the final edits. On the StyleGAN pools of 30 candidates from three training trajectories, lower anchor disagreement goes with lower DR and higher TS, with Spearman correlations of $+0.87$ and -0.75 (Fig. 4a). This relation is consistent with the second condition of Thm. 2 on these pools. Since the disagreement is computed through the scoring renderer while DR and TS are measured on the anchor's final edits, the ranking carries over to final editing. MSMS therefore attains a smaller DR regret, the excess DR over the best candidate in the pool, than selection by preservation, observational fit, or random selection (Fig. 4b).

5.2 Counterfactual Editing

MSMS attains the highest target success on all three benchmarks, with non-descendant drift and TID within 0.03 and 0.02 of the best method. It also gives the best FID on SDXL and the best DR on StyleGAN (Tab. 1). First, on the two diffusion benchmarks, MSMS and BD-CLS both restrict the edit to counterfactually consistent factor responses. The anchor-guided choice among these responses gives MSMS higher target success and lower FID than BD-CLS on both benchmarks, equal drift on CBMNIST, and lower DR on SDXL. BD-CLS retains lower DR on CBMNIST and lower drift on SDXL. Second, the latent editors invert the image into a diffusion latent space and therefore do not apply to the GAN benchmark. MSMS requires only the query interface of the anchor, and on StyleGAN it attains the highest target success and the lowest drift and DR. The full-training models must learn a causal generator from the benchmark data, and the labeled face images do not suffice to train a generator of the quality of SDXL or StyleGAN. ANCM therefore produces smoothed faces with the highest FID on both face benchmarks. CGN generates a face for the requested factors without conditioning on the source, which yields the best FID on StyleGAN together with higher

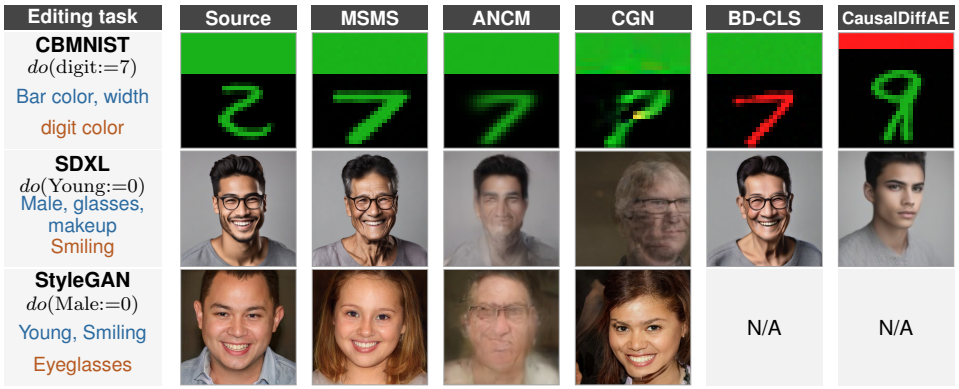


Figure 6: **Counterfactual edits across three benchmarks.** Each row shares a source and intervention. Labels mark non-descendants in blue and descendants in orange. N/A denotes unavailable pairings.

Table 1: Editing performance: mean±SD over 3 runs. Bold marks the best values per benchmark.

Benchmark	Method	TS↑	NDD↓	DR↓	TID↓	FID↓
CBMNIST	CGN	0.830±0.008	0.053±0.004	0.998±0.000	0.092±0.000	57.1
	ANCM	0.970±0.000	**0.010±0.000**	0.870±0.000	0.069±0.000	41.9
	BD-CLS	0.977±0.001	0.011±0.001	**0.397±0.007**	**0.058±0.001**	20.8
	CausalDiffAE	0.534±0.006	0.176±0.005	0.476±0.019	0.074±0.001	**8.6**
	MSMS (ours)	**0.980±0.004**	0.011±0.003	0.715±0.019	0.073±0.001	13.0
SDXL	CGN	0.498±0.009	0.456±0.004	0.317±0.011	0.225±0.001	142.8
	ANCM	0.399±0.017	0.402±0.004	**0.276±0.006**	**0.190±0.000**	180.2
	BD-CLS	0.624±0.005	**0.046±0.001**	0.439±0.001	0.209±0.000	66.4
	CausalDiffAE	0.654±0.014	0.405±0.002	0.370±0.013	0.208±0.001	97.3
	MSMS (ours)	**0.660±0.010**	0.075±0.002	0.370±0.006	0.210±0.000	**45.6**
StyleGAN	CGN	0.716±0.014	0.204±0.015	0.465±0.014	0.258±0.001	**15.0**
	ANCM	0.407±0.011	0.396±0.012	0.515±0.009	**0.234±0.000**	158.8
	MSMS (ours)	**0.931±0.010**	**0.098±0.003**	**0.260±0.008**	0.254±0.000	68.4

drift than MSMS. CausalDiffAE regenerates the image from its causal latent and thereby alters non-descendants such as the bar color and the glasses.

In Fig. 6, MSMS edits the source, so the digit becomes a 7 above the unchanged bar and the SDXL face ages with wrinkles and graying hair while its glasses, smile, and background stay. Under $do(\texttt{Male}:=0)$, the StyleGAN face changes hairstyle and clothing while its pose, smile, and framing remain. This preservation depends on holding the non-descendants at their factual values. Recomputing them changes the digit in a bar-color edit and the glasses in face edits and raises NDD four- to thirty-seven-fold, with TS and DR moving by at most 0.024 (Fig. 5, Tab. 7, and §H.3).

6 Conclusion

In this work, we introduce MSMS to select among observationally equivalent causal models that imply different counterfactual images. Our approach compares their responses in a shared mechanism space and uses a pretrained generator to rank candidates subject to observational fit and non-descendant preservation. We establish finite-query score-recovery bounds and a conditional counterfactual-risk guarantee, including sufficient conditions for L_3-consistency. Separating factor abduction from rendering enables reuse of pretrained diffusion and GAN generators without image–factor invertibility or full generative retraining. Experiments across three benchmarks demonstrate MSMS's capability in targeted editing and non-descendant preservation, and examine the role of pretrained responses in model selection. Future directions we leave open include correction queries that leave the descendants to the anchor, candidate pools that span independent initializations and alignment weights, in which selection can depart from the last training iterate, and response kernels that compare distributions beyond their means.

AI USE STATEMENT

Generative AI tools were used for code editing, experiment orchestration, and proofreading (both english and math) of the manuscript. The research idea, the method, and all formal statements and derivations of MSMS are the authors' own. All AI-assisted code was reviewed and tested by the authors, and all AI-suggested edits to the text were checked before being accepted. The authors take full responsibility for the final content, including text, claims, and experimental artifacts regarding this paper.

ETHICS STATEMENT

This work uses public pretrained generators (SDXL (Podell et al., 2024) and StyleGAN3-T (Karras et al., 2021)), the ColoredBarMNIST benchmark (Pan & Bareinboim, 2024; 2025), and synthetic samples. The face experiments use synthetic portraits. We release neither an identity-targeting model, a new high-risk generator, nor a scraped dataset. The contribution is a selection rule over proxy causal models rather than a new image generator. §F.4 discusses risks inherited from the underlying generators, including identity misuse, misinformation, and non-consensual editing. The study recruits no participants or crowdworkers and collects no personal or demographic data. Two blinded internal research-team members perform the predeclared output-quality review under a frozen rubric. A distinct third blinded member adjudicates disagreements, and the review records model-output rankings only. We therefore treat institutional review board approval as inapplicable. §E.6 credits the third-party assets CGN (Sauer & Geiger, 2021), ANCM (Pan & Bareinboim, 2024), BD-CLS (Pan & Bareinboim, 2025), and CausalDiffAE (Komanduri et al., 2024). SDXL is used under the CreativeML Open RAIL++-M terms and StyleGAN3 under the NVIDIA Source Code License. CelebA-HQ (Karras et al., 2018) is used only to train the attribute predictor that screens generated faces and the encoder that inverts them, under its research-use terms.

REPRODUCIBILITY STATEMENT

Appendix D provides complete proofs of the main-text results and states the supplementary results with their assumptions. The causal and kernel background appears in §§B and C, and Tab. 2 in Appendix F maps each theoretical object to its implementation counterpart and assumptions. Within §E, §E.1 describes the interfaces, §E.5 the evaluation metrics, §E.2 candidate generation and selection, §E.3 the stable discrepancy threshold, §E.6 the benchmarks and adapters, §E.4 the query banks, and §E.8 the compute resources. Tabs. 3 and 4 list candidate-generation settings and benchmark-specific queries. Source images, evaluation predictors, selection rules, and run settings were fixed and recorded with their SHA-256 hashes before the reported results were computed. Anonymized code for candidate generation, selection, the baselines, and evaluation is provided as supplementary material.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407. PMLR, 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.1090/S0002-9947-1950-0051437-7.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. Technical Report R-60, CausalAI Lab, Columbia University, 2020.

Malek Ben Alaya, Daniel M. Lang, Benedikt Wiestler, Julia A. Schnabel, and Cosmin I. Bercea. MedEdit: Counterfactual diffusion-based image editing on brain MRI. In *Simulation and Synthesis in Medical Imaging (MICCAI 2024 SASHIMI Workshop)*. Springer, 2024. arXiv:2407.15270.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1801.01401.

Patrick Chao, Patrick Blöbaum, Sapan Kirit Patel, and Shiva Kasiviswanathan. Modeling causal mechanisms with diffusion models for interventional and counterfactual queries. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=EDHQDsqiSe.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7390–7425. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/de-sousa-ribeiro23a.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Guang-Yuan Hao, Jiji Zhang, Biwei Huang, Hao Wang, and Kun Zhang. Natural counterfactuals with necessary backtracking. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/1b10264c77a2a1e0ef8abfbd68d36583-Abstract-Conference.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.08500.

Shanshan Huang, Haoxuan Li, Chunyuan Zheng, Lei Wang, Guorui Liao, Zhili Gong, Huayi Yang, and Li Liu. Visual representation learning through causal intervention for controllable image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *5th International Conference on Learning Representations (ICLR 2017)*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022. arXiv:2203.15636.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 852–863, 2021.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020. arXiv:1907.04809.

Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *ECAI 2024*, pp. 2516–2523. IOS Press, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. arXiv:1811.12359.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, 2003. ISBN 978-0-387-00454-5. doi: 10.1007/b97478.

Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Benchmarking counterfactual image generation. In *Advances in Neural Information Processing Systems 37: Datasets and Benchmarks Track*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/f04351c9fa1e22797c7d32c1f6d23948-Abstract-Datasets_and_Benchmarks_Track.html.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017. doi: 10.1561/2200000060.

Yushu Pan and Elias Bareinboim. Counterfactual image editing. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39087–39101. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/pan24a.html.

Yushu Pan and Elias Bareinboim. Counterfactual image editing with disentangled causal latent space. In *Advances in Neural Information Processing Systems*, 2025. URL https://neurips.cc/virtual/2025/poster/115533. NeurIPS 2025.

Junhyung Park, Simon Buchholz, Bernhard Schölkopf, and Krikamol Muandet. A measure-theoretic axiomatisation of causality. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sPLTQSf6GI.

Junhyung Park, Fanny Yang, and Thomas Icard. Counterfactual spaces, 2026. URL https://arxiv.org/abs/2601.00507.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2006.06485.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Maxime Peyrard and Robert West. A ladder of causal distances. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2012–2018. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/277. URL https://doi.org/10.24963/ijcai.2021/277. Main Track.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 647–668. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/sanchez22a.html.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *9th International Conference on Learning Representations (ICLR 2021)*, 2021.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(64):1941–1979, 2008.

Xue Song, Jiequan Cui, Hanwang Zhang, Jingjing Chen, Richang Hong, and Yu-Gang Jiang. Doubly abductive counterfactual inference for text-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9162–9171, 2024. arXiv:2403.02981.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL https://jmlr.org/papers/v11/sriperumbudur10a.html.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pp. 567–573, 2002.

Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing (eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 177–196. PMLR, 2023. URL https://proceedings.mlr.press/v213/kugelgen23a.html.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: expressiveness, learnability, and inference. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Kevin Muyuan Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*, 2023.

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. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yang_CausalVAE_Disentangled_Representation_Learning_via_Neural_Structural_Causal_Models_CVPR_2021_paper.html.

Appendix

Table of Contents

A Notation **17**

B Causal background **18**

B.1 SCMs, induced graphs, and c-components 19

B.2 Interventions and the three causal layers 19

B.3 Neural causal models and the proxy family 20

B.4 Partial identification and counterfactual consistency 20

B.5 Intervention-relative partition and block factorization 21

B.6 Counterfactual ambiguity under observational and interventional agreement 22

C Kernel background **22**

C.1 Kernels and reproducing kernel Hilbert spaces 22

C.2 Kernel mean embeddings and maximum mean discrepancy 23

C.3 Relation to the operational score and to Markov kernels 23

D Proofs for the methodology **24**

D.1 Structural factorization of the editing response 24

D.2 Full axiomatic definition of the mechanism space 24

D.3 Metric structure of mechanism space 25

D.4 Kernelized realization on joint query-response distributions 26

D.5 Proxy classes and population selector 28

D.6 Observational feasibility and optional bounded functionals 28

D.7 Nonemptiness under exact realizability 29

D.8 Operational score and response-sample estimation 30

D.9 Finite-query near-optimal recovery 31

D.10 Sample complexity 32

D.11 Conditional L_3 risk guarantee 34

D.12 Conditional risk control from reference proximity 35

E Implementation and evaluation details **36**

E.1 Proxy, renderer, and inference 36

E.2 Candidate generation and held-out selection 37

E.3 Choosing the stable discrepancy threshold 39

E.4 Query banks and empirical discrepancies 40

E.5 Evaluation protocol 42

E.6 Benchmarks and baselines 43

E.7 Computational cost of selection 45
E.8 Compute resources 45

F Further discussion **45**
F.1 Plausibility of the assumptions 45
F.2 Relation to prior assumptions 46
F.3 Limitations 48
F.4 Broader impacts 48

G Additional editing results **49**
G.1 Sensitivity to the number of renders 49
G.2 Holding non-descendant factors 49
G.3 Distributional image quality 50

H Additional figures **50**
H.1 Benchmarks and reference images 50
H.2 Edits by every method 52
H.3 Non-descendant invariance on additional cases 52
H.4 Eight ordered renders of one case 52
H.5 Failure cases 52

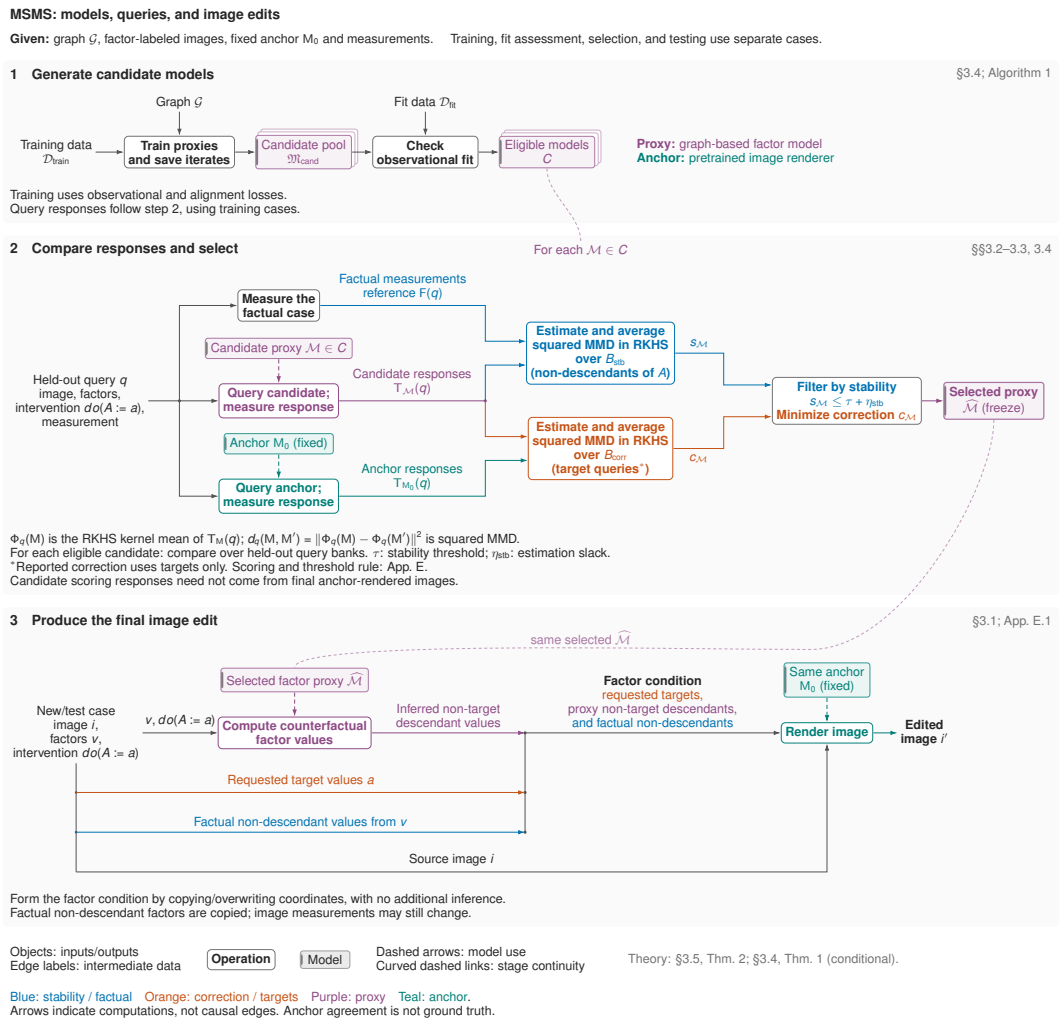


Figure 7: Overview of MSMS: candidate generation, held-out model selection, and counterfactual image editing. *Please zoom in to read the detailed labels and mathematical notation.*

A NOTATION

This appendix collects the principal notation of the paper, organized by conceptual domain, with the location where each symbol is introduced. Proof-local variables and ephemeral indices are omitted.

General conventions. Uppercase letters (V) denote random variables, bold uppercase letters ($\boldsymbol{V}$) denote tuples of variables, and lowercase letters (v, $\boldsymbol{v}$) denote realized values, with $\mathcal{D}_{\boldsymbol{V}}$ the domain of $\boldsymbol{V}$. Calligraphic letters ($\mathcal{M}$) denote SCMs, fraktur letters ($\mathfrak{M}$) denote collections of models, and sans-serif letters (M) denote models that need not be SCMs. Hats ($\hat{\mathcal{M}}$, $\widehat{D}$) mark learned proxies and empirical estimates, $\mathcal{M}^\star$ marks the true model, and $\mathcal{M}^\dagger$ and $\widehat{\mathcal{M}}$ mark the population and selected models.

A. STRUCTURAL CAUSAL MODELS AND GRAPHS

Symbol	Meaning	Introduced
$\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$	Structural causal model	§2; App. B.1
$\mathcal{M}^\star$	True target-generative SCM	Def. 1
$\mathcal{G}$, $\mathcal{G}_{\mathcal{M}}$	Causal graph, graph induced by $\mathcal{M}$	§2; Def. 5
$\mathrm{Pa}_{\mathcal{G}}$, $\mathrm{De}_{\mathcal{G}}$, $\mathrm{ND}_{\mathcal{G}}$	Parents, descendants (including the set itself), and non-descendants	§2; Def. 5
$\mathcal{M}_{\boldsymbol{x}}$, $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$	Mutilated SCM under $do(\boldsymbol{X} = \boldsymbol{x})$ and the potential response	§2; App. B.2
L_1, L_2, L_3	Observational, interventional, and counterfactual layers	§2; Eqs. (8)–(9)
$\mathcal{N}_X$, $\mathcal{D}_X$	Non-descendants and strict descendants of the intervention targets $\boldsymbol{X}$	App. B.5

B. EDITING TASK AND CAUSAL QUERIES

Symbol	Meaning	Introduced
I, $\boldsymbol{A}$, $\boldsymbol{a}$, $\boldsymbol{Z}$, $\boldsymbol{z}$	Image, intervened variables with requested values, and context variables with factual values	§2
$(\mathrm{Pa}_I, f_I, U_I)$	Rendering substructure: rendering parents, renderer, and image-specific noise	Def. 1
$q = (\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{x}, \boldsymbol{Z}, \boldsymbol{z})$	Causal query	Def. 2
$\mathcal{D}_{\mathcal{Q}}$, $\mathcal{D}_{\mathcal{Q},\mathrm{edit}}$	Query domain and editing family	Def. 2
$\mathsf{T}^I_{\mathcal{M}}(q)$	Image response of $\mathcal{M}$ to an editing query	§3.1; Prop. 1
$\boldsymbol{u}_{\mathrm{fac}}$, u_I	Factor-side and image-side exogenous states	Eq. (2)

C. PROXIES AND MODEL CLASSES

Symbol	Meaning	Introduced
$\hat{\mathcal{M}}(\boldsymbol{\theta})$, $\Omega_{\boldsymbol{\theta}}(\mathcal{G})$	Neural proxy with parameters $\boldsymbol{\theta}$ and the graph-constrained NCM family	§2; App. B.3
M_0	Anchor, a pretrained generator that edits a source image toward requested attributes	§2
M	Queryable model	Def. 3
$\Omega(\mathcal{G}, P(\boldsymbol{V}))$	SCMs compatible with the graph and the observational distribution	Eq. (11)
$\mathfrak{M}_{\mathrm{feas}}$	Observationally feasible proxy class	Eq. (3)
$\mathfrak{M}_\tau$	Retained class under the stable threshold τ	Eq. (4)
$\mathfrak{M}_{\mathrm{cf}}$	Zero-risk subclass	§3.5
$\mathfrak{M}_{\mathrm{cand}}$, C	Saved models, and the candidate pool after the observational-fit screen	Alg. 1; App. E.2

D. Mechanism space and discrepancies

Symbol	Meaning	Introduced
$\mathsf{T}_{\mathsf{M}}(q), \mathcal{S}$	Measured response distribution and the response space	Def. 3
$\mathcal{Q}, \mu, \mathfrak{T}_\mu$	Common query domain, query distribution, and shared mechanism space	Def. 3
$\mu_{\mathrm{stb}}, \mu_{\mathrm{corr}}$	Stable and correction query distributions	§3.2
F	Stable reference, the query-indexed point mass at the factual measurement of a stable query	§3.2
$\Phi_q, \mathcal{H}_{k_q}, d_q$	Kernel mean embedding, RKHS, and squared MMD of one query	§3.2; Def. 11
$D_{\mathrm{stb}}(\mathcal{M}, \mathsf{F}), D_{\mathrm{corr}}(\mathcal{M}, \mathsf{M}_0)$	Stable and correction discrepancies	§3.2

E. Selection

Symbol	Meaning	Introduced
τ	Stable discrepancy threshold	Eq. (4)
$\mathcal{M}^\dagger$	Population selector	Eq. (4)
$B_{\mathrm{stb}}, B_{\mathrm{corr}}$	Held-out stable and correction query banks	§3.4; App. E.4
U_q, w_q	Per-query estimate of d_q and its weight	§3.4; App. D.8
$s_{\mathcal{M}}, c_{\mathcal{M}}$	Stable and correction scores of a candidate	§3.4
η_{stb}	Slack of the empirical stable constraint	§3.4; App. E.3
$\widehat{C}_\tau, C_\tau, g_{\mathrm{stb}}$	Empirical and population retained pools and the stable margin	§3.4
$\varepsilon_{\mathrm{stb}}, \varepsilon_{\mathrm{corr}}, \delta$	Uniform score accuracies and failure probability	Thm. 1
$\widehat{\mathcal{M}}$	Selected model	Thm. 1

F. Counterfactual risk

Symbol	Meaning	Introduced
$\nu_{\mathrm{edit}}, d_{\mathcal{P}}$	Editing query distribution and metric on image distributions	Def. 4
$\mathcal{R}_{L_3}(\mathcal{M})$	Counterfactual risk	Eq. (5)
α, β	Constants of the risk-control condition	Eq. (6); App. D.12
ϵ_Z, ϵ_U	Factor-side and image-side recovery slacks	App. D.12

G. Training and pool construction

Symbol	Meaning	Introduced
$\mathcal{L}(\boldsymbol{\theta})$	Training objective	Eq. (7)
$\lambda_{\mathrm{data}}, \lambda_{\mathrm{stb}}, \lambda_{\mathrm{corr}}, \gamma, R$	Multipliers of the relaxation and the regularizer	Eq. (7)
$\mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{fit}}$	Training split and observational-fit split	Alg. 1
$N, \mathcal{J}, \epsilon_{\mathrm{obs}}, \widehat{D}_{\mathrm{obs}}$	Iteration count, pool indices, tolerance, and factor-prediction discrepancy	Alg. 1; App. E.2

B Causal background

This appendix collects the causal background deferred from §2 and §3. It first reviews SCMs, their induced graphs, and c-components, then restates the semantics of the three causal layers and defines

counterfactual image editing. It then introduces neural causal models as the proxy family, recalls partial identification and counterfactual consistency, and derives the block factorization that underlies the stable and correction families of §3.2. It closes with a binary example of the ambiguity that motivates selection. Bold symbols denote tuples, and $\mathcal{D}_{\boldsymbol{V}}$ denotes the domain of $\boldsymbol{V}$. Variable spaces are standard Borel and conditional distributions are regular conditional probabilities, so equalities between conditional response distributions hold almost everywhere under the stated query measure.

B.1 SCMs, INDUCED GRAPHS, AND C-COMPONENTS

An SCM $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$ is defined in §2, with each endogenous variable satisfying $V_j = f_{V_j}(\mathrm{Pa}(V_j), \boldsymbol{U}_{V_j})$. Every SCM induces a causal graph over its endogenous variables.

Definition 5 (Causal graph induced by an SCM)**.** The causal graph induced by $\mathcal{M}$ is the mixed graph $\mathcal{G}_{\mathcal{M}}$ over $\boldsymbol{V}$ that contains $V_i \to V_j$ when V_i is an argument of f_{V_j} and $V_i \leftrightarrow V_j$ when the two assignments share an exogenous parent. For $\boldsymbol{S} \subseteq \boldsymbol{V}$, $\mathrm{Pa}_{\mathcal{G}}(\boldsymbol{S})$ collects the parents of its members outside $\boldsymbol{S}$, $\mathrm{De}_{\mathcal{G}}(\boldsymbol{S}) = \bigcup_{S \in \boldsymbol{S}} \mathrm{De}_{\mathcal{G}}(S)$ its descendants, where $\mathrm{De}_{\mathcal{G}}(S)$ contains S and every variable reachable from it by a directed path, and $\mathrm{ND}_{\mathcal{G}}(\boldsymbol{S}) = \boldsymbol{V} \setminus \mathrm{De}_{\mathcal{G}}(\boldsymbol{S})$ its non-descendants.

Bidirected edges induce a second partition of $\boldsymbol{V}$.

Definition 6 (C-components (Tian & Pearl, 2002))**.** A c-component of $\mathcal{G}$ is a maximal subset of $\boldsymbol{V}$ connected by bidirected paths. The c-component of a variable therefore collects the variables whose assignments share exogenous influence with it.

The two partitions play different roles below. The descendant relation underlies the intervention-relative decomposition of §B.5, whereas c-components describe the latent-confounding topology that the neural parametrization of §B.3 follows.

B.2 INTERVENTIONS AND THE THREE CAUSAL LAYERS

An SCM assigns precise semantics to interventions and counterfactuals, which a generic generative model does not. The observational layer is the induced distribution $P^{\mathcal{M}}(\boldsymbol{V})$. The intervention $do(\boldsymbol{X} = \boldsymbol{x})$ induces the mutilated SCM $\mathcal{M}_{\boldsymbol{x}} = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}_{\boldsymbol{x}}, P(\boldsymbol{U}) \rangle$, where $\boldsymbol{F}_{\boldsymbol{x}} = \{f_{V_j} : V_j \in \boldsymbol{V} \setminus \boldsymbol{X}\} \cup \{X_k \leftarrow x_k : X_k \in \boldsymbol{X}\}$ replaces only the assignments for $\boldsymbol{X}$ and preserves the exogenous distribution. Its interventional distribution is

$$P^{\mathcal{M}}(\boldsymbol{y} \mid do(\boldsymbol{x})) = P^{\mathcal{M}_{\boldsymbol{x}}}(\boldsymbol{y}) = \int \mathbf{1}[\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}] \, dP^{\mathcal{M}}(\boldsymbol{u}), \tag{8}$$

which reduces to the observational distribution when $\boldsymbol{X} = \varnothing$. The counterfactual layer first updates the exogenous background on factual evidence and then evaluates the intervened model under that same abducted background. Given evidence $\boldsymbol{Z} = \boldsymbol{z}$ with $P^{\mathcal{M}}(\boldsymbol{z}) > 0$,

$$P^{\mathcal{M}}(\boldsymbol{y}_{\boldsymbol{x}} \mid \boldsymbol{z}) = \int P^{\mathcal{M}_{\boldsymbol{x}}}(\boldsymbol{y} \mid \boldsymbol{u}) \, dP^{\mathcal{M}}(\boldsymbol{u} \mid \boldsymbol{z}), \tag{9}$$

under the regular-conditional-probability convention used throughout the paper, and the counterfactual layer reduces to the interventional layer when $\boldsymbol{Z} = \varnothing$.

Definition 7 (Counterfactual image editing)**.** Let $I \in \boldsymbol{V}$ denote the image, $\varnothing \neq \boldsymbol{A} \subseteq \boldsymbol{V} \setminus \{I\}$ the editable attributes, and $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{I\}$ the evidence variables. Given a factual image i, a target assignment $\boldsymbol{a}$, and evidence $\boldsymbol{z}$, counterfactual image editing approximates $P^{\mathcal{M}}(I_{\boldsymbol{a}} \mid I = i, \boldsymbol{Z} = \boldsymbol{z})$ and samples edited images from it.

Counterfactual image editing is therefore an L_3 task, since the target attributes are intervened upon while the edited image is defined relative to the factual observation (Pan & Bareinboim, 2024; 2025). Def. 1 equips the SCM with a designated image variable and a rendering specification $(\mathrm{Pa}_I, \mathcal{U}_I, P(U_I))$ shared between $\mathcal{M}^\star$ and every candidate, each of which specifies its own renderer and factor mechanisms. The exogenous variable U_I enters only the assignment of I and is independent of the exogenous variables of the factor mechanisms under $P(\boldsymbol{U})$, whereas factor assignments may share exogenous variables, represented by bidirected edges in the factor subgraph.

B.3 NEURAL CAUSAL MODELS AND THE PROXY FAMILY

In practice the true SCM is unknown and cannot be searched over directly. To address this, Xia et al. (2021) introduced neural causal models (NCMs), a family of proxy causal models that encode structural constraints within neural parameterizations of SCM mechanisms.

Definition 8 (Neural causal model (Xia et al., 2021))**.** A neural causal model (NCM) is an SCM $\hat{\mathcal{M}}(\boldsymbol{\theta}) = \langle \hat{\boldsymbol{U}}, \boldsymbol{V}, \hat{\boldsymbol{F}}, P(\hat{\boldsymbol{U}}) \rangle$ such that (i) $\hat{\boldsymbol{U}} = \{\hat{U}_{\boldsymbol{C}} : \boldsymbol{C} \subseteq \boldsymbol{V}\}$, where each $\hat{U}_{\boldsymbol{C}}$ is associated with a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ and has domain $\mathcal{D}_{\hat{U}_{\boldsymbol{C}}} = [0, 1]$; (ii) $\hat{\boldsymbol{F}} = \{\hat{f}_{V_j,\theta_j} : V_j \in \boldsymbol{V}\}$, where each $\hat{f}_{V_j,\theta_j}$ is a feedforward neural network defining the structural assignment $V_j \leftarrow \hat{f}_{V_j,\theta_j}\big(\boldsymbol{V}_{\mathrm{Pa}(V_j)}, \boldsymbol{U}_{V_j}\big)$ with $\boldsymbol{U}_{V_j} := \{\hat{U}_{\boldsymbol{C}} : V_j \in \boldsymbol{C}\}$; and (iii) $P(\hat{\boldsymbol{U}})$ is defined so that each $\hat{U}_{\boldsymbol{C}} \sim \mathrm{Unif}(0, 1)$.

An NCM can be viewed as a normalized SCM with uniformly distributed exogenous noise, and it therefore induces the same three causal layers.

Definition 9 ($\mathcal{G}$-constrained neural causal model (Xia et al., 2021))**.** Let $\mathcal{G}$ be a causal diagram over $\boldsymbol{V}$. An NCM $\hat{\mathcal{M}}(\boldsymbol{\theta})$ is $\mathcal{G}$-constrained if, for every $V_j \in \boldsymbol{V}$, its structural assignment takes the form $V_j \leftarrow \hat{f}_{V_j,\theta_j}\big(\boldsymbol{V}_{\mathrm{Pa}_{\mathcal{G}}(V_j)}, \boldsymbol{U}_{V_j}\big)$, where $\boldsymbol{U}_{V_j} = \hat{U}_{\boldsymbol{C}(V_j)}$ and $\boldsymbol{C}(V_j)$ is the c-component of V_j.

$\mathcal{G}$-constrained NCMs are our searchable class of proxy models because they combine consistency with flexibility. Xia et al. (2021) showed that, for L_2 queries identifiable from $\mathcal{G}$, an L_1-consistent $\mathcal{G}$-constrained NCM recovers the true L_2 quantities of the underlying SCM, while NCMs are expressive enough to represent all three layers of a given SCM. The neural proxy $\hat{\mathcal{M}}(\boldsymbol{\theta})$ of §2 is such a model, and we denote the resulting search space by

$$\Omega_{\boldsymbol{\theta}}(\mathcal{G}) := \left\{ \hat{\mathcal{M}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}},\ \mathcal{G}_{\hat{\mathcal{M}}(\boldsymbol{\theta})} = \mathcal{G} \right\}, \tag{10}$$

which collects the proxies whose induced mixed graph is exactly $\mathcal{G}$, independently of how their exogenous variables are grouped.

Neural construction and implementation scope. For finite endogenous domains, the graph-constrained NCM construction of Xia et al. (2023, Definition 2) assigns an independent normalized exogenous variable to each maximal bidirected clique C and lets the mechanism for V_j consume the variables for cliques containing V_j:

$$V_j \leftarrow \hat{f}_{V_j,\theta_j}\big(\mathrm{Pa}_{\mathcal{G}}(V_j), \{\hat{U}_C : V_j \in C\}\big).$$

The cited construction covers finite-domain variables and maximal bidirected cliques. The implementation uses one Gaussian block per bidirected district (c-component), a broader continuous-domain engineering choice. Sharing one block across a non-clique district can induce extra bidirected edges. For example, one block shared by A, B, and C in $A \leftrightarrow B \leftrightarrow C$ also permits A–C latent dependence. The parametrization belongs to $\Omega_{\boldsymbol{\theta}}(\mathcal{G})$ exactly when its induced mixed graph equals $\mathcal{G}$, as it does for singleton or clique districts. Theorem-to-implementation claims require this equality.

B.4 PARTIAL IDENTIFICATION AND COUNTERFACTUAL CONSISTENCY

The usefulness of graph-constrained NCMs does not extend to the counterfactual layer. Graph and observational agreement generally leave counterfactual distributions unidentified (Shpitser & Pearl, 2008; Bareinboim et al., 2020). Let

$$\Omega(\mathcal{G}, P(\boldsymbol{V})) = \{\mathcal{M} : \mathcal{G}_{\mathcal{M}} = \mathcal{G},\ P^{\mathcal{M}}(\boldsymbol{V}) = P(\boldsymbol{V})\}. \tag{11}$$

For a scalar counterfactual functional Q, its sharp admissible range is

$$\underline{Q} = \inf_{\mathcal{M} \in \Omega(\mathcal{G}, P(\boldsymbol{V}))} Q(\mathcal{M}), \qquad \overline{Q} = \sup_{\mathcal{M} \in \Omega(\mathcal{G}, P(\boldsymbol{V}))} Q(\mathcal{M}). \tag{12}$$

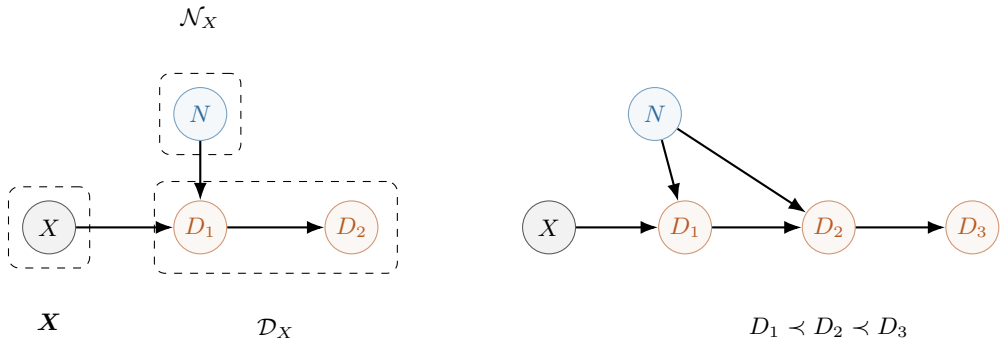
These are the optimal counterfactual bounds (Manski, 2003; Shpitser & Pearl, 2008). Under this non-identifiability, Pan & Bareinboim (2024) propose counterfactual consistency as a feasibility criterion. It requires the estimated quantity to lie within the admissible range instead of demanding a unique value.

Definition 10 (Counterfactual consistency (Pan & Bareinboim, 2024)). An estimate is counterfactually consistent for Q when it lies in $[\underline{Q}, \overline{Q}]$. A proxy is consistent for a family of functionals when it matches the observational distribution and satisfies the corresponding bound for every functional in that family.

Counterfactual consistency is a feasibility criterion and not a solution concept. Several proxies can satisfy the same graphical and observational constraints while inducing different valid counterfactual responses. MSMS selects among such proxies with held-out response queries, restricting the candidates to those that pass the observational-fit screen of §3.6.

B.5 Intervention-relative partition and block factorization

Fix intervention targets $\boldsymbol{X} \subseteq \boldsymbol{V}$ and write $\mathcal{N}_X := \mathrm{ND}_{\mathcal{G}}(\boldsymbol{X})$ and $\mathcal{D}_X := \mathrm{De}_{\mathcal{G}}(\boldsymbol{X}) \setminus \boldsymbol{X}$, so that $\boldsymbol{V} = \boldsymbol{X} \dot\cup \mathcal{N}_X \dot\cup \mathcal{D}_X$. This partition is induced by the directed descendant relation and differs from the c-component partition, which bidirected connectivity induces. The interventional and counterfactual layers factorize along it, as Fig. 8a illustrates.



(a) Partition induced by the intervention targets. (b) Markovian descendant subgraph.

Figure 8: Causal graphs underlying the block factorization. (a) The partition $\boldsymbol{V} = \boldsymbol{X} \dot\cup \mathcal{N}_X \dot\cup \mathcal{D}_X$ induced by the intervention targets. (b) The descendant subgraph used in the Markovian reduction of Lem. 1.

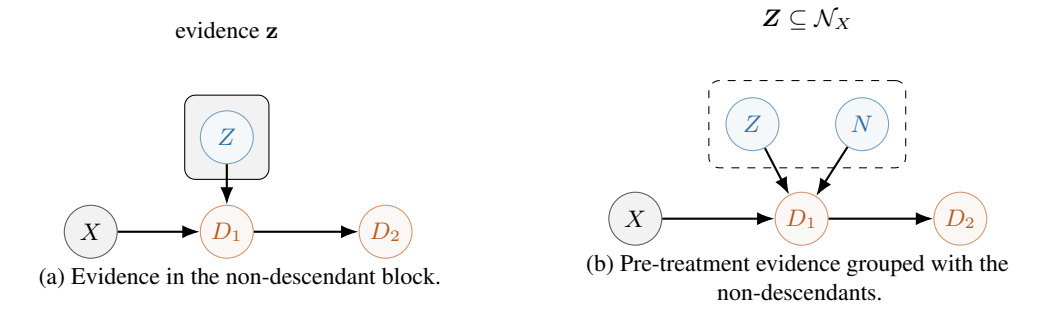
Lemma 1 (Block factorization). *For any SCM $\mathcal{M}$, intervention $do(\boldsymbol{X} = \boldsymbol{x})$, and evidence $\boldsymbol{Z} = \boldsymbol{z}$ with $P^{\mathcal{M}}(\boldsymbol{z}) > 0$, the counterfactual distribution of $\boldsymbol{V}$ factorizes as*

$$P^{\mathcal{M}}(\boldsymbol{v}_{\boldsymbol{X}}, \boldsymbol{v}_{\mathcal{N}_X}, \boldsymbol{v}_{\mathcal{D}_X} \mid do(\boldsymbol{x}), \boldsymbol{z}) = \mathbf{1}[\boldsymbol{v}_{\boldsymbol{X}} = \boldsymbol{x}]\, P^{\mathcal{M}}(\boldsymbol{v}_{\mathcal{N}_X} \mid \boldsymbol{z})\, P^{\mathcal{M}}(\boldsymbol{v}_{\mathcal{D}_X} \mid do(\boldsymbol{x}), \boldsymbol{v}_{\mathcal{N}_X}, \boldsymbol{z}), \quad (13)$$

where the conditional distributions are regular conditional probabilities.

Proof. Apply the chain rule to the joint distribution in the order $\boldsymbol{X}$, $\mathcal{N}_X$, $\mathcal{D}_X$. The intervention block equals $\boldsymbol{x}$ almost surely under $do(\boldsymbol{X} = \boldsymbol{x})$. Every $V \in \mathcal{N}_X$ is a non-descendant of $\boldsymbol{X}$, so $V_{\boldsymbol{x}}(\boldsymbol{u}) = V(\boldsymbol{u})$ for every $\boldsymbol{u}$, and the counterfactual distribution of the non-descendant block given the evidence equals its factual conditional distribution $P^{\mathcal{M}}(\boldsymbol{v}_{\mathcal{N}_X} \mid \boldsymbol{z})$. The descendant block retains its dependence on $do(\boldsymbol{x})$ and on the realized values of the other blocks. □

The three factors are the structural content of the stable and correction families in §3.2. The intervention block is a point mass at the requested values, and the non-descendant block keeps its factual conditional distribution. Responses read on $\boldsymbol{X} \cup \mathcal{N}_X$ are therefore determined by the query and the observational distribution, and they coincide for every SCM compatible with the graph. The descendant block $P^{\mathcal{M}}(\boldsymbol{v}_{\mathcal{D}_X} \mid do(\boldsymbol{x}), \boldsymbol{v}_{\mathcal{N}_X}, \boldsymbol{z})$ depends on the mechanisms of $\mathcal{M}$ beyond their observational implications, and observationally equivalent SCMs may differ on it. In the Markovian case it factorizes further along a topological order of $\mathcal{D}_X$ into the local conditionals $P^{\mathcal{M}}(v_j \mid \mathrm{pa}_j)$ evaluated at the intervened and realized parent values, with pre-treatment evidence absorbed into the non-descendant conditional. Fig. 9 depicts the two evidence configurations, and Fig. 10 visualizes the reduction.



(a) Evidence in the non-descendant block.

(b) Pre-treatment evidence grouped with the non-descendants.

Figure 9: Evidence configurations in the counterfactual decomposition. (a) Factual evidence that lies in the non-descendant block. (b) The pre-treatment setting $\boldsymbol{Z} \subseteq \mathcal{N}_X$, in which the descendant block of Lem. 1 reduces to L_1-level local conditionals.

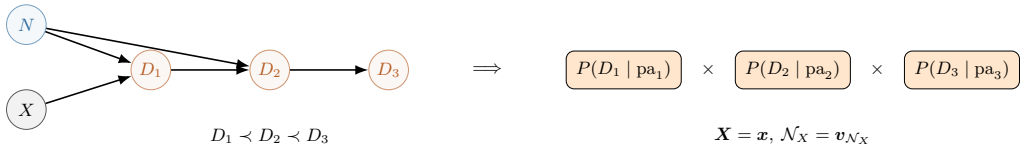


Figure 10: Markovian reduction of the descendant block. Once the intervention values and the non-descendant context are fixed, the descendant distribution factorizes into local L_1-level conditionals in topological order.

B.6 Counterfactual ambiguity under observational and interventional agreement

The ambiguity in Fig. 1 persists even when models agree exactly on observational and interventional distributions. Consider $A \to B$ with independent fair bits A, U_1, U_2 and assignments

$$\mathcal{M}_1 : \ B = A \oplus U_1, \qquad \mathcal{M}_2 : \ B = (1 - A)U_1 + AU_2,$$

where $\oplus$ denotes exclusive-or. Both models assign probability $1/4$ to each observed pair (A, B) and make B a fair bit under either intervention on A. For a factual case with $A = 0, B = 0$, however, both infer $U_1 = 0$, while U_2 remains uniformly distributed. Setting $A := 1$ then gives $B = 1$ with certainty under $\mathcal{M}_1$ and with probability $1/2$ under $\mathcal{M}_2$. A renderer that visibly distinguishes the values of B consequently produces different counterfactual image distributions. Thus, matching the observed and interventional distributions leaves unresolved how a particular instance changes across interventions. Selecting between such models requires additional information.

C Kernel background

This appendix collects the kernel background behind the response distances of §3.2. It recalls positive semidefinite kernels and reproducing kernel Hilbert spaces and defines kernel mean embeddings and the maximum mean discrepancy. It then states when the discrepancy separates distributions and relates these objects to the operational mean score and to the Markov kernels of causal-space formalisms.

C.1 Kernels and reproducing kernel Hilbert spaces

A positive semidefinite kernel on a set $\mathcal{X}$ is a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$ for every finite collection of points x_i and coefficients c_i.

Proposition 2 (Feature-map kernel)**.** *For any map $\phi : \mathcal{X} \to \mathcal{H}$ into a Hilbert space, $k_\phi(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ is positive semidefinite.*

Proof. $\sum_{i,j} c_i c_j k_\phi(x_i, x_j) = \| \sum_i c_i \phi(x_i) \|^2_{\mathcal{H}} \geq 0.$ □

A Hilbert space $\mathcal{H}_k$ of functions $f : \mathcal{X} \to \mathbb{R}$ is a reproducing kernel Hilbert space (RKHS) with kernel k if $k(x, \cdot) \in \mathcal{H}_k$ for every x and $f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}_k}$ for every $f \in \mathcal{H}_k$ (Aronszajn, 1950).

Theorem 3 (Moore–Aronszajn (Aronszajn, 1950))**.** *Every positive semidefinite kernel k on $\mathcal{X}$ induces a unique RKHS $\mathcal{H}_k$ with the reproducing property above.*

The space is the completion of the span of the sections $k(x, \cdot)$ under the inner product $\langle k(x, \cdot), k(x', \cdot)\rangle = k(x, x')$. For the linear kernel used in selection, ϕ is the response feature map itself and $\mathcal{H}_k$ is the feature space.

C.2 Kernel mean embeddings and maximum mean discrepancy

Kernel methods extend from points to probability measures by averaging feature elements.

Definition 11 (Kernel mean embedding and maximum mean discrepancy (Muandet et al., 2017; Sriperumbudur et al., 2010; Gretton et al., 2012))**.** For a bounded measurable kernel k with separable RKHS $\mathcal{H}_k$ and a probability measure P, the kernel mean embedding is the Bochner integral $\mu_P = \int k(x, \cdot)\, dP(x) \in \mathcal{H}_k$. The canonical feature map is strongly measurable under these conditions, and boundedness gives $\int \|k(x, \cdot)\|_{\mathcal{H}_k}\, dP(x) < \infty$. The maximum mean discrepancy between probability measures P and Q is $\mathrm{MMD}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}$.

Theorem 4 (MMD expansion (Gretton et al., 2012))**.** *For any probability measures P and Q,*

$$\mathrm{MMD}_k^2(P, Q) = \mathbb{E}_{P\otimes P}k(X, X') + \mathbb{E}_{Q\otimes Q}k(Y, Y') - 2\mathbb{E}_{P\otimes Q}k(X, Y).$$

Proof. Expand $\|\mu_P - \mu_Q\|_{\mathcal{H}_k}^2 = \langle\mu_P, \mu_P\rangle + \langle\mu_Q, \mu_Q\rangle - 2\langle\mu_P, \mu_Q\rangle$ and use $\langle\mu_P, \mu_Q\rangle = \iint\langle k(x, \cdot), k(y, \cdot)\rangle_{\mathcal{H}_k}\, dP(x)\, dQ(y) = \iint k(x, y)\, dP(x)\, dQ(y)$, which follows from the reproducing property. □

The measurability condition matters, since for $k(x, y) = \mathbf{1}[x = y]$ on an uncountable domain with a non-atomic distribution the canonical feature map is not essentially separably valued and the Bochner integral is undefined. A kernel is *characteristic* when $P \mapsto \mu_P$ is injective (Sriperumbudur et al., 2010), and consequently

$$\mathrm{MMD}_k(P, Q) = 0 \quad \Longleftrightarrow \quad P = Q \tag{14}$$

for characteristic k, which supports the distributional construction in Thm. 5. For each query, the main text embeds a response distribution by $\Phi_q(\mathsf{M}) = \mu_{\mathsf{T}_\mathsf{M}(q)}$ and sets $d_q = \mathrm{MMD}_{k_q}^2$, so the discrepancies D_{stb} and D_{corr} are averages of squared MMDs over the stable and correction query distributions.

C.3 Relation to the operational score and to Markov kernels

Relation to the operational score. The response-distribution geometry in Def. 12 compares the joint distributions of common queries and their measured responses with a characteristic-kernel maximum mean discrepancy (MMD), denoted D_μ^{char}. A characteristic kernel distinguishes probability distributions, whereas a finite vector of response means can leave distributional differences undetected. This is stronger than the operational linear-feature target D_b^{lin} of §3.2: a zero D_b^{lin} establishes equality of the selected feature means, not equality of response distributions or membership in the counterfactually correct class.

Finite-bank correspondence. Uniformly weighted, distinct queries with a common linear response feature space give a finite-bank joint-MMD proportional to the unnormalized mean score (Lem. 4), but arbitrary weights and repeated queries do not. The implementation and finite-query result therefore use the weighted per-query average directly.

The term *kernel* is also used for a Markov kernel $\Lambda : \mathcal{X} \times \Sigma_\mathcal{Y} \to [0, 1]$, which maps each input to a probability measure. This is the sense used by causal-space formalisms (Park et al., 2023; 2026), and it is distinct from a positive semidefinite RKHS kernel. In MSMS, response operators are Markov kernels, while k supplies the geometry used to compare their induced distributions.

D PROOFS FOR THE METHODOLOGY

This appendix proves the results of §3 in the order of the main text and collects the objects that the proofs use. §D.1 proves Prop. 1, §§D.2–D.4 define the mechanism space with its metric and kernelized structure, §§D.5–D.7 treat the proxy classes, and §§D.8–D.12 prove the finite-query, sample-complexity, and risk results.

D.1 STRUCTURAL FACTORIZATION OF THE EDITING RESPONSE

An editing query $q = (\{I\}, \{A\}, a, \{I\}, i)$ requests the edited-image distribution $\mathsf{T}^I_{\mathsf{M}}(q)$, and an image measurement map ψ_q induces the response distribution $\mathsf{T}_{\mathsf{M}}(q) := (\psi_q)_{\#}\mathsf{T}^I_{\mathsf{M}}(q)$, whose measurement index we suppress. The operational response distribution of §§3.4 and E.4 uses the scoring path of a candidate rather than its final rendered edits. These query-indexed distributions define the mechanism space of Def. 12, in which models are compared through responses without equating their parameters. The statement appears in Prop. 1 in the main text.

Proof of Proposition 1. Fix an editing query $q = (\{I\}, \boldsymbol{A}, \boldsymbol{a}, \{I\} \cup \boldsymbol{Z}_{\mathrm{aux}}, (i, \boldsymbol{z}_{\mathrm{aux}}))$ and write $u_{\mathrm{fac}} := U_{\boldsymbol{V}\setminus\{I\}}$, so that $U = (u_{\mathrm{fac}}, U_I)$ with U_I independent of u_{fac}. The counterfactual image arises from abduction of $P^{\mathcal{M}}(U \mid I = i, \boldsymbol{Z}_{\mathrm{aux}} = \boldsymbol{z}_{\mathrm{aux}})$, replacement of the assignments of $\boldsymbol{A}$ by the constants $\boldsymbol{a}$, and prediction $i' = f_I(\mathrm{Pa}_I^{\boldsymbol{a}}(u_{\mathrm{fac}}), U_I)$, where f_I and $P(U_I)$ are retained because I is terminal in $\mathcal{G}_{\mathcal{M}}$ and U_I enters only its assignment (Def. 1).

Since $\boldsymbol{Z}_{\mathrm{aux}}$ is determined by u_{fac}, the chain rule and the independence of U_I give, for a regular-conditional version and almost every evidence,

$$P^{\mathcal{M}}(U \mid I = i, \boldsymbol{z}_{\mathrm{aux}}) = P^{\mathcal{M}}(u_{\mathrm{fac}} \mid i, \boldsymbol{z}_{\mathrm{aux}})\, P^{\mathcal{M}}(U_I \mid i, u_{\mathrm{fac}}),$$

and $I = f_I(\mathrm{Pa}_I(u_{\mathrm{fac}}), U_I)$ makes the conditional law of (I, U_I) given u_{fac} depend on u_{fac} only through $\mathrm{Pa}_I(u_{\mathrm{fac}})$, so $P^{\mathcal{M}}(U_I \mid i, u_{\mathrm{fac}}) = P^{\mathcal{M}}(U_I \mid i, \mathrm{Pa}_I(u_{\mathrm{fac}}))$. Pushing this posterior forward through the prediction map and applying Fubini yields, for measurable $\mathcal{B} \subseteq \mathcal{D}_I$,

$$\mathsf{T}^I_{\mathcal{M}}(q)(\mathcal{B}) = \iint \mathbf{1}\big\{ f_I\big(\mathrm{Pa}_I^{\boldsymbol{a}}(u_{\mathrm{fac}}), u_I\big) \in \mathcal{B} \big\}\; dP^{\mathcal{M}}(u_I \mid i, \mathrm{Pa}_I(u_{\mathrm{fac}}))\, dP^{\mathcal{M}}(u_{\mathrm{fac}} \mid i, \boldsymbol{z}_{\mathrm{aux}}),$$

which is the displayed identity with the indicator written as a Dirac mass. □

D.2 FULL AXIOMATIC DEFINITION OF THE MECHANISM SPACE

Response operators are Markov kernels on standard Borel spaces with bounded response features, and Def. 3 defines the response-equivalence quotient. The definition below realizes this space with a characteristic kernel on the query domain of Def. 2, with (M1)–(M5) giving the construction and (M6) the operational linear target, which §D.9 uses directly. The common domain $\mathcal{Q}$ of Def. 3 is written $\mathcal{D}_{\mathcal{Q},\mathrm{align}}$ and consists of supported causal queries with their measurement specifications, and $\mathcal{S}$ is the disjoint union of the response spaces of the measurement types. For the stable and correction distributions of §3.2 we take $\mu = (\mu_{\mathrm{stb}} + \mu_{\mathrm{corr}})/2$, so both are absolutely continuous with respect to μ and the two discrepancies descend to $\mathfrak{T}_\mu$.

Definition 12 (Query domain and axiomatic mechanism space)**.** The query domain $\mathcal{D}_{\mathcal{Q}}$ and the editing family $\mathcal{D}_{\mathcal{Q},\mathrm{edit}}$ are those of Def. 2, with the context variables written $\boldsymbol{Z}_{\mathrm{aux}}$ when the image I is listed separately.

Fix a measurable response space $\mathcal{S} \supseteq \mathcal{D}_{\boldsymbol{Y}}$ for each $\boldsymbol{Y} \subseteq \boldsymbol{V}$, and an alignment domain $\mathcal{D}_{\mathcal{Q},\mathrm{align}} \subseteq \mathcal{D}_{\mathcal{Q}}(\mathsf{M}_0)$ with the stable/correction query families of §3. A mechanism space on $\mathcal{D}_{\mathcal{Q},\mathrm{align}}$ and $\mathcal{S}$ is specified by:

(M1) *Carrier*: a collection $\mathfrak{M}$ of queryable models. Each M supplies a measurable domain $\mathcal{D}_{\mathcal{Q}}(\mathsf{M})$ and a Markov-kernel response operator $\mathsf{T}_{\mathsf{M}} : \mathcal{D}_{\mathcal{Q}}(\mathsf{M}) \to \mathcal{P}(\mathcal{S})$. For a general SCM response, $\mathsf{T}_{\mathcal{M}}(q) = P^{\mathcal{M}}(\boldsymbol{Y}_{\boldsymbol{x}} \mid \boldsymbol{z})$. For an editing query we reserve $\mathsf{T}^I_{\mathcal{M}}(q)$ for the raw image kernel and set $\mathsf{T}_{\mathcal{M}}(q) = (\psi_q)_{\#}\mathsf{T}^I_{\mathcal{M}}(q)$ for the image-measurement specialization.

(M2) *Query measure*: a probability measure μ on $\mathcal{D}_{\mathcal{Q},\mathrm{align}}$ whose support lies in $\bigcap_{\mathsf{M}\in\mathfrak{M}} \mathcal{D}_{\mathcal{Q}}(\mathsf{M})$.

(M3) *Product kernel*: a bounded measurable positive-definite kernel $k = k_{\mathcal{Q}} \otimes k_{\mathcal{S}}$ that is characteristic on the joint query–response domain. Characteristicness of the two factors alone is not used as a shortcut for this joint condition; the required product-kernel conditions are those in Szabó & Sriperumbudur (2018). On non-atomic μ the Kronecker kernel $k_{\mathcal{Q}}(q, q') = \mathbf{1}[q = q']$ is excluded because it collapses the corresponding joint-MMD to 0; see Rem. 1 for the empirical-bank case.

(M4) *Joint embedding*: $\Psi_\mu(\mathsf{M}) := \mu_{\Gamma^\mu_{\mathsf{M}}} \in \mathcal{H}_k$, where $\Gamma^\mu_{\mathsf{M}} := \mu \otimes \mathsf{T}_{\mathsf{M}}$ is the joint query–response distribution.

(M5) *Characteristic discrepancy*: the response-distribution discrepancy is the squared joint-MMD on Γ^μ,

$$D^{\mathrm{char}}_\mu(\mathsf{M}, \mathsf{M}') := \left\|\Psi_\mu(\mathsf{M}) - \Psi_\mu(\mathsf{M}')\right\|^2_{\mathcal{H}_k}.$$

D^{char}_μ is non-negative but is not a metric because it is a squared RKHS norm, and Prop. 3 gives the integrated unsquared metric on the response-distribution quotient.

(M6) *Operational linear target and estimator*: for a finite bank with normalized weights $\pi_r = w_r / \sum_s w_s$ and response feature map φ_r, define

$$D^{\mathrm{lin}}_{\mathrm{avg}}(\mathsf{M}, \mathsf{M}') := \sum_{r=1}^{m} \pi_r \, \frac{\|\mathbb{E}\varphi_r(S_{\mathsf{M}}) - \mathbb{E}\varphi_r(S_{\mathsf{M}'})\|_2^2}{d_{\varphi_r}}.$$

The operational $\widehat{D}^{\mathrm{avg}}$ replaces each population term by its per-query V- or U-statistic. It targets $D^{\mathrm{lin}}_{\mathrm{avg}}$, not D^{char}_μ. With uniform weights, a distinct bank, and a common linear response kernel, the unnormalized version is proportional to finite-bank joint-MMD by Lem. 4. Arbitrary weights enter the per-query average linearly but enter joint-MMD quadratically, and repeated queries have the same mismatch; no general proportionality is claimed. The finite-query theorem is stated directly for $D^{\mathrm{lin}}_{\mathrm{avg}}$ and $\widehat{D}^{\mathrm{avg}}$. Its per-query U-statistic terms are unbiased when response pairs are conditionally i.i.d. across their sample index. Queries from one observational case need not be independent; Cor. 3 therefore groups them into per-case averages and assumes independence only across held-out cases.

The **mechanism space** is the quotient of queryable models under Ψ_μ-equality.

D.3 Metric structure of mechanism space

Lemma 2 (Integral metrics descend to the quotient)**.** *Let $(\mathcal{A}, \Sigma, \nu)$ be a probability space, let $(\mathcal{X}, d)$ be a* separable *metric space, and let $f, g, h : \mathcal{A} \to \mathcal{X}$ be measurable. Define*

$$D(f, g) := \int_{\mathcal{A}} d(f(a), g(a))\, d\nu(a) \in [0, \infty].$$

Separability makes $a \mapsto d(f(a), g(a))$ measurable, so D is well defined as an extended pseudometric. It is finite on the class of maps that are ν-integrable relative to a base point, $\mathcal{L}^1_\nu(\mathcal{X}) := \{f : \int d(f(a), x_0)\, d\nu(a) < \infty\}$ for some (equivalently every) $x_0 \in \mathcal{X}$, and when d is bounded this class contains every measurable map. Then D is a pseudometric on $\mathcal{L}^1_\nu(\mathcal{X})$ (and an extended pseudometric on all measurable maps). Moreover,

$$D(f, g) = 0 \iff f = g \quad \nu\text{-a.e.}$$

Hence D induces a metric on the quotient by ν-a.e. equality.

Proof. Separability of $\mathcal{X}$ makes the product σ-algebra on $\mathcal{X} \times \mathcal{X}$ Borel, so $a \mapsto d(f(a), g(a))$ is measurable, and $d(f, g) \le d(f, x_0) + d(x_0, g)$ gives finiteness on $\mathcal{L}^1_\nu(\mathcal{X})$. Integrating the pointwise triangle inequality gives $D(f, h) \le D(f, g) + D(g, h)$. If $D(f, g) = 0$, Markov's inequality gives $\nu\{d(f, g) \ge 1/n\} \le n\, D(f, g) = 0$ for every n, so $f = g$ ν-a.e., and the converse is immediate. □

Proposition 3 (Mechanism space is a metric space)**.** *Let $d_{\mathcal{P}}$ be a metric on $\mathcal{P}(\mathcal{S})$ that is separable and either bounded or restricted to response operators in the integrable class $\mathcal{L}^1_\mu(\mathcal{P}(\mathcal{S}))$ of Lem. 2*

(the MMD metric of a bounded kernel, used below, is bounded), and let μ be a probability measure on $\mathcal{D}_{\mathcal{Q},\mathrm{align}}$. Then the integrated response discrepancy

$$\Delta_\mu(\mathsf{M},\mathsf{M}') = \int_{\mathcal{D}_{\mathcal{Q},\mathrm{align}}} d_{\mathcal{P}}(\mathsf{T}_{\mathsf{M}}(q),\mathsf{T}_{\mathsf{M}'}(q))\; d\mu(q) \tag{15}$$

depends only on the μ-a.e. equivalence classes of the response operators and induces a metric on the quotient space $\mathfrak{T}_\mu$.

Proof. Apply Lem. 2 with $\mathcal{A} = \mathcal{D}_{\mathcal{Q},\mathrm{align}}$, $\nu = \mu$, $\mathcal{X} = \mathcal{P}(\mathcal{S})$, $d = d_{\mathcal{P}}$, and $f(q) = \mathsf{T}_{\mathsf{M}}(q)$, $g(q) = \mathsf{T}_{\mathsf{M}'}(q)$. Then the integral discrepancy is a pseudometric on measurable response operators and becomes a metric after quotienting by μ-a.e. equality. This quotient is precisely $\mathfrak{T}_\mu$. □

D.4 Kernelized realization on joint query-response distributions

For the distributional construction, we pass from response operators to their joint query-response distributions. This branch studies a richer comparison than the operational linear-mean score, and Lem. 4 and Rem. 2 give their precise finite-bank relationship. Fix μ on $\mathcal{D}_{\mathcal{Q},\mathrm{align}}$. For a queryable model M, define the joint query-response distribution Γ_{M} on $\mathcal{D}_{\mathcal{Q},\mathrm{align}} \times \mathcal{S}$ by

$$\Gamma_{\mathsf{M}}(A \times B) := \int_A \mathsf{T}_{\mathsf{M}}(q)(B)\, d\mu(q)$$

for measurable rectangles $A \times B$.

Lemma 3 (Equivalence of response operators and joint distributions)**.** *Assume that $\mathcal{S}$ is standard Borel. Then*

$$\mathsf{T}_{\mathsf{M}}(q) = \mathsf{T}_{\mathsf{M}'}(q) \quad \mu\text{-a.e. } q \in \mathcal{D}_{\mathcal{Q},\mathrm{align}} \iff \Gamma_{\mathsf{M}} = \Gamma_{\mathsf{M}'}.$$

Proof. If $\mathsf{T}_{\mathsf{M}}(q) = \mathsf{T}_{\mathsf{M}'}(q)$ for μ-a.e. q, then for every measurable rectangle $A \times B$,

$$\Gamma_{\mathsf{M}}(A \times B) = \int_A \mathsf{T}_{\mathsf{M}}(q)(B)\, d\mu(q) = \int_A \mathsf{T}_{\mathsf{M}'}(q)(B)\, d\mu(q) = \Gamma_{\mathsf{M}'}(A \times B).$$

Hence $\Gamma_{\mathsf{M}} = \Gamma_{\mathsf{M}'}$.

Conversely, assume $\Gamma_{\mathsf{M}} = \Gamma_{\mathsf{M}'}$. Let $\mathcal{C}$ be a countable π-system generating the Borel σ-field of $\mathcal{S}$. For each $B \in \mathcal{C}$ and measurable $A \subseteq \mathcal{D}_{\mathcal{Q},\mathrm{align}}$,

$$\int_A \mathsf{T}_{\mathsf{M}}(q)(B)\, d\mu(q) = \Gamma_{\mathsf{M}}(A \times B) = \Gamma_{\mathsf{M}'}(A \times B) = \int_A \mathsf{T}_{\mathsf{M}'}(q)(B)\, d\mu(q).$$

Therefore $\mathsf{T}_{\mathsf{M}}(q)(B) = \mathsf{T}_{\mathsf{M}'}(q)(B)$ for μ-a.e. q, and the standing measurability assumption in Def. 12 ensures that $q \mapsto \mathsf{T}_{\mathsf{M}}(q)(B)$ is a measurable $[0,1]$-valued function (so T_{M} is a Markov kernel in the standard sense), so the μ-a.e. equality is well-posed. Intersecting over the countable class $\mathcal{C}$, there exists a μ-full set on which the two probability measures agree on $\mathcal{C}$. Since $\mathcal{S}$ is standard Borel, equality on a countable generating π-system implies equality of the full probability measures. Hence $\mathsf{T}_{\mathsf{M}}(q) = \mathsf{T}_{\mathsf{M}'}(q)$ for μ-a.e. q. □

Theorem 5 (Kernelized realization of mechanism discrepancy)**.** *Let k be a bounded measurable kernel on $\mathcal{D}_{\mathcal{Q},\mathrm{align}} \times \mathcal{S}$ whose RKHS $\mathcal{H}_k$ is separable (so its canonical feature map $x \mapsto k(\cdot, x)$ is strongly measurable; a continuous kernel on a separable domain suffices), and let $\Phi_k(\mathsf{M}) := \mu_{\Gamma_{\mathsf{M}}} \in \mathcal{H}_k$ denote the kernel mean embedding of the joint query-response distribution Γ_{M} (well-defined as a Bochner integral under these conditions; cf. Def. 11). Then the induced model kernel*

$$\kappa_k(\mathsf{M},\mathsf{M}') := \langle \Phi_k(\mathsf{M}), \Phi_k(\mathsf{M}') \rangle_{\mathcal{H}_k}$$

is positive semidefinite, and

$$\|\Phi_k(\mathsf{M}) - \Phi_k(\mathsf{M}')\|^2_{\mathcal{H}_k} = \mathrm{MMD}^2_k(\Gamma_{\mathsf{M}}, \Gamma_{\mathsf{M}'}).$$

If, in addition, k is characteristic on the joint domain $\mathcal{D}_{\mathcal{Q},\mathrm{align}} \times \mathcal{S}$ (which for product kernels $k = k_{\mathcal{Q}} \otimes k_{\mathcal{S}}$ requires both $k_{\mathcal{Q}}$ and $k_{\mathcal{S}}$ to be characteristic together with the joint characteristic-ness conditions of Szabó & Sriperumbudur, 2018), then

$$\|\Phi_k(\mathsf{M}) - \Phi_k(\mathsf{M}')\|^2_{\mathcal{H}_k} = 0 \iff \mathsf{T}_{\mathsf{M}}(q) = \mathsf{T}_{\mathsf{M}'}(q) \quad \mu\text{-a.e. } q.$$

The implemented Kronecker query kernel $k_{\mathcal{Q}}(q,q') = \mathbf{1}[q = q']$ does not satisfy this hypothesis on a continuous query domain, as Rem. 1 discusses.

Proof. Boundedness of k gives $\int \|k(\cdot,x)\|_{\mathcal{H}_k}\, d\Gamma_{\mathsf{M}}(x) = \int \sqrt{k(x,x)}\, d\Gamma_{\mathsf{M}}(x) < \infty$, and separability of $\mathcal{H}_k$ together with measurability of k makes the feature map strongly measurable (Pettis), so the kernel mean embedding $\Phi_k(\mathsf{M}) = \int k(\cdot,(q,s))\, d\Gamma_{\mathsf{M}}(q,s)$ exists in $\mathcal{H}_k$ as a Bochner integral (Def. 11). Boundedness and measurability alone do not suffice: for the Kronecker kernel $k(x,y) = \mathbf{1}[x=y]$ on an uncountable domain with a non-atomic distribution, the feature vectors $k(\cdot,x)$ are mutually orthogonal, the feature map is not essentially separably valued, and the Bochner integral is undefined.

Under the stated separability conditions, positive semidefiniteness of κ_k follows from Prop. 2 applied to the feature map Φ_k.

By Def. 11, the MMD between Γ_{M} and $\Gamma_{\mathsf{M}'}$ is the RKHS distance between their kernel mean embeddings, hence

$$\mathrm{MMD}_k^2(\Gamma_{\mathsf{M}}, \Gamma_{\mathsf{M}'}) = \|\Phi_k(\mathsf{M}) - \Phi_k(\mathsf{M}')\|_{\mathcal{H}_k}^2.$$

If k is characteristic on the joint domain (in the sense above), Eq. (14) gives

$$\|\Phi_k(\mathsf{M}) - \Phi_k(\mathsf{M}')\|_{\mathcal{H}_k}^2 = 0 \iff \Gamma_{\mathsf{M}} = \Gamma_{\mathsf{M}'}.$$

The final equivalence then follows from Lem. 3. The closing claim about the Kronecker $k_{\mathcal{Q}}$ is the failure-of-characteristic-ness assertion in Rem. 1: $k_{\mathcal{Q}}(q,q') = \mathbf{1}[q = q']$ corresponds to a counting-measure embedding that does not separate continuous-support measures. □

Remark 1 (Characteristic kernel and the implementation)**.** The reported implementation (§E) uses the Kronecker query kernel $k_{\mathcal{Q}}(q,q') = \mathbf{1}[q = q']$ and the linear response kernel $k_{\mathcal{S}}(s,s') = \langle \varphi(s), \varphi(s') \rangle$. The former distinguishes the finite empirical query support without being characteristic on the continuous domain $\mathcal{D}_{\mathcal{Q},\mathrm{align}}$, and the latter identifies response means only. For instance, with $\varphi(s) = s$ the distributions $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ and δ_0 have zero linear discrepancy although a characteristic response kernel separates them, so the equivalence in Thm. 5 does not apply to the reported scores.

Lemma 4 (Mean-MSE on a fixed feature map equals empirical squared joint-MMD with a Kronecker query kernel)**.** *Let $\varphi : \mathcal{S} \to \mathbb{R}^{d_\varphi}$ be a fixed measurable feature map with $\sup_s \|\varphi(s)\|_2 < \infty$. Define*

$$k_{\mathcal{Q}}(q,q') := \mathbf{1}[q = q'], \qquad k_{\mathcal{S}}(s,s') := \langle \varphi(s), \varphi(s') \rangle_{\mathbb{R}^{d_\varphi}}.$$

Let $\{q_r\}_{r=1}^m \subseteq \mathcal{D}_{\mathcal{Q},\mathrm{align}}$ be a finite query bank with distinct *entries and empirical query measure $\hat{\mu}_m = \frac{1}{m}\sum_r \delta_{q_r}$, and let $\Gamma_{\mathsf{M}}^{\hat{\mu}_m} := \hat{\mu}_m \otimes \mathsf{T}_{\mathsf{M}}$. Write $\mu_{\mathsf{T}_{\mathsf{M}}(q_r)} := \mathbb{E}_{S\sim\mathsf{T}_{\mathsf{M}}(q_r)}[\varphi(S)] \in \mathbb{R}^{d_\varphi}$ for the per-query population mean embedding. Then*

$$\left\|\mu_{\Gamma_{\mathsf{M}}^{\hat{\mu}_m}} - \mu_{\Gamma_{\mathsf{M}_0}^{\hat{\mu}_m}}\right\|_{\mathcal{H}_{k_{\mathcal{Q}}\otimes k_{\mathcal{S}}}}^2 = \frac{1}{m^2}\sum_{r=1}^m \left\|\mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_r)}\right\|_2^2. \tag{16}$$

In particular, the population per-query mean-MSE $D_{\hat{\mu}_m}^{\text{mean-MSE}}(\mathsf{M},\mathsf{M}_0) := \frac{1}{m}\sum_{r=1}^m \left\|\mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_r)}\right\|_2^2$ satisfies

$$D_{\hat{\mu}_m}^{\text{mean-MSE}}(\mathsf{M},\mathsf{M}_0) = m \cdot \left\|\mu_{\Gamma_{\mathsf{M}}^{\hat{\mu}_m}} - \mu_{\Gamma_{\mathsf{M}_0}^{\hat{\mu}_m}}\right\|_{\mathcal{H}_{k_{\mathcal{Q}}\otimes k_{\mathcal{S}}}}^2. \tag{17}$$

Proof. The product kernel $k_{\mathcal{Q}} \otimes k_{\mathcal{S}}$ has reproducing kernel mean embedding

$$\mu_\Gamma\big((q,s)\big) = \mathbb{E}_{(Q,S)\sim\Gamma}\big[k_{\mathcal{Q}}(Q,q)\, k_{\mathcal{S}}(S,s)\big].$$

Writing μ_Γ as an element of the tensor-product RKHS $\mathcal{H}_{k_{\mathcal{Q}}} \otimes \mathcal{H}_{k_{\mathcal{S}}}$, for $\Gamma_{\mathsf{M}}^{\hat{\mu}_m} = \hat{\mu}_m \otimes \mathsf{T}_{\mathsf{M}}$ we have

$$\mu_{\Gamma_{\mathsf{M}}^{\hat{\mu}_m}} = \frac{1}{m}\sum_{r=1}^m k_{\mathcal{Q}}(q_r,\cdot) \otimes \mu_{\mathsf{T}_{\mathsf{M}}(q_r)}.$$

Using bilinearity of the inner product and $\langle k_{\mathcal{Q}}(q_r,\cdot), k_{\mathcal{Q}}(q_s,\cdot)\rangle_{\mathcal{H}_{k_{\mathcal{Q}}}} = k_{\mathcal{Q}}(q_r,q_s)$,

$$\left\|\mu_{\Gamma_{\mathsf{M}}^{\hat{\mu}_m}} - \mu_{\Gamma_{\mathsf{M}_0}^{\hat{\mu}_m}}\right\|_{\mathcal{H}_k}^2 = \frac{1}{m^2}\sum_{r,s=1}^m k_{\mathcal{Q}}(q_r,q_s)\left\langle \mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_r)},\, \mu_{\mathsf{T}_{\mathsf{M}}(q_s)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_s)}\right\rangle_2.$$

The Kronecker structure $k_{\mathcal{Q}}(q_r, q_s) = \mathbf{1}[q_r = q_s]$ together with distinctness of the bank removes the cross-query terms, so only $r = s$ survives,

$$= \frac{1}{m^2} \sum_{r=1}^{m} \left\| \mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_r)} \right\|_2^2,$$

which is (16). Multiplying both sides by m gives (17). □

Remark 2 (Unequal weights and repeats in the query bank)**.** Lem. 4 assumes a uniform bank of distinct queries. With R distinct queries of multiplicities n_r and $m = \sum_r n_r$, the empirical squared joint-MMD weights the response-mean differences quadratically in the multiplicities,

$$\left\| \mu_{\Gamma_{\mathsf{M}}^{\hat{\mu}_m}} - \mu_{\Gamma_{\mathsf{M}_0}^{\hat{\mu}_m}} \right\|_{\mathcal{H}_{k_{\mathcal{Q}} \otimes k_{\mathcal{S}}}}^2 = \frac{1}{m^2} \sum_{r=1}^{R} n_r^2 \left\| \mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_r)} \right\|_2^2,$$

whereas the empirical mean-MSE on the same bank weights them linearly,

$$D_{\hat{\mu}_m}^{\text{mean-MSE}}(\mathsf{M}, \mathsf{M}_0) = \frac{1}{m} \sum_{r=1}^{R} n_r \left\| \mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_r)} \right\|_2^2.$$

so the identity (17) fails unless all n_r are equal, and the same holds for distinct queries with unequal masses π_r. Training samples query mini-batches with replacement, so the per-query averaged mean-MSE $\widehat{D}_{\text{avg}}(\mathsf{M}, \mathsf{M}_0) := \mathbb{E}_{q \sim \hat{\mu}_m} \left\| \hat{\mu}_{\mathsf{T}_{\mathsf{M}}(q)} - \hat{\mu}_{\mathsf{T}_{\mathsf{M}_0}(q)} \right\|_2^2$ is a separate functional from the bank-level joint-MMD. The finite-query results are stated for $\widehat{D}_{\text{avg}}$, whose per-query U-statistic is unbiased and whose concentration argument treats each case as one bounded summand.

D.5 Proxy classes and population selector

The empirical screen of Algorithm 1 tests finite-data eligibility, while the population risk result in §3.5 assumes exact observational fit and a counterfactually correct model. Thm. 6 analyzes selection in a finite eligible pool, and Thm. 2 concerns this idealized population selector.

The analysis uses the classes of the main text. The exact feasible class $\mathfrak{M}_{\text{feas}}$ is that of Eq. (3), the finite pool C contains the saved models that pass the held-out factor-fit screen $\widehat{D}_{\text{obs}} \le \epsilon_{\text{obs}}$ of Algorithm 1, the zero-risk subclass $\mathfrak{M}_{\text{cf}}$ is that of §3.5, and the retained class $\mathfrak{M}_\tau$ and the population selector are those of Eq. (4). Throughout the proofs, D_{stb} and D_{corr} denote the linear-feature discrepancies D^{lin} of §D.8, and whenever a population selector is invoked the infimum is assumed to be attained.

D.6 Observational feasibility and optional bounded functionals

These results connect exact observational feasibility to counterfactual-consistency bounds. The reported selector does not use the optional bounded-functional penalty.

Lemma 5 (Exact observational feasibility implies bounded counterfactual feasibility)**.** *Let Q be a scalar counterfactual functional whose optimal bounds are defined under $(\mathcal{G}, P(\boldsymbol{V}))$. If $\hat{\mathcal{M}}(\boldsymbol{\theta}) \in \mathfrak{M}_{\text{feas}}$, then*

$$Q(\hat{\mathcal{M}}(\boldsymbol{\theta})) \in \left[\underline{Q}(\mathcal{G}, P(\boldsymbol{V})), \overline{Q}(\mathcal{G}, P(\boldsymbol{V})) \right].$$

Proof. If $\hat{\mathcal{M}}(\boldsymbol{\theta}) \in \mathfrak{M}_{\text{feas}}$, then by definition

$$\hat{\mathcal{M}}(\boldsymbol{\theta}) \in \Omega_{\boldsymbol{\theta}}(\mathcal{G}) \quad \text{and} \quad P^{\hat{\mathcal{M}}(\boldsymbol{\theta})}(\boldsymbol{V}) = P(\boldsymbol{V}).$$

By Def. 9, every graph-constrained NCM is an SCM with graph $\mathcal{G}$, so $\Omega_{\boldsymbol{\theta}}(\mathcal{G}) \subseteq \Omega(\mathcal{G})$, and combined with the observational match just shown, $\hat{\mathcal{M}}(\boldsymbol{\theta}) \in \Omega(\mathcal{G}, P(\boldsymbol{V}))$. By the definitions of $\underline{Q}(\mathcal{G}, P(\boldsymbol{V}))$ and $\overline{Q}(\mathcal{G}, P(\boldsymbol{V}))$ as the infimum and supremum of $Q(\mathcal{M})$ over $\Omega(\mathcal{G}, P(\boldsymbol{V}))$, it follows immediately that

$$\underline{Q}(\mathcal{G}, P(\boldsymbol{V})) \le Q(\hat{\mathcal{M}}(\boldsymbol{\theta})) \le \overline{Q}(\mathcal{G}, P(\boldsymbol{V})).$$

□

Corollary 1 (Bounded-functional penalty vanishes on the exact feasible set)**.** *Let $\mathfrak{Q}_{\mathrm{bnd}}$ be any finite family of scalar counterfactual functionals whose optimal bounds are defined under $(\mathcal{G}, P(\boldsymbol{V}))$, and define*

$$\mathcal{L}_{\mathrm{ctf}}^{\mathfrak{Q}_{\mathrm{bnd}}}(\boldsymbol{\theta}) := \sum_{Q \in \mathfrak{Q}_{\mathrm{bnd}}} \left([\underline{Q} - Q(\hat{\mathcal{M}}(\boldsymbol{\theta}))]_+ + [Q(\hat{\mathcal{M}}(\boldsymbol{\theta})) - \overline{Q}]_+ \right).$$

Then

$$\hat{\mathcal{M}}(\boldsymbol{\theta}) \in \mathfrak{M}_{\mathrm{feas}} \implies \mathcal{L}_{\mathrm{ctf}}^{\mathfrak{Q}_{\mathrm{bnd}}}(\boldsymbol{\theta}) = 0.$$

Proof. Fix $\hat{\mathcal{M}}(\boldsymbol{\theta}) \in \mathfrak{M}_{\mathrm{feas}}$. By Lem. 5,

$$\underline{Q}(\mathcal{G}, P(\boldsymbol{V})) \le Q(\hat{\mathcal{M}}(\boldsymbol{\theta})) \le \overline{Q}(\mathcal{G}, P(\boldsymbol{V})) \qquad \forall Q \in \mathfrak{Q}_{\mathrm{bnd}}.$$

Hence each positive-part violation term in the bounded-functional penalty vanishes, and summing over $Q \in \mathfrak{Q}_{\mathrm{bnd}}$ gives $\mathcal{L}_{\mathrm{ctf}}^{\mathfrak{Q}_{\mathrm{bnd}}}(\boldsymbol{\theta}) = 0$. □

D.7 Nonemptiness under exact realizability

Exact task realizability supplies equality of joint factual–counterfactual distributions. The second hypothesis below is needed to transfer that equality to the editing queries being evaluated.

Lemma 6 (Nonemptiness under exact task realizability)**.** *Let $\mathcal{G} = \mathcal{G}_{\mathcal{M}^\star}$. Assume:*

(H1) Exact task realizability. *The idealized family $\Omega_{\boldsymbol{\theta}}(\mathcal{G})$ (Def. 9) contains a finite-parameter exact L_3-realizer of $\mathcal{M}^\star$. The expressiveness result of Xia et al. (2023) motivates this assumption for finite domains and does not establish it for the continuous image domain or the district-level parametrization used here.*

(H2) Compatibility of the editing measure. *Write $\pi_{\mathrm{cond}}(q) := (I = i, \boldsymbol{Z}_{\mathrm{aux}} = \boldsymbol{z}_{\mathrm{aux}})$ and $\pi_{\mathrm{int}}(q) := (\boldsymbol{A}, \boldsymbol{a})$ for an editing query $q = (\{I\}, \boldsymbol{A}, \boldsymbol{a}, \{I\} \cup \boldsymbol{Z}_{\mathrm{aux}}, (i, \boldsymbol{z}_{\mathrm{aux}}))$. The conditioning coordinate must be compatible with the factual distribution* for each intervention value*, not only marginally: either (a) the set of intervention values $\{(\boldsymbol{A}, \boldsymbol{a})\}$ charged by ν_{edit} is countable and $(\pi_{\mathrm{cond}})_* \nu_{\mathrm{edit}}(\cdot \mid \boldsymbol{A}, \boldsymbol{a}) \ll P^{\mathcal{M}^\star}(I, \boldsymbol{Z}_{\mathrm{aux}})$ for every charged value (this is the case for the categorical interventions of the reported benchmarks); or (b) the regular conditional distributions below admit a version that is jointly measurable in $(\boldsymbol{a}, i, \boldsymbol{z}_{\mathrm{aux}})$, and the disintegration of ν_{edit} along π_{int} satisfies $(\pi_{\mathrm{cond}})_* \nu_{\mathrm{edit}}(\cdot \mid \boldsymbol{A}, \boldsymbol{a}) \ll P^{\mathcal{M}^\star}(I, \boldsymbol{Z}_{\mathrm{aux}})$ for $(\pi_{\mathrm{int}})_* \nu_{\mathrm{edit}}$-almost every $(\boldsymbol{A}, \boldsymbol{a})$. Marginal absolute continuity of $(\pi_{\mathrm{cond}})_* \nu_{\mathrm{edit}}$ alone is not sufficient: with a continuous conditioning coordinate, a query distribution that couples the intervention value to the case (for instance $\boldsymbol{a} = \boldsymbol{z}_{\mathrm{aux}}$) can concentrate on the union of the intervention-dependent null sets on which two versions of the conditionals may disagree. Regular conditional distributions of $I_{\boldsymbol{a}}$ given $(I, \boldsymbol{Z}_{\mathrm{aux}})$ exist and are unique up to a $P^{\mathcal{M}^\star}(I, \boldsymbol{Z}_{\mathrm{aux}})$-null set by the standing standard-Borel hypothesis on $\mathcal{D}_{\boldsymbol{V}}$.*

Then there exists $\boldsymbol{\theta}^\star \in \mathcal{D}_{\boldsymbol{\theta}}$ with $\hat{\mathcal{M}}(\boldsymbol{\theta}^\star) \in \Omega_{\boldsymbol{\theta}}(\mathcal{G})$ such that, for every intervention $\boldsymbol{A} \subseteq \boldsymbol{V} \setminus \{I\}$ and $\boldsymbol{a} \in \mathcal{D}_{\boldsymbol{A}}$,

$$P^{\hat{\mathcal{M}}(\boldsymbol{\theta}^\star)}(\boldsymbol{V}_{\boldsymbol{a}}, \boldsymbol{V}) = P^{\mathcal{M}^\star}(\boldsymbol{V}_{\boldsymbol{a}}, \boldsymbol{V}). \tag{18}$$

Its raw editing kernels satisfy $\mathsf{T}^I_{\hat{\mathcal{M}}(\boldsymbol{\theta}^\star)}(q) = \mathsf{T}^I_{\mathcal{M}^\star}(q)$ for ν_{edit}-almost every $q \in \mathcal{D}_{\mathcal{Q},\mathrm{edit}}$. Consequently, $\hat{\mathcal{M}}(\boldsymbol{\theta}^\star) \in \mathfrak{M}_{\mathrm{cf}}$ and $\mathfrak{M}_{\mathrm{cf}} \neq \emptyset$.

Proof. **Step 1: realizability.** Hypothesis (H1) supplies $\boldsymbol{\theta}^\star \in \mathcal{D}_{\boldsymbol{\theta}}$ for which Eq. (18) holds.

Step 2: feasibility. Taking the empty intervention $\boldsymbol{A} = \varnothing$ collapses Eq. (18) to the observational identity $P^{\hat{\mathcal{M}}(\boldsymbol{\theta}^\star)}(\boldsymbol{V}) = P^{\mathcal{M}^\star}(\boldsymbol{V}) = P(\boldsymbol{V})$, so $\hat{\mathcal{M}}(\boldsymbol{\theta}^\star) \in \mathfrak{M}_{\mathrm{feas}}$.

Step 3: conditional descent (ν_{edit}-a.e.). For an editing query $q = (\{I\}, \boldsymbol{A}, \boldsymbol{a}, \{I\} \cup \boldsymbol{Z}_{\mathrm{aux}}, (i, \boldsymbol{z}_{\mathrm{aux}})) \in \mathcal{D}_{\mathcal{Q},\mathrm{edit}}$, the raw image kernel $\mathsf{T}^I_{\mathcal{M}}(q)$ is the regular conditional distribution of $I_{\boldsymbol{a}}$ given $(I, \boldsymbol{Z}_{\mathrm{aux}}) = (i, \boldsymbol{z}_{\mathrm{aux}})$ under the joint twin-network distribution. Standard-Borel structure on $\mathcal{D}_{\boldsymbol{V}}$ guarantees that such regular conditionals exist and are unique only up to a $P^{\mathcal{M}^\star}(I, \boldsymbol{Z}_{\mathrm{aux}})$-null set. Since the joint twin-network distributions under $\hat{\mathcal{M}}(\boldsymbol{\theta}^\star)$ and $\mathcal{M}^\star$ coincide by Step 1, versions

of their regular conditionals can therefore be chosen to agree $P^{\mathcal{M}^\star}(I, \boldsymbol{Z}_{\mathrm{aux}})$-almost everywhere, not necessarily at every point of its topological support. For each intervention value $(\boldsymbol{A}, \boldsymbol{a})$ let $N_{\boldsymbol{a}}$ denote the $P^{\mathcal{M}^\star}(I, \boldsymbol{Z}_{\mathrm{aux}})$-null set outside which the two chosen versions agree, and let $E := \{q : \pi_{\mathrm{cond}}(q) \in N_{\pi_{\mathrm{int}}(q)}\}$ be the set of queries on which the kernels may differ. Under (H2a), E is a countable union of the sets $\{\pi_{\mathrm{int}}(q) = (\boldsymbol{A}, \boldsymbol{a})\} \cap \pi_{\mathrm{cond}}^{-1}(N_{\boldsymbol{a}})$, each of which is ν_{edit}-null by the conditional absolute continuity in (H2a); hence $\nu_{\mathrm{edit}}(E) = 0$. Under (H2b), joint measurability makes E measurable and Fubini's theorem gives $\nu_{\mathrm{edit}}(E) = \int (\pi_{\mathrm{cond}})_* \nu_{\mathrm{edit}}(N_{\boldsymbol{a}} \mid \boldsymbol{A}, \boldsymbol{a})\, (\pi_{\mathrm{int}})_* \nu_{\mathrm{edit}}(\mathrm{d}(\boldsymbol{A}, \boldsymbol{a})) = 0$. In either case $\mathsf{T}^I_{\hat{\mathcal{M}}(\boldsymbol{\theta}^\star)}(q) = \mathsf{T}^I_{\mathcal{M}^\star}(q)$ for ν_{edit}-a.e. $q \in \mathcal{D}_{\mathcal{Q},\mathrm{edit}}$.

Step 4: closure. Steps 2 and 3 give $\hat{\mathcal{M}}(\boldsymbol{\theta}^\star) \in \mathfrak{M}_{\mathrm{feas}}$ and the ν_{edit}-a.e. equality of the editing kernels, which is the membership condition of $\mathfrak{M}_{\mathrm{cf}}$ in §D.5, so $\mathfrak{M}_{\mathrm{cf}} \neq \varnothing$. □

Remark 3 (Approximate implementation class)**.** The exact class of Lem. 6 is an idealization. An implementation satisfies instead an observational tolerance $\widehat{D}_{\mathrm{obs}} \leq \epsilon_{\mathrm{obs}}$ and an editing-response tolerance

$$\sup_{q \in \mathcal{D}_{\mathcal{Q},\mathrm{edit}}} d_{\mathcal{P}}\big(\mathsf{T}^I_{\hat{\mathcal{M}}(\boldsymbol{\theta}^\star_\epsilon)}(q), \mathsf{T}^I_{\mathcal{M}^\star}(q)\big) \leq \epsilon_{\mathrm{ua}},$$

for an assumed $\boldsymbol{\theta}^\star_\epsilon$, and extending the risk theorem to such models requires a risk-control condition on them and the propagation of both slacks.

D.8 Operational score and response-sample estimation

The finite-pool argument uses the weighted linear-feature population target

$$D_b^{\mathrm{lin}}(\mathsf{M}, \mathsf{R}_b) := \int \|\mathbb{E}\varphi_q(S_{\mathsf{M}}) - \mathbb{E}\varphi_q(S_{\mathsf{R}_b})\|_2^2 / d_{\varphi_q}\, d\mu_b(q),$$

where $\mathsf{R}_{\mathrm{stb}} = \mathsf{F}$, $\mathsf{R}_{\mathrm{corr}} = \mathsf{M}_0$, $S_{\mathsf{M}} \sim \mathsf{T}_{\mathsf{M}}(q)$, and μ_b is the query distribution induced by the bank-construction rule, including its normalized family weights. This quantity equals the main-text scores D_{stb} and D_{corr} in §3.2, and the superscript lin distinguishes it from the characteristic-kernel discrepancy of §D.4. The pool and its query construction are fixed independently of the held-out scores, response replicates are conditionally i.i.d. across sample indices, and queries from one case may be dependent.

For finite query sets Q_b with predeclared weights $w_q > 0$, the operational estimator averages the per-query discrepancies as

$$\widehat{D}_b^{e,\mathrm{avg}}(\mathcal{M}, \mathsf{R}_b) = \frac{\sum_{q \in Q_b} w_q\, \widehat{d}_q^{\,e}(\mathcal{M}, \mathsf{R}_b)}{\sum_{q \in Q_b} w_q}, \qquad b \in \{\mathrm{stb}, \mathrm{corr}\}, \quad e \in \{V, U\}, \tag{19}$$

where $b \in \{\mathrm{stb}, \mathrm{corr}\}$ identifies the reference and $e \in \{V, U\}$ identifies the response-sample estimator.

Remark 4 (V- and U-statistic estimates of the per-query term)**.** From a per-query response sample $\{s_r^{(i)}\}_{i=1}^{S_r}$, the implementation forms the sample mean $\hat{\mu}_{\mathsf{M}}(q_r) := \frac{1}{S_r}\sum_i \varphi(s_r^{(i)})$ and divides squared distances by d_φ, which rescales the identities of Lem. 4 by a constant. The plug-in $V_r := \|\hat{\mu}_{\mathsf{M}}(q_r) - \hat{\mu}_{\mathsf{M}_0}(q_r)\|_2^2 / d_\varphi$ is biased upward by the sampling variance, whereas the U-statistic

$$U_r := \frac{1}{d_\varphi S_r(S_r-1)} \sum_{i \neq j} \big\langle \varphi(s_r^{(i),\mathsf{M}}) - \varphi(s_r^{(i),\mathsf{M}_0}),\ \varphi(s_r^{(j),\mathsf{M}}) - \varphi(s_r^{(j),\mathsf{M}_0}) \big\rangle$$

satisfies $\mathbb{E}[U_r] = \|\mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - \mu_{\mathsf{T}_{\mathsf{M}_0}(q_r)}\|_2^2 / d_\varphi$. Training uses V_r and held-out selection uses U_r with $S_r \geq 2$. For stable queries the reference is the fixed vector $a_r := \varphi(s_r^{\mathrm{fact}})$ of the factual observation, and the one-sided estimator

$$U_r^{\mathrm{fix}} := \frac{1}{d_\varphi S_r(S_r-1)} \sum_{i \neq j} \big\langle \varphi(s_r^{(i),\mathsf{M}}) - a_r,\ \varphi(s_r^{(j),\mathsf{M}}) - a_r \big\rangle$$

is unbiased for $\|\mu_{\mathsf{T}_{\mathsf{M}}(q_r)} - a_r\|_2^2 / d_\varphi$ by the same argument. Which estimator applies is decided by the source of the reference.

D.9 FINITE-QUERY NEAR-OPTIMAL RECOVERY

The operational rule screens the saved models by factor fit, $C = \{\mathcal{M} \in \mathfrak{M}_{\text{cand}} : \widehat{D}_{\text{obs}}(\mathcal{M}) \le \epsilon_{\text{obs}}\}$ with ϵ_{obs} fixed before scoring (§E.2), and for a threshold τ and slack η_{stb} selects

$$\widehat{C}_{\tau,\eta} = \left\{\mathcal{M} \in C : \widehat{D}^{U,\text{avg}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le \tau + \eta_{\text{stb}}\right\}, \tag{20}$$

$$\widehat{\mathcal{M}} \in \arg\min_{\mathcal{M} \in \widehat{C}_{\tau,\eta}} \widehat{D}^{U,\text{avg}}_{\text{corr}}(\mathcal{M}, \mathsf{M}_0). \tag{21}$$

where $\widehat{C}_{\tau,\eta}$ makes the slack dependence of the main-text $\widehat{C}_\tau$ explicit and an empty retained set returns FAIL. For a fixed pool C and threshold τ, write $C_\tau = \{\mathcal{M} \in C : D^{\text{lin}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le \tau\}$ and let $D^\star_{\text{corr}}$ be the minimum population reference discrepancy over C_τ.

Theorem 6 (Finite-query near-optimal recovery, full statement of Thm. 1)**.** *Suppose $C_\tau \neq \varnothing$ and, with probability at least $1-\delta$, the preservation estimates have uniform error at most ε_{stb} on C, while the reference-discrepancy estimates have uniform error at most $\varepsilon_{\text{corr}}$ on C_τ. If*

$$g_{\text{stb}} := \inf_{\mathcal{M} \in C \setminus C_\tau} \left[D^{\text{lin}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) - \tau\right] > 2\varepsilon_{\text{stb}}, \qquad \inf \varnothing := +\infty,$$

then every $\eta_{\text{stb}} \in [\varepsilon_{\text{stb}}, g_{\text{stb}} - \varepsilon_{\text{stb}})$ satisfies, on the same event,

$$\widehat{C}_{\tau,\eta} = C_\tau, \qquad D^{\text{lin}}_{\text{corr}}(\widehat{\mathcal{M}}, \mathsf{M}_0) \le D^\star_{\text{corr}} + 2\varepsilon_{\text{corr}}.$$

A unique minimizer separated from all other candidates in C_τ by more than $2\varepsilon_{\text{corr}}$ is recovered exactly.

Proof of Theorem 6. Work on the event of probability at least $1-\delta$ on which both stated uniform approximation bounds hold.

Step 1: recovery of the candidate set satisfying preservation. We show $\widehat{C}_{\tau,\eta_{\text{stb}}} = C_\tau$.

If $\mathcal{M} \in C_\tau$, then $D^{\text{lin}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le \tau$. By the uniform preservation bound,

$$\widehat{D}^{U,\text{avg}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le D^{\text{lin}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) + \varepsilon_{\text{stb}} \le \tau + \eta_{\text{stb}},$$

where the last inequality uses $\eta_{\text{stb}} \ge \varepsilon_{\text{stb}}$. Thus $\mathcal{M} \in \widehat{C}_{\tau,\eta_{\text{stb}}}$.

Conversely, let $\mathcal{M} \in \widehat{C}_{\tau,\eta_{\text{stb}}}$. Then

$$\widehat{D}^{U,\text{avg}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le \tau + \eta_{\text{stb}}.$$

Applying the same uniform bound in the opposite direction gives

$$D^{\text{lin}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) \le \widehat{D}^{U,\text{avg}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) + \varepsilon_{\text{stb}} \le \tau + \eta_{\text{stb}} + \varepsilon_{\text{stb}}.$$

If $\mathcal{M} \notin C_\tau$, then $\mathcal{M} \in C \setminus C_\tau$, so by the preservation margin assumption,

$$D^{\text{lin}}_{\text{stb}}(\mathcal{M}, \mathsf{F}) - \tau \ge g_{\text{stb}} > \eta_{\text{stb}} + \varepsilon_{\text{stb}},$$

where the strict inequality follows from $\eta_{\text{stb}} < g_{\text{stb}} - \varepsilon_{\text{stb}}$. This contradicts the preceding inequality, so $\mathcal{M} \in C_\tau$.

Step 2: near-optimal reference agreement. Let $\hat{\mathcal{M}}^\star \in \arg\min_{\mathcal{M} \in C_\tau} D^{\text{lin}}_{\text{corr}}(\mathcal{M}, \mathsf{M}_0)$, so $D^{\text{lin}}_{\text{corr}}(\hat{\mathcal{M}}^\star, \mathsf{M}_0) = D^\star_{\text{corr}}$. By Step 1, $\widehat{\mathcal{M}}$ minimizes $\widehat{D}^{U,\text{avg}}_{\text{corr}}$ over C_τ. Therefore

$$\widehat{D}^{U,\text{avg}}_{\text{corr}}(\widehat{\mathcal{M}}, \mathsf{M}_0) \le \widehat{D}^{U,\text{avg}}_{\text{corr}}(\hat{\mathcal{M}}^\star, \mathsf{M}_0).$$

Applying the uniform reference-discrepancy bound on both sides,

$$\begin{aligned} D^{\text{lin}}_{\text{corr}}(\widehat{\mathcal{M}}, \mathsf{M}_0) &\le \widehat{D}^{U,\text{avg}}_{\text{corr}}(\widehat{\mathcal{M}}, \mathsf{M}_0) + \varepsilon_{\text{corr}} \\ &\le \widehat{D}^{U,\text{avg}}_{\text{corr}}(\hat{\mathcal{M}}^\star, \mathsf{M}_0) + \varepsilon_{\text{corr}} \\ &\le D^{\text{lin}}_{\text{corr}}(\hat{\mathcal{M}}^\star, \mathsf{M}_0) + 2\varepsilon_{\text{corr}} \\ &= D^\star_{\text{corr}} + 2\varepsilon_{\text{corr}}. \end{aligned}$$

Step 3: recovery of a separated minimizer. If the minimizer is unique and every other candidate in C_τ has reference score more than $2\varepsilon_{\mathrm{corr}}$ above its score, the bound in Step 2 excludes every other candidate. Hence the selected candidate is that minimizer, proving the final claim of the theorem. □

Corollary 2 (Uniform guarantee over stable discrepancy thresholds)**.** *Let the nonempty finite eligible pool C and its query-construction rule be fixed independently of the held-out score bank. Suppose that, with probability at least $1-\delta$, where $0<\delta<1$, both bounds*

$$\sup_{\mathcal{M}\in C}\left|\widehat{D}^{U,\mathrm{avg}}_{\mathrm{stb}}(\mathcal{M},\mathsf{F})-D^{\mathrm{lin}}_{\mathrm{stb}}(\mathcal{M},\mathsf{F})\right|\le\varepsilon_{\mathrm{stb}},$$
$$\sup_{\mathcal{M}\in C}\left|\widehat{D}^{U,\mathrm{avg}}_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0)-D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0)\right|\le\varepsilon_{\mathrm{corr}}$$

hold simultaneously for $\varepsilon_{\mathrm{stb}},\varepsilon_{\mathrm{corr}}\ge 0$. For each $\tau\ge 0$, let $C_\tau=\{\mathcal{M}\in C: D^{\mathrm{lin}}_{\mathrm{stb}}(\mathcal{M},\mathsf{F})\le\tau\}$ and define

$$g_{\mathrm{stb}}(\tau):=\inf_{\mathcal{M}\in C\setminus C_\tau}\left[D^{\mathrm{lin}}_{\mathrm{stb}}(\mathcal{M},\mathsf{F})-\tau\right],\qquad \inf\varnothing:=+\infty.$$

With probability at least $1-\delta$, the following implication holds simultaneously for every threshold τ and slack η_{stb}. Whenever

$$C_\tau\neq\varnothing,\qquad g_{\mathrm{stb}}(\tau)>2\varepsilon_{\mathrm{stb}},\qquad \varepsilon_{\mathrm{stb}}\le\eta_{\mathrm{stb}}<g_{\mathrm{stb}}(\tau)-\varepsilon_{\mathrm{stb}},$$

the retained set in Eq. (20) satisfies $\widehat{C}_{\tau,\eta}=C_\tau$, and every empirical minimizer $\widehat{\mathcal{M}}_{\tau,\eta}$ in Eq. (21) obeys

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\widehat{\mathcal{M}}_{\tau,\eta},\mathsf{M}_0)\le\min_{\mathcal{M}\in C_\tau}D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0)+2\varepsilon_{\mathrm{corr}}.$$

In particular, the implication permits τ and η_{stb} to be selected from the same held-out scores.

Proof. Work on the event on which the two uniform bounds hold. For any pair $(\tau,\eta_{\mathrm{stb}})$ satisfying the displayed conditions, every $\mathcal{M}\in C_\tau$ has estimated preservation score at most $\tau+\varepsilon_{\mathrm{stb}}\le\tau+\eta_{\mathrm{stb}}$. Every $\mathcal{M}\in C\setminus C_\tau$ instead has estimated preservation score at least $\tau+g_{\mathrm{stb}}(\tau)-\varepsilon_{\mathrm{stb}}>\tau+\eta_{\mathrm{stb}}$. Thus $\widehat{C}_{\tau,\eta}=C_\tau$. Let $\mathcal{M}^\star_\tau$ minimize the population reference discrepancy on this nonempty finite set. Empirical optimality and the uniform reference bound give

$$\begin{aligned}D^{\mathrm{lin}}_{\mathrm{corr}}(\widehat{\mathcal{M}}_{\tau,\eta},\mathsf{M}_0)&\le\widehat{D}^{U,\mathrm{avg}}_{\mathrm{corr}}(\widehat{\mathcal{M}}_{\tau,\eta},\mathsf{M}_0)+\varepsilon_{\mathrm{corr}}\\&\le\widehat{D}^{U,\mathrm{avg}}_{\mathrm{corr}}(\mathcal{M}^\star_\tau,\mathsf{M}_0)+\varepsilon_{\mathrm{corr}}\\&\le D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}^\star_\tau,\mathsf{M}_0)+2\varepsilon_{\mathrm{corr}}.\end{aligned}$$

The uniform event controls both scores for every candidate before a threshold is chosen, so these deterministic inequalities hold for all admissible threshold/slack pairs on that same event. Its probability is at least $1-\delta$. □

Cor. 3 supplies the uniform event by bounding both scores over C. Applying the corollary to the empirical-quantile rule requires a population stable margin for which the selected slack lies in the displayed interval, which the reported pools do not verify.

D.10 Sample complexity

Corollary 3 (Sample complexity)**.** *Under the assumptions of Thm. 6, let the held-out bank consist of n_c cases drawn independently from the validation fold, let $\bar{d}_\bullet(c)$ denote the per-case average of the per-query U-statistic estimates of the linear-feature discrepancy over the queries built from case c, with bounded per-case range $2B_\bullet$, $\bullet\in\{\mathrm{stb},\mathrm{corr}\}$, and let $\widehat{D}^{U,\mathrm{avg}}_\bullet$ be the uniform average of $\bar{d}_\bullet(c)$ over cases. Queries within a case share its factual observation and may be dependent, and the cases are independent. The stable discrepancy threshold τ is fixed before these scores are evaluated. Assume $0<g_{\mathrm{stb}}<\infty$. For any $\epsilon'\in(0,g_{\mathrm{stb}}/2)$ and $\Delta>0$, use an estimation slack*

$$\eta_{\mathrm{stb}}\in[g_{\mathrm{stb}}/2-\epsilon',\ g_{\mathrm{stb}}/2+\epsilon'),$$

corresponding to $\varepsilon_{\mathrm{stb}}=g_{\mathrm{stb}}/2-\epsilon'$ in Thm. 6. If

$$n_c>\frac{2B^2_{\mathrm{stb}}}{(g_{\mathrm{stb}}/2-\epsilon')^2}\log\frac{4|C|}{\delta},\qquad n_c\ge\frac{8B^2_{\mathrm{corr}}}{\Delta^2}\log\frac{4|C|}{\delta},$$

then with probability at least $1-\delta$, $D^{\mathrm{lin}}_{\mathrm{corr}}(\widehat{\mathcal{M}}, \mathsf{M}_0) \le D^{\star}_{\mathrm{corr}} + \Delta$. *Define the* finite selection margin *over the eligible candidate set as* $g_{\mathrm{sel}}(C_\tau) := \min_{\mathcal{M}\in C_\tau,\, \mathcal{M}\neq\mathcal{M}^\star_C}[D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}, \mathsf{M}_0) - D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}^\star_C, \mathsf{M}_0)]$, *where* $\mathcal{M}^\star_C$ *is the unique* $D^{\mathrm{lin}}_{\mathrm{corr}}$*-minimizer in* C_τ *(assumed to exist). If* $g_{\mathrm{sel}} > 0$ *and* $\Delta < g_{\mathrm{sel}}$, *then* $\widehat{\mathcal{M}} = \mathcal{M}^\star_C$ *exactly. This margin concerns* C, *not the full population class unless* C *already contains its unique minimizer.*

When $\varnothing \neq C_\tau = C$, Theorem 6 applies directly under its uniform score-error bounds with finite $\varepsilon_{\mathrm{stb}}$ and any finite $\eta_{\mathrm{stb}} \ge \varepsilon_{\mathrm{stb}}$.

Proof of Corollary 3. We verify the uniform approximation bounds of Theorem 6 with Hoeffding's inequality and a union bound over the fixed finite pool C, applied to the per-case averages $\bar{d}_\bullet(c)$, which are independent across the n_c validation cases and bounded. The resulting event also supports Cor. 2.

Boundedness of the per-query summands. The held-out per-query estimator is the unbiased U-statistic of Rem. 4. Let $\mathsf{R}_{\mathrm{stb}} = \mathsf{F}$ and $\mathsf{R}_{\mathrm{corr}} = \mathsf{M}_0$. For a query in block $\bullet$ with response dimension d_φ, $U_r = \frac{1}{S_r(S_r-1)d_\varphi}\sum_{i\neq j}\langle \varphi(s_r^{(i),\mathsf{M}}) - \varphi(s_r^{(i),\mathsf{R}_\bullet}),\, \varphi(s_r^{(j),\mathsf{M}}) - \varphi(s_r^{(j),\mathsf{R}_\bullet})\rangle$. Under $\sup_{s\in\mathcal{S}}\|\varphi(s)\| \le R_\bullet$, each feature difference has norm at most $2R_\bullet$, so Cauchy–Schwarz gives $U_r \in [-B_\bullet, B_\bullet]$ with $B_\bullet = 4R_\bullet^2$. Queries from the same case are averaged into one bounded summand $\bar{d}_\bullet(c)$, and independence across cases is an assumption that disjoint folds alone do not establish (§E.2).

Preservation bound. Fix $\mathcal{M} \in C$. By the construction above, the per-case summands $\bar{d}_{\mathrm{stb}}(c)$ defining $\widehat{D}^{U,\mathrm{avg}}_{\mathrm{stb}}(\mathcal{M}, \mathsf{F})$ are independent across cases, unbiased, and bounded in $[-B_{\mathrm{stb}}, B_{\mathrm{stb}}]$. In the uniform-case-weight setting, Hoeffding's inequality yields

$$P\left(\left|\widehat{D}^{U,\mathrm{avg}}_{\mathrm{stb}}(\mathcal{M}, \mathsf{F}) - D^{\mathrm{lin}}_{\mathrm{stb}}(\mathcal{M}, \mathsf{F})\right| > \varepsilon_{\mathrm{stb}}\right) \le 2\exp\left(-\frac{n_c\varepsilon^2_{\mathrm{stb}}}{2B^2_{\mathrm{stb}}}\right).$$

Applying a union bound over the $|C|$ candidates gives

$$P\left(\sup_{\mathcal{M}\in C}\left|\widehat{D}^{U,\mathrm{avg}}_{\mathrm{stb}}(\mathcal{M}, \mathsf{F}) - D^{\mathrm{lin}}_{\mathrm{stb}}(\mathcal{M}, \mathsf{F})\right| > \varepsilon_{\mathrm{stb}}\right) \le 2|C|\exp\left(-\frac{n_c\varepsilon^2_{\mathrm{stb}}}{2B^2_{\mathrm{stb}}}\right).$$

For any $\epsilon' \in (0, g_{\mathrm{stb}}/2)$, setting $\varepsilon_{\mathrm{stb}} = g_{\mathrm{stb}}/2 - \epsilon'$ preserves Thm. 6's strict margin $g_{\mathrm{stb}} > 2\varepsilon_{\mathrm{stb}}$ and, requiring the right-hand side to be at most $\delta/2$, yields

$$n_c > \frac{2B^2_{\mathrm{stb}}}{(g_{\mathrm{stb}}/2 - \epsilon')^2}\log\frac{4|C|}{\delta} \xrightarrow[\epsilon'\downarrow 0]{} \frac{8B^2_{\mathrm{stb}}}{g^2_{\mathrm{stb}}}\log\frac{4|C|}{\delta}. \tag{22}$$

Every $\epsilon' > 0$ gives an attainable condition, and the limit $8B^2_{\mathrm{stb}}/g^2_{\mathrm{stb}}$ is not attained because the theorem requires $g_{\mathrm{stb}} > 2\varepsilon_{\mathrm{stb}}$.

Reference-discrepancy bound. Applying the same argument to $\widehat{D}^{U,\mathrm{avg}}_{\mathrm{corr}}$, with terms bounded in $[-B_{\mathrm{corr}}, B_{\mathrm{corr}}]$, gives

$$P\left(\sup_{\mathcal{M}\in C}\left|\widehat{D}^{U,\mathrm{avg}}_{\mathrm{corr}}(\mathcal{M}, \mathsf{M}_0) - D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}, \mathsf{M}_0)\right| > \varepsilon_{\mathrm{corr}}\right) \le 2|C|\exp\left(-\frac{n_c\varepsilon^2_{\mathrm{corr}}}{2B^2_{\mathrm{corr}}}\right).$$

Setting $\varepsilon_{\mathrm{corr}} = \Delta/2$ and requiring this to be at most $\delta/2$ yields the sufficient condition

$$n_c \ge \frac{8B^2_{\mathrm{corr}}}{\Delta^2}\log\frac{4|C|}{\delta}.$$

With probability at least $1-\delta$, both uniform approximation bounds hold simultaneously. Thm. 6 then gives

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\widehat{\mathcal{M}}, \mathsf{M}_0) \le D^{\star}_{\mathrm{corr}} + \Delta.$$

If $g_{\mathrm{sel}} > 0$ and $\Delta < g_{\mathrm{sel}}$, no other candidate in C_τ lies within Δ of $D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}^\star_C, \mathsf{M}_0)$, so $\widehat{\mathcal{M}} = \mathcal{M}^\star_C$. □

D.11 CONDITIONAL L_3 RISK GUARANTEE

The main-text theorem relates reference agreement to the counterfactual image risk in Def. 4 through an explicit risk-control condition on the exact feasible class.

The statement appears in Thm. 2 in the main text.

Proof of Theorem 2. Let $\mathcal{M}^\dagger \in \arg\min_{\mathcal{M}\in\mathfrak{M}_\tau} D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}, \mathsf{M}_0)$ be any population selector, whose existence follows from the attainment assumption.

By the assumption that a correct model satisfies preservation in Thm. 2, there exists $\bar{\mathcal{M}} \in \mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau$. Since $\mathcal{M}^\dagger$ minimizes $D^{\mathrm{lin}}_{\mathrm{corr}}$ over $\mathfrak{M}_\tau$,

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}^\dagger, \mathsf{M}_0) \le D^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}}, \mathsf{M}_0).$$

Taking the infimum over $\bar{\mathcal{M}} \in \mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau$ yields

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}^\dagger, \mathsf{M}_0) - \inf_{\bar{\mathcal{M}}\in\mathfrak{M}_{\mathrm{cf}}\cap\mathfrak{M}_\tau} D^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}}, \mathsf{M}_0) \le 0.$$

The reference-informativeness condition of Thm. 2, applied to $\mathcal{M}^\dagger \in \mathfrak{M}_\tau$, gives

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}^\dagger, \mathsf{M}_0) - \inf_{\bar{\mathcal{M}}\in\mathfrak{M}_{\mathrm{cf}}\cap\mathfrak{M}_\tau} D^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}}, \mathsf{M}_0) \ge \alpha\, \mathcal{R}_{L_3}(\mathcal{M}^\dagger) - \beta.$$

Combining the two inequalities yields

$$0 \ge \alpha\, \mathcal{R}_{L_3}(\mathcal{M}^\dagger) - \beta,$$

hence

$$\mathcal{R}_{L_3}(\mathcal{M}^\dagger) \le \frac{\beta}{\alpha}.$$

If $\beta = 0$, then $\mathcal{R}_{L_3}(\mathcal{M}^\dagger) = 0$. Since $\mathcal{R}_{L_3}$ is the integral of the nonnegative function $q \mapsto d_{\mathcal{P}}(\mathsf{T}^I_{\mathcal{M}^\dagger}(q), \mathsf{T}^I_{\mathcal{M}^\star}(q))$, Lem. 2 implies

$$\mathsf{T}^I_{\mathcal{M}^\dagger}(q) = \mathsf{T}^I_{\mathcal{M}^\star}(q) \qquad \nu_{\mathrm{edit}}\text{-a.e. } q \in \mathcal{D}_{\mathcal{Q},\mathrm{edit}}.$$

□

Corollary 4 (Finite-query counterfactual-risk bound)**.** *Suppose the conditions of Theorems 6 and 2 hold on the uniform-estimation event and, additionally, $C \subseteq \mathfrak{M}_{\mathrm{feas}}$. The last condition is idealized, since the implementation-level eligibility rule does not establish it. Define the candidate-set reference offset*

$$\zeta_{\mathrm{cand}} := D^\star_{\mathrm{corr}} - \inf_{\bar{\mathcal{M}}\in\mathfrak{M}_{\mathrm{cf}}\cap\mathfrak{M}_\tau} D^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}}, \mathsf{M}_0).$$

Then the empirical selector satisfies

$$\mathcal{R}_{L_3}(\widehat{\mathcal{M}}) \le \frac{\beta + \zeta_{\mathrm{cand}} + 2\varepsilon_{\mathrm{corr}}}{\alpha}.$$

The three numerator terms are the risk-control slack, candidate-set reference offset, and finite-query estimation error. The offset need not be non-negative because the finite candidate set may contain proxies satisfying preservation outside the counterfactually correct subclass.

Proof. Reference informativeness gives $\alpha\mathcal{R}_{L_3}(\widehat{\mathcal{M}}) \le \beta + D^{\mathrm{lin}}_{\mathrm{corr}}(\widehat{\mathcal{M}}, \mathsf{M}_0) - \inf_{\bar{\mathcal{M}}\in\mathfrak{M}_{\mathrm{cf}}\cap\mathfrak{M}_\tau} D^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}}, \mathsf{M}_0)$. Theorem 6 bounds the empirical selector's reference discrepancy by $D^\star_{\mathrm{corr}} + 2\varepsilon_{\mathrm{corr}}$. Substitution and division by $\alpha > 0$ yield the result. □

D.12 CONDITIONAL RISK CONTROL FROM REFERENCE PROXIMITY

Reference proximity can control risk only if differences in the selected mean responses already control raw-image edit risk on the surviving class. The proposition assumes that control and computes the constants it implies, without deriving it from pretraining.

Proposition 4 (Conditional Hilbert-space risk bound)**.** *For $S_{\mathsf{M}} \sim \mathsf{T}_{\mathsf{M}}(q)$, define the linear-mean embedding in the Hilbert direct integral*

$$\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathsf{M})(q) := d_{\varphi_q}^{-1/2}\mathbb{E}\varphi_q(S_{\mathsf{M}}), \qquad \mathcal{H}_{\mathrm{corr}} := \int^{\oplus} \mathbb{R}^{d_{\varphi_q}}\, d\mu_{\mathrm{corr}}(q).$$

Then $D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathcal{M}') = \|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}')\|^2_{\mathcal{H}_{\mathrm{corr}}}$. Let $\epsilon_Z, \epsilon_U \geq 0$ be declared factor- and renderer-side recovery slacks. Suppose there exist $\bar{\mathcal{M}} \in \mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau$, $C_{\mathrm{edit}} > 0$, and $C_{\mathrm{sb}}, \epsilon_0 \geq 0$ such that

$$\mathcal{R}_{L_3}(\mathcal{M}) \leq C_{\mathrm{edit}}\|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}})\|^2_{\mathcal{H}_{\mathrm{corr}}} + C_{\mathrm{sb}}(\epsilon_Z^2 + \epsilon_U^2) \quad \forall \mathcal{M} \in \mathfrak{M}_\tau,$$

and $\|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathsf{M}_0) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}})\|_{\mathcal{H}_{\mathrm{corr}}} \leq \epsilon_0$. Then the reference-informativeness condition of Thm. 2 holds with

$$\alpha = \tfrac{1}{2C_{\mathrm{edit}}}, \qquad \beta = 2\epsilon_0^2 + \tfrac{C_{\mathrm{sb}}}{2C_{\mathrm{edit}}}(\epsilon_Z^2 + \epsilon_U^2).$$

The edit-risk domination inequality is a premise, and the proposition propagates it and the reference-proximity bound into the constants of the risk-control condition without establishing that either term is small in the reported pipeline.

Proof of Proposition 4. Fix $\mathcal{M} \in \mathfrak{M}_\tau$. Let $\bar{\mathcal{M}} \in \mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau$ be the model appearing in the hypothesis, and let (ϵ_Z, ϵ_U) denote its factor- and image-side recovery slacks.

By the triangle inequality in $\mathcal{H}_{\mathrm{corr}}$,

$$\|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}})\|_{\mathcal{H}_{\mathrm{corr}}} \leq \|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathsf{M}_0)\|_{\mathcal{H}_{\mathrm{corr}}} + \|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathsf{M}_0) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}})\|_{\mathcal{H}_{\mathrm{corr}}}.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$, together with the hypotheses

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathcal{M}') = \|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}')\|^2_{\mathcal{H}_{\mathrm{corr}}}$$

and

$$\|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathsf{M}_0) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}})\|_{\mathcal{H}_{\mathrm{corr}}} \leq \epsilon_0,$$

we obtain

$$\|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}})\|^2_{\mathcal{H}_{\mathrm{corr}}} \leq 2D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0) + 2\epsilon_0^2.$$

By the extended edit-risk control hypothesis,

$$\begin{aligned}\mathcal{R}_{L_3}(\mathcal{M}) &\leq C_{\mathrm{edit}}\|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}})\|^2_{\mathcal{H}_{\mathrm{corr}}} + C_{\mathrm{sb}}(\epsilon_Z^2 + \epsilon_U^2) \\ &\leq 2C_{\mathrm{edit}}D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0) + 2C_{\mathrm{edit}}\epsilon_0^2 + C_{\mathrm{sb}}(\epsilon_Z^2 + \epsilon_U^2).\end{aligned}$$

Rearranging gives

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0) \geq \frac{1}{2C_{\mathrm{edit}}}\mathcal{R}_{L_3}(\mathcal{M}) - \epsilon_0^2 - \frac{C_{\mathrm{sb}}}{2C_{\mathrm{edit}}}(\epsilon_Z^2 + \epsilon_U^2).$$

Next, since $\bar{\mathcal{M}} \in \mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau$,

$$\inf_{\widetilde{\mathcal{M}} \in \mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau} D^{\mathrm{lin}}_{\mathrm{corr}}(\widetilde{\mathcal{M}},\mathsf{M}_0) \leq D^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}},\mathsf{M}_0) = \|\Psi^{\mathrm{lin}}_{\mathrm{corr}}(\bar{\mathcal{M}}) - \Psi^{\mathrm{lin}}_{\mathrm{corr}}(\mathsf{M}_0)\|^2_{\mathcal{H}_{\mathrm{corr}}} \leq \epsilon_0^2.$$

Subtracting this upper bound on the infimum from the previous lower bound yields

$$D^{\mathrm{lin}}_{\mathrm{corr}}(\mathcal{M},\mathsf{M}_0) - \inf_{\widetilde{\mathcal{M}} \in \mathfrak{M}_{\mathrm{cf}} \cap \mathfrak{M}_\tau} D^{\mathrm{lin}}_{\mathrm{corr}}(\widetilde{\mathcal{M}},\mathsf{M}_0) \geq \frac{1}{2C_{\mathrm{edit}}}\mathcal{R}_{L_3}(\mathcal{M}) - 2\epsilon_0^2 - \frac{C_{\mathrm{sb}}}{2C_{\mathrm{edit}}}(\epsilon_Z^2 + \epsilon_U^2).$$

This is exactly the reference-informativeness condition of Thm. 2 with $\alpha = \frac{1}{2C_{\mathrm{edit}}}$ and $\beta = 2\epsilon_0^2 + \frac{C_{\mathrm{sb}}}{2C_{\mathrm{edit}}}(\epsilon_Z^2 + \epsilon_U^2)$. Setting $C_{\mathrm{sb}} = 0$ recovers the original reference-only bound $\beta = 2\epsilon_0^2$. □

E IMPLEMENTATION AND EVALUATION DETAILS

This appendix records how the method of §3 is instantiated and how the results of §5 are measured. §E.1 describes the proxy, the scoring renderer, and the rendering interfaces of the three anchors. §E.2 gives the training objective, the saved candidate pools, the observational-fit screen, and the data splits, and §E.3 the rule that chooses the stable discrepancy threshold. §E.4 specifies the query banks and the estimators of the two discrepancies. §E.5 defines the metrics and the evaluation protocol, §E.6 the benchmarks and the baseline adapters, and §§E.7 and E.8 the cost of selection and the compute used. Attribute measurements use a fixed image predictor on faces and an exact factor parser on ColoredBar-MNIST, and final image evaluation uses a separate attribute predictor.

E.1 PROXY, RENDERER, AND INFERENCE

Inference pipeline. For one factual image and intervention, the proxy infers a factor-level background state, sets the target, and propagates its factor mechanisms. Its responses are measured at two points, in training and selection and at test time. During training and selection, a learned *scoring renderer* maps the proxy's full factor rollout to an image for the fixed face attribute predictor, whereas CBMNIST queries read the factor fields through the benchmark's exact factor interface. At test time, the pipeline holds the non-descendant factors at their factual values and sends the resulting tuple and the source image to the anchor's rendering interface. The scoring renderer is part of the candidate model, and §5.1 examines whether rankings computed through it transfer to the final edits rendered by the anchor.

Proxy SCM. The implementation uses a neural-SCM interface over the observed factors and delegates image formation to the benchmark renderer. Its directed factor mechanisms follow the supplied graph, with one Gaussian exogenous block per c-component, MLP mechanisms, and Bernoulli, categorical, or Gaussian heads according to variable type. This district-level noise sharing is exactly graph-constrained for singleton or clique districts. A general non-clique district can introduce extra connections and lies outside the finite-domain G-NCM expressiveness result. The final benchmark lineages use singleton c-components. Discrete mechanisms use teacher-forced cross-entropy on the observational path, straight-through Gumbel samples (Jang et al., 2017) during alignment, and hard samples at inference.

Abduction. Abduction is handled separately on the factor side and the image side, and neither side computes the exact posterior $P^{\mathcal{M}}(\boldsymbol{u} \mid \boldsymbol{z})$ of Eq. (9). When every factor value of a case is observed and the deterministic-encoding entry of Tab. 3 applies (CBMNIST), each factor-level exogenous code is a deterministic function of the observed factor tuple in the proxy's parameterization, so observing the complete tuple fixes the abducted factor state without a posterior computation. Otherwise (SDXL and StyleGAN), an amortized network conditioned on the observed factors supplies an approximate posterior over the factor-level exogenous state. The image-side noise U_I is not abduced explicitly, as the anchor's image-conditioned interface takes the factual image as input and no posterior over U_I is computed. The anchor serves as a rendering reference and carries no SCM or counterfactual semantics.

Scoring renderer and response measurement. The learned scoring renderer is an MLP followed by a convolutional decoder. Alignment losses update this module and, where gradients propagate through its inputs, the factor mechanisms, which keeps image-based alignment differentiable without differentiating through the anchor. Face parser queries read its output with a fixed attribute predictor. CBMNIST parser queries instead use the benchmark's exact factor interface and read the factor fields supplied by the proxy or the anchor, even when the response also carries an image. This structured-factor route is therefore not a measurement inferred from final pixels, whereas image-feature queries, when active, use the scoring renderer's output.

Final rendering interfaces. The final image path uses the image-conditioned interface of the benchmark anchor. SDXL img2img conditions generation on a noised source image and the requested factor condition. StyleGAN3-T first finds a latent code in W^+ that reconstructs the source, then changes it along attribute directions. ColoredBar-DDPM renders the final images by DDIM sampling from the noised source image with strength 0.85 and 50 steps. During candidate generation it instead

used paired-noise DDIM sampling with 25 steps, whose images do not enter the CBMNIST scores because those queries read the exact factor interface.

Inputs and references. Factor annotations specify the observed states used to fit the proxy and construct interventions. Candidate generation and held-out model selection then compute stable and correction responses through the same frozen parser and feature interfaces, using disjoint cases. The correction reference in both stages is the anchor edit, and stable queries use the factual reference F. The pipeline thus uses a supplied graph, factor-labeled observational data, and fixed measurement functions, and neither candidate generation nor selection is trained against ground-truth counterfactual images.

Three distinct objects. This separation distinguishes the desired image counterfactual, the proxy's inferred factor state, and the renderer's output. Correct factors need not produce a correct image when the renderer is biased. Likewise, holding non-descendant factors at their factual values guarantees zero drift only at the factor level, not in image-based measurements. The selection rule, consisting of the finite pool, the stable constraint, the correction ranking, and the held-out protocol, is shared across benchmarks. Each domain supplies its causal graph, intervention semantics, stable and correction query roles, and measurements, and each anchor backend supplies its generation and editing interface, its inversion or conditioning route, its prompt adapter, and its preprocessing. "Architecture independent" thus describes the rule's query-response interface rather than a plug-and-play implementation with no backend adapter.

Tab. 2 lists each theoretical object of §3 with its operational counterpart and the reading that the correspondence supports.

E.2 Candidate generation and held-out selection

Training objective. We train graph-constrained neural proxies on factor-labeled observational data and save models as a finite pool $\mathfrak{M}_{\mathrm{cand}}$, from which the observational-fit screen of Algorithm 1 retains the eligible pool C. The anchor and the measurements remain fixed, and training query banks are resampled. Candidate generation uses

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) =& \lambda_{\mathrm{data}}\widehat{\mathcal{L}}_{\mathrm{obs}}^{\mathrm{train}}(\boldsymbol{\theta}) + \lambda_{\mathrm{stb}}\widehat{D}_{\mathrm{stb}}^{V,\mathrm{avg}}(\hat{\mathcal{M}}(\boldsymbol{\theta}), \mathsf{F}) \\ &+ \lambda_{\mathrm{corr}}\widehat{D}_{\mathrm{corr}}^{\mathrm{train}}(\hat{\mathcal{M}}(\boldsymbol{\theta}), \mathsf{M}_0) + \gamma R(\boldsymbol{\theta}).\end{aligned} \tag{23}$$

The stable training score $\widehat{D}_{\mathrm{stb}}^{\mathrm{train}}$ in Eq. (7) is the V-statistic $\widehat{D}_{\mathrm{stb}}^{V,\mathrm{avg}}$ above. The λ coefficients weight observational fit and the two alignment terms, while γ weights the regularizer R. In the reported pools, R is the sum of squared non-renderer proxy parameters and learned abduction parameters, with $\gamma = 10^{-4}$. Deterministic abduction contributes no learned-parameter term, and renderer regularization has zero weight. The observational loss combines teacher-forced and rollout factor predictions. Def. 9 and §E.4 specify the graph constraint, losses, and query counts. This soft objective supplies candidates for the held-out selection rule.

Training loop. Training samples a small pool of observational cases, instantiates a small alignment bank with stable and correction blocks, evaluates the loss in Eq. (23), and saves models at fixed intervals. Stable responses compare the proxy response with the same measurement of the factual case. Correction responses compare proxy and anchor responses. The anchor and all measurement functions remain frozen. Teacher forcing predicts each factor from its observed parents, and rollout prediction uses parents generated by preceding mechanisms. Both losses compare factor predictions with observational labels. The observational discrepancy is

$$\widehat{\mathcal{L}}_{\mathrm{obs}}^{\mathrm{train}} = \rho_{\mathrm{tf}}\widehat{D}_{\mathrm{tf}} + \rho_{\mathrm{roll}}\widehat{D}_{\mathrm{roll}}.$$

The reported pools use these two terms. Optional noise-utilization and abduction-reconstruction weights are zero. The candidate-generation correction term is

$$\widehat{D}_{\mathrm{corr}}^{\mathrm{train}} = \widehat{D}_{\mathrm{corr}}^{V},$$

where the superscript V denotes the response-mean V-statistic used during training. The observational and alignment losses affect which models are produced but do not enter the held-out selector, which recomputes its scores with held-out U-statistics.

Table 2: Theoretical objects and implementation counterparts. Correspondence holds only under the listed assumptions.

Object or result	Mathematical scope	Deployed counterpart	Supported interpretation
$\mathsf{T}^{I}_{\mathcal{M}}(q)$ / $\mathsf{T}_{\mathcal{M}}(q)$	Raw edited-image distribution / its pushforward through a fixed measurement map	Face scoring: learned query-renderer images; CBMNIST parser scoring: factor fields; final edits: pretrained renderer	Exact L_3 risk concerns raw image distributions; deployed score recovery concerns the specified scoring responses, not final-image distributions
D^{char}_{μ}	Joint query–response distribution with a characteristic joint kernel	No training or selection score; Gaussian-RBF MMD is diagnostic only	Response-distribution geometry, not a guarantee for the operational selector
D^{lin}_{b}	Weighted average of squared differences between fixed feature means	Mean-MSE V-statistic in training and per-query U-statistic in selection	Finite-pool score recovery under the conditions of Thm. 6
$\mathfrak{M}_{\mathrm{feas}}$	Exact graph and observational-distribution equality	No finite operational test certifies membership	Idealized domain of the counterfactual-risk result
C (eligible candidates)	Predeclared held-out observational-fit statistic and tolerance	Required held-out model filter before reselection; the recorded field `theory_eligible` duplicates `observational_fit_pass`	Eligibility for the finite-pool selector only; never evidence of exact-class membership or certification of the counterfactual-risk result
G-NCM expressiveness	Finite endogenous domains and maximal-bidirected-clique Uniform noise (Xia et al., 2023)	District-level Gaussian blocks; current benchmark pools have singleton districts	The citation does not certify the operational continuous architecture
Finite-pool selector	Fixed-threshold theorem; simultaneous threshold extension under uniform errors and margin/slack conditions	Predeclared quantile-threshold rule over candidate preservation scores, with the threshold selected using reference-disagreement scores and never final test outcomes (§E.3); empty candidate sets recorded as FAIL	Cor. 2 covers score-dependent thresholds satisfying its conditions; operational bootstrap slack and population margins remain uncertified

Observational-fit screen. Before reselection, a separate observational-role bank filters the saved models. It contains disjoint, factor-covered evaluation and null-reference subsets. The model loss is averaged over the evaluation subset. The null-reference subset alone estimates the probabilities of a parent-agnostic marginal predictor. Their ratio is $\widehat{D}_{\mathrm{obs}} = \widehat{L}^{\mathrm{tf}}_{\mathrm{factor}}/\widehat{L}_{\mathrm{marginal}}$, where values below 1 improve on the reference. Both subsets, their label-coverage requirements, and $\epsilon_{\mathrm{obs}} = 0.95$ are fixed before scoring. The rule establishes factor-level empirical eligibility on held-out cases. It neither tests the full image distribution nor proves membership in $\mathfrak{M}_{\mathrm{feas}}$. The training observational loss stays outside held-out ranking, while this independent filter restricts its candidate pool.

Data splits. Candidate generation and model selection use different cases. An earlier benchmark test fold supported strength and sample-count diagnostics, so it no longer qualifies as an untouched final set. The SDXL rendering strength of 0.70 was chosen on that fold by per-image review of matched edits at three strengths over three training seeds, before the reported runs. The frozen final bank instead contains 50 sources and two interventions per source for each dataset. Before final scoring, we verified zero source-identity overlap with candidate generation, selection, observational-fit assessment, parser training and validation, threshold selection, and tuning. Set disjointness alone cannot establish independence among cases within the selection fold. The concentration result therefore assumes independent observational cases in the held-out bank. Every saved model is rescored on the validation-selection fold, and the stable discrepancy threshold is set by the quantile

rule of §E.3, which never consults final outcomes. An empty eligible or feasible set is recorded as FAIL, and none occurred in the reported pools.

The candidate model pools use two-layer MLP factor mechanisms with hidden width 64. Candidate generation uses Adam for 100 steps on CBMNIST and 200 steps on the face benchmarks, with learning rates 10^{-3} for factor mechanisms and 10^{-4} for the learned query renderer, optional image adapters, and amortized abduction networks when present. The teacher-forced and rollout terms use weights $(\rho_{\mathrm{tf}}, \rho_{\mathrm{roll}}) = (0.25, 1)$; noise-utilization, abduction-reconstruction, and structural-penalty weights are zero. These settings describe candidate generation only. The final stable discrepancy threshold, observational eligibility rule, source bank, evaluation predictor identities, and evaluation settings are fixed separately in the approved pre-result protocol and execution manifests.

Table 3: Candidate-generation settings of the model pools used in evaluation.

Setting	CBMNIST	SDXL	StyleGAN
Loss weights, observational	1	1	1
preservation	1	5	5
reference disagreement	1	5	5
Alignment context cases	4	2	16
Stable queries per step	8	8	8
Correction queries per step	4	4	4
Training responses per query	4	2	4
Factor-state inference	deterministic	amortized	amortized
Candidate-scoring interval	20	3	20
Saved model count	5	67	10
Rendering interface	DDIM twin-noise	SDXL img2img	matched encoder $\to W^+$
Rendering strength	0.85	0.70	–
Guidance scale	5	7.5	–
Rendering steps	25	50	StyleGAN3-T at 256^2

E.3 Choosing the stable discrepancy threshold

Reported rule. Every reported selection uses a predeclared rule that adapts the stable discrepancy threshold to the candidate pool. Eligible candidates respect the graph architecture and pass the observational-fit screen. Under exact SCM semantics, non-descendants retain their values when the same background state is held fixed. In the implementation, however, the inferred state may not reproduce the factual factors when their mechanisms are rerun. Selection scores use these full rollouts, so stable scores can reflect both factor reconstruction error and changes introduced by rendering. The final editor separately holds non-descendant factors at their factual values.

The threshold τ ranges over the $\{10, 25, 50, 75, 90\}\%$ empirical quantiles of the candidates' stable scores, each widened by bootstrap slack η_{stb}. For each threshold, we find the retained candidate with the smallest correction score. Within a threshold, ties favor lower stable scores and then earlier optimization iterates. Across thresholds, we compare correction scores, stable scores, thresholds, and optimization iterates in that order, preferring smaller values. Since these sets are nested, loosening the constraint is preferred whenever it improves the minimum correction score. When stable scores vary by about as much as the estimation slack, the constraint separates few candidates and selection mainly follows the correction score. A wider spread can make the constraint more restrictive.

The shared slack η_{stb} is the standard deviation of 100 bootstrap weighted means. Each resample draws uniformly, with replacement, from the pooled candidate–stable-query score and weight pairs, retaining the original number of entries. The standard deviation uses Bessel's correction. The same slack is used for all candidates. The fixed-threshold and simultaneous-threshold results below state the error and margin conditions under which this choice yields score recovery.

Recorded pools. The frozen pools contain 5 (CBMNIST), 10 (StyleGAN) and 67 (SDXL) eligible candidates per run. Their stable scores span 0.088–0.167 (CBMNIST), 0.197–0.220 (StyleGAN) and 0.195–0.221 (SDXL), with slacks η_{stb} of 0.004, 0.002 and 0.001. On CBMNIST and SDXL every grid threshold selects the same model (steps 100 and 200/200/198), so the constraint does not change the choice there. On StyleGAN the constraint interacts with the ranking: the tighter

thresholds select step 140 or 160 and the loosest selects step 200 on runs 0 and 2, and the rule keeps the latter. On CBMNIST every candidate has correction score exactly 0, because the categorical anchor measurement agrees with all candidates, so the CBMNIST choice is therefore a tie-break within the feasible set, and the CBMNIST rows test the complete pipeline rather than the ranking.

Relation to the fixed-threshold theorem. The theorem treats τ as a fixed threshold. A threshold independent of the correction scores is the mean stable score of the anchor's single-attribute flip edits plus one source-cluster bootstrap standard error (10,000 resamples, random state 1729), and it equals 0.24953 (SDXL), 0.06225 (StyleGAN) and 1.09×10^{-10} (CBMNIST, whose anchor measurement is nearly exact). Applied to the recorded scores, this threshold admits all 67 SDXL candidates and yields the same models as the operational rule, and admits no StyleGAN or CBMNIST candidate. The reported rows therefore all come from the quantile rule. Both the empirical quantiles and their comparison by correction score use the selection data. Cor. 2 handles this dependence through one event controlling both scores uniformly over the full candidate pool. On that event, its conclusion holds simultaneously for every threshold whose nonempty population retained set, stable margin, and selected slack satisfy the stated conditions. The comparison is with the population optimum at the chosen threshold. Applying this result to a reported run requires uniform error bounds and a population margin compatible with its bootstrap slack, and these conditions have not been verified for the experimental pools. The independent unit in Cor. 3 is a held-out case, equivalently a per-case query average in the Hoeffding bound. The ranges and margins needed by the bound are not verified for these pools, so the available case counts do not by themselves establish its sufficient condition. All MSMS rows, including the CBMNIST stress selectors, follow this rule, and no selection consulted final test outcomes.

E.4 Query banks and empirical discrepancies

Graph roles and measured coordinates. The SDXL graph contains *Male*→*Heavy_Makeup*, and glasses, age, and smile are non-descendants under $do(\textit{Male}{:=}0)$. The general criterion can compare both the target and makeup with an anchor response. In the reported implementation, the anchor condition replaces only the target in the factual tuple. As descendants remain in that condition, correction parser queries read the target alone, and whole-image correction features are disabled when the target has descendants. This avoids rewarding the factual descendant conditioning, but leaves descendant quality outside that direct comparison. The graph is an assumption of this synthetic task, not inferred from names.

Worked example. The reported query banks instantiate the stable and correction families of §3.2 as follows. Under the StyleGAN request $do(\mathit{Male}{:=}0)$, the anchor receives the factual factor tuple with only *Male* replaced, so its other conditioning values, including glasses, remain factual. The candidate's rollout supplies a second tuple, which the scoring renderer maps to a scoring image. Since the anchor's descendant values are fixed by this conditioning, the reported correction queries read only the target, *Male*, against the anchor's edit. The stable queries read the non-descendants, age and smile, against their factual values. At final rendering, the candidate supplies its propagated glasses value together with the factual age and smile to the anchor. Anchor conditioning, candidate scoring, and final rendering are thus three different operations.

Stable and correction banks. The stable reference F is the query-indexed point mass of §3.2, and queries without a factual response are excluded. Stable parser queries read non-descendants and compare them with factual values. Stable feature queries are included only when the factual comparison does not penalize the intended edit, and context-matched whole-image stable features are disabled. Correction features are available for targets without descendants. The face runs use parser response space, whereas CBMNIST uses a hybrid parser/feature configuration. The actual response paths are specified in §E.1, where CBMNIST's exact factor parser and the face image parsers do not observe the same type of object.

Tab. 4 gives the graph, the interventions and the query roles of each benchmark.

Bank sizes. Held-out selection uses eight factual cases on CBMNIST and sixteen on each face benchmark. After query exclusions, the scored banks contain 56 stable and 24 correction queries on CBMNIST, and 96 stable and 32 correction queries on SDXL and StyleGAN. Each query uses

Table 4: Query construction per benchmark. Each row lists the supplied graph, the interventions and the measurements assigned to the reported stable and correction blocks. Stable responses use factual values; correction responses use pretrained target responses. Candidate face parsers read learned query-renderer images; CBMNIST's exact parser reads factor fields.

Benchmark	Graph and interventions	Preservation attributes	Reference attributes	Measurement	Renderer
CBMNIST	$d \to d_c$, $b_c \to b_w$; $do(d)$ or $do(b_c)$	opposite branch: (b_c, b_w) or (d, d_c)	target only: d or b_c	exact factors + image statistics	DDPM twin-noise
SDXL	$M \to H$, $Y \to S$; flip M or Y	(E, S, Y) or (E, H, M)	target only: M or Y	attribute predictor	SDXL img2img
StyleGAN	$M \to E$, $Y \to S$; flip M or Y	(S, Y) or (E, M)	target only: M or Y	attribute predictor	inversion + direction

d, d_c, b_c, b_w denote digit, digit color, bar color, and bar width; E, H, M, S, Y denote Eyeglasses, Heavy Makeup, Male, Smiling, and Young. Candidate-generation budgets are in Tab. 3. Final-bank sizes are in §E.5.

32 response samples. These counts are shared across candidates within a pool and apply to all three reported runs.

Training banks. Training rebuilds a small conditional alignment bank at each step and uses eight stable and four correction queries. The candidate-generation response budget is 4 samples per query for CBMNIST, 2 for SDXL, and 4 for StyleGAN. The held-out selector uses 32 response samples per query, independent of these training budgets. Tab. 3 reports the candidate-generation settings of the candidate model pools used in evaluation. Selection thresholds, observational eligibility, source images, evaluation predictor models, renderer settings, and file hashes are read from the records of the reported runs, not inferred from earlier runs, and a table row is produced only when those records are complete and consistent.

Estimators. For response feature map φ, the operational target averages normalized squared distances between response means. The U-statistic analysis assumes conditionally i.i.d. response replicates across the sample index. The run record therefore fixes the sampling procedure and its initial random state, and violating this condition removes the theorem guarantee. Candidate and reference responses may remain dependent within a paired replicate.

Writing $\mathcal{I}_{\mathrm{stb}}$ and $\mathcal{I}_{\mathrm{corr}}$ for the query indices, the target's V-statistic plug-in is

$$\widehat{D}^{V,\mathrm{avg}}_{\mathrm{stb}}(\mathsf{M}, \mathsf{F}) = \frac{\sum_{r \in \mathcal{I}_{\mathrm{stb}}} w_r \|\hat{\mu}_{\mathsf{M}}(q_r) - \hat{\mu}_{\mathsf{F}}(q_r)\|_2^2 / d_\varphi}{\sum_{r \in \mathcal{I}_{\mathrm{stb}}} w_r},$$

$$\widehat{D}^{V,\mathrm{avg}}_{\mathrm{corr}}(\mathsf{M}, \mathsf{M}_0) = \frac{\sum_{r \in \mathcal{I}_{\mathrm{corr}}} w_r \|\hat{\mu}_{\mathsf{M}}(q_r) - \hat{\mu}_{\mathsf{M}_0}(q_r)\|_2^2 / d_\varphi}{\sum_{r \in \mathcal{I}_{\mathrm{corr}}} w_r}.$$

For $S \geq 2$ response samples of one query, write $p_i, a_i \in \mathbb{R}^{d_\varphi}$ for candidate and reference feature vectors and $\bar{p}, \bar{a}$ for their sample means. The per-query V- and U-statistics are

$$\widehat{d}^{V}_q = \frac{\|\bar{p} - \bar{a}\|_2^2}{d_\varphi}, \qquad \widehat{d}^{U}_q = \frac{\sum_{i \neq j} \langle p_i - a_i, p_j - a_j \rangle}{d_\varphi S(S-1)}. \tag{24}$$

For stable queries the factual measurement is fixed, so $a_i = a$ and the one-sided estimator debiases only the candidate responses. Correction queries use Eq. (24) when responses are paired and its standard two-sample form otherwise. Active parser and feature subfamilies receive equal within-block mass, renormalized when a subfamily is unavailable. With a Kronecker query kernel and linear response kernel, the unnormalized per-query mean-MSE is proportional to empirical joint-MMD on a distinct, uniformly weighted bank (Lem. 4). The operational family-weighted average is generally a different functional. The linear kernel aligns response means rather than full response distributions. Candidate generation uses the displayed V-statistic. Held-out selection retains the same weights but replaces every squared sample-mean term with the paired, two-sample, or fixed-reference U-statistic in Eq. (24), yielding $\widehat{D}^{U,\mathrm{avg}}_b$. The Gaussian-RBF MMD routine is used in selector diagnostics, not in the reported candidate-generation or held-out-selection scores. The evaluation-time squared discrepancy between response-feature means is a diagnostic only, and although it and the selector operate on response features, they are not numerically identical estimators at finite sample size.

E.5 EVALUATION PROTOCOL

Metrics. The final evaluation reads generated pixels for every method, including on CBMNIST, and scores them through the same benchmark-specific parser. A method's internal factor traces therefore play no role in the TS, NDD, and DR values of Tab. 1. Write Φ_V for the evaluation predictor's output for coordinate V, $\hat{I}_i^{(1)}, \dots, \hat{I}_i^{(S)}$ for the $S = S_{\text{eval}}$ ordered renders of case i under $do(\boldsymbol{A}{:=}\boldsymbol{a})$, $\boldsymbol{z}_i$ for the recorded factual factors of its source, and $(I_i^{\text{ref}}, \boldsymbol{z}_i^{\text{ref}})$ for the matched reference image and its factors. With $\boldsymbol{N_A} = \text{ND}_{\mathcal{G}}(\boldsymbol{A}) \setminus \{I\}$ and $\boldsymbol{D_A} = \text{De}_{\mathcal{G}}(\boldsymbol{A}) \setminus (\boldsymbol{A} \cup \{I\})$, the per-case metrics are

$$\text{TS}_i = \frac{1}{S\,|\boldsymbol{A}|} \sum_{s=1}^{S} \sum_{A_j \in \boldsymbol{A}} \mathbf{1}\Big[\Phi_{A_j}(\hat{I}_i^{(s)}) = a_j\Big], \quad \text{NDD}_i = \frac{1}{S\,|\boldsymbol{N_A}|} \sum_{s=1}^{S} \sum_{V \in \boldsymbol{N_A}} \delta_V\Big(\Phi_V(\hat{I}_i^{(s)}), z_{i,V}\Big), \tag{25}$$

$$\text{DR}_i = \frac{1}{S\,|\boldsymbol{D_A}|} \sum_{s=1}^{S} \sum_{V \in \boldsymbol{D_A}} \delta_V\Big(\Phi_V(\hat{I}_i^{(s)}), z_{i,V}^{\text{ref}}\Big), \quad \text{TID}_i = \frac{1}{S} \sum_{s=1}^{S} \frac{\big\|\mathcal{R}(\hat{I}_i^{(s)}) - I_i^{\text{ref}}\big\|_1}{3HW}, \tag{26}$$

where $\mathcal{R}$ resizes a render bilinearly to the reference resolution $3 \times H \times W$ and the per-coordinate discrepancy maps every factor to $[0, 1]$,

$$\delta_V(x, y) = \begin{cases} \mathbf{1}[\,\text{round}(x) \neq \text{round}(y)\,] & V \text{ categorical with more than two classes,} \\ |x - y| & V \text{ binary,} \\ \min\{|x - y|/r_V,\, 1\} & V \text{ bounded continuous with range } r_V. \end{cases} \tag{27}$$

Without this normalization a ten-way coordinate such as *digit* would contribute up to 9 while a binary coordinate contributes at most 1, and the average would leave $[0, 1]$. A categorical coordinate is scored by mismatch rather than by index distance, so *digit* 3 against 8 is one wrong answer. Parser outputs are discrete labels, so the target match in Eq. (25) is exact equality (a continuous coordinate would use $|\cdot| \leq \frac{1}{2}$). A table entry is the mean of these per-case values over the final bank.

Reference for DR. The reference that DR compares against comes from outside the anchor used for selection. On CBMNIST, the procedural generator renders the intervened factor vector directly. For faces, we retrieve a corpus image that carries the requested target value together with the source's non-descendant values, using the nearest Hamming distance when no exact match exists. Ties favor the descendant values predicted by a linear structural model fitted on the observational split. This retrieval is deterministic, so it does not sample an individual counterfactual distribution, and DR measures agreement with the retrieved reference.

Image quality. TID is auxiliary when methods render at different resolutions, since resizing changes pixel distances. FID (Heusel et al., 2017) measures distributional image quality rather than preservation of the source, comparing Inception-v3 pool features of a cell's generated images with those of 2,000 benchmark training images, both resized to 299×299. Each run contributes one hundred cases with eight ordered renders per case. To keep sample counts matched within a benchmark, the CBMNIST and StyleGAN cells pool 2,400 images over three runs, and each SDXL estimate is computed per run from its 800 images. SDXL BD-CLS reports the mean of its two per-run FIDs, the other SDXL FIDs use run 0, and KID uses run 0 for all SDXL methods. All methods on a benchmark share the same reference set, and the 28×28 CBMNIST images are upsampled, so their FID is interpreted only within that benchmark. Tab. 8 also reports the Kernel Inception Distance (Bińkowski et al., 2018), computed with the same reference set.

Runs and uncertainty. Unless stated otherwise, Table 1 and the remaining tables report mean and sample standard deviation over three runs. The standard deviation is taken over run means, since renders within a run are not independent replicates. For MSMS, a run indexes both candidate-pool training and evaluation, whereas a baseline trained once per dataset reuses its model, so its runs vary the benchmark split and any stochastic editing. These different sources of variation mean that overlapping uncertainty intervals do not support an equivalence claim. Such a claim would require a paired comparison on common source cases and interventions with prespecified margins, which the reported comparisons do not include.

Parsers on SDXL. On SDXL, two parser models serve separate roles in optimization and evaluation. The training-selection parser supplies fixed measurement functions during candidate generation and held-out model selection, and the evaluation parser is reserved for final scoring. The frozen final source bank has no identity overlap with the training and validation splits of either parser, and both models pass their predeclared labeled acceptance tests.

Renders per case. Each case is rendered $S_{\mathrm{eval}} = 8$ times, a count the protocol fixed before any final outcome was generated, and $S = 1$ and $S = 4$ use prefixes of the same ordered sample bank as sensitivity analyses. Every method returns an explicit batch of eight renders per case, and a deterministic method may return pixel-identical renders while retaining eight distinct sample records. For the StyleGAN renderer, rendering is deterministic conditional on the sampled factors, so variation across its ordered batch comes from the stochastic descendant mechanisms of the proxy, and the nested prefixes reweight those outcomes.

Qualitative figures. The qualitative figures draw on run 0, whereas every table value uses the full bank. Typical method comparisons and candidate-disagreement examples were chosen by inspecting the recorded images and are illustrative rather than random samples. The repeated-render figure uses the first case with nonidentical renders, and the failure figure shows individually inspected renders with target or preservation errors. A method-comparison case appears only when every displayed method has rendered it. Sources, references, and identical outputs may recur across panels, and allocations, per-case scores, and image hashes are recorded with the assets.

Source review and visual assessment. The source review used for the reported runs rejects an image in full whenever any assigned factor, the single-person requirement, or portrait quality is not jointly satisfied. Replacements preserve the complete factor vector, and the aggregate rejection and replacement counts of this review are not part of the reported results. The visual assessment, in turn, hides method labels and numerical evaluation scores and assigns every arm the same display area, although inspection of original-resolution files can still reveal a method family through resolution or artifacts.

E.6 Benchmarks and baselines

We use three benchmarks, each paired in distribution with its anchor, and compare MSMS with four baselines that span the main design patterns of causal image editing. Each anchor is a frozen generator accessed only through its image interface, and it is never assumed to provide interventional or counterfactual semantics. One paragraph below describes each benchmark with its anchor, and one paragraph each baseline with the adapter that connects it to the common evaluation interface.

ColoredBar-MNIST. **ColoredBar-MNIST (CBMNIST)** (Pan & Bareinboim, 2024; 2025) is a 28×28 synthetic benchmark whose observed tuple pairs an image with the factors digit, digit color, bar color, and bar width, with the graph of prior causal-editing work. Its anchor **ColoredBar-DDPM** is a class-conditional UNet at 28×28 with an 80-class joint condition, a linear 1000-step noise schedule, classifier-free guidance, and DDIM inference. Final renders use img2img sampling from the noised source with strength 0.85, guidance scale 5, and 50 steps, and the benchmark's exact factor interface supplies the CBMNIST parser queries.

SDXL-synthetic. **SDXL-synthetic** is generated by SDXL itself from structured prompts over *Eyeglasses*, *Heavy_Makeup*, *Male*, *Smiling*, and *Young*, so the factor labels are the conditioning inputs used at generation. Its anchor is **SDXL** (Podell et al., 2024) through the img2img pipeline at 1024×1024 with strength 0.70, guidance scale 7.5, and 50 steps, and the same generator is shared with the diffusion baselines BD-CLS and CausalDiffAE. A real-face attribute predictor trained on CelebA-HQ (Karras et al., 2018) screens the generated images during benchmark construction, and a separate predictor trained on the generated images scores the edits.

StyleGAN-synthetic. **StyleGAN-synthetic** is the analogous in-distribution benchmark for StyleGAN3-T. It draws W^+ codes from an attribute-direction bank over *Male*, *Young*, *Eyeglasses*, and *Smiling*. Its anchor is the **StyleGAN3-T** FFHQ-U generator (Karras et al., 2021) at 256×256 under the inversion realization $A_{\mathrm{gan_invert}}$. An e4e encoder trained on CelebA-HQ maps the source to

W^+, and the edit moves the code along the calibrated attribute directions before rendering. BD-CLS and CausalDiffAE are diffusion-based editors and do not run on this GAN-paired benchmark, so its row compares MSMS with CGN and ANCM.

CGN. **CGN** (Sauer & Geiger, 2021) factors an image into object, mask, and background generators, $I = M \odot O + (1 - M) \odot B$, and is trained per benchmark. The modified SDXL and StyleGAN face-CGN adapters synthesize from the intervened condition and a case-and-sample-deterministic random latent. They do not invert or reconstruct the factual pixels. Their NDD and TID thus compare a new face with the factual face and do not measure source-conditioned reconstruction quality. For CBMNIST CGN, a deterministic check of the predeclared final model's saved mask grid classified 37 of 80 joint-condition rows as inverted and two as missing a digit-region mask. We retain these artifacts instead of choosing a replacement model after inspection.

ANCM. **ANCM** (Pan & Bareinboim, 2024) fixes one $\mathcal{G}$-constrained NCM with its own image mechanism and edits inside it, so it does not select among proxies. It is trained per benchmark with the paper's variational image mechanism and, on the face benchmarks, its diffusion refinement, and its edits are read through the same evaluation predictor as every other method.

BD-CLS. **BD-CLS** (Pan & Bareinboim, 2025) edits in a disentangled causal latent space on top of a pretrained generator. The SDXL BD-CLS row uses the training-free pretrained-diffusion editor with one numerical-domain repair. After each optimization step, we project its latent coefficient to the $[-0.2, 1.2]$ interval already applied by the vendored implementation after the loop. Under the vendored $2000\times$ loss scaling, unprojected intermediate iterates produced nonfinite UNet predictions for some cases. The repair adds neither an interval nor an optimization hyperparameter, although it can change previously successful outputs. The reported SDXL row uses the completed reruns for runs 0 and 2 and is labeled as a modified adapter.

The CBMNIST BD-CLS row follows the paper's learned synthetic realization as a task-specific adaptation. Separate models handle digit and bar-color interventions. The digit route excludes digit color, while the bar-color route excludes bar width. Training otherwise uses the vendored upstream-default WGAN/RMSprop path. We apply a narrow-semantics qualifier for this route decomposition because it differs from both the training-free SDXL procedure and an exact upstream-equivalent implementation.

CausalDiffAE. **CausalDiffAE** (Komanduri et al., 2024) uses a diffusion autoencoder, $z = E(I)$, $z' = do(z, \boldsymbol{a}')$, $I' = D(z')$, and is retrained per diffusion-paired benchmark. CausalDiffAE rows follow the paper's Algorithm 2. The decoder is conditioned on the causal representation alone, an intervention writes the label value into the exogenous slot of the target as the mean of the alignment prior (the convention the upstream code uses for its raw MorphoMNIST measurements), the learned causal layer propagates it over the benchmark graph, and DDIM decodes with the benchmark-specific noise coefficient described below. The benchmark's propagated attribute tuple is never read. The CBMNIST model uses the paper's MorphoMNIST recipe (base width 128, channel multipliers 1/2/2, linear 1000-step diffusion, 10K updates at batch 768) on the three-channel benchmark, and the non-averaged parameters are used because the EMA weights (0.9999) still carry about a third of their initialization at this budget. Under this recipe the CBMNIST decoder does not realize interventions on the ten-way digit: the parsed digit saturates at 9 for target values of 2 or more (target success 0.10 on digit against 0.97 on bar color), which is reported as measured.

The DDIM noise coefficient is fixed per benchmark before any scored render by an outcome-free check on the twenty-source null-edit reconstructions of the final model. Here $\eta = 0$ (the paper's default) is kept where those reconstructions show no speckle (CBMNIST, $0/20$), and $\eta = 1$ is used where $\eta = 0$ produces speckle (SDXL, 6–$8/20$). The SDXL $\eta = 0$ run gives TS 0.629 ± 0.015, NDD 0.406 ± 0.007, DR 0.345 ± 0.026 and TID 0.218 ± 0.002 over the same three runs, with 276 of its $2{,}400$ renders above the sources' maximum high-frequency energy and 44 washed out, against none at $\eta = 1$. CausalDiffAE reports no face benchmark. Our SDXL row uses the same faithful path with the model trained on the frozen face pool at the paper's 128×128 CausalCircuit RGB model scale, with 4000-step cosine diffusion and a 20K-update budget. It is not an upstream CelebA-HQ reproduction.

Ablations. The render-count sensitivity uses the $S = 1$ and $S = 4$ prefixes of the same ordered sample bank as the $S = 8$ tables. The preservation ablation keeps the selected model and every render random state and only recomputes the non-descendant factors instead of holding them at their factual values. The selection-rule comparison reselects from the same saved pools under the alternative rules of §5.1. FID and KID are reported with matched sample counts. §G gives the tables.

Repetition and fairness. Every cell of Tab. 1 aggregates three runs, except SDXL BD-CLS with runs 0 and 2. For MSMS a run indexes both candidate-pool training and evaluation, whereas a baseline trained once per benchmark reuses its model across runs. Methods are compared on identical source cases, interventions, render counts, and evaluation predictors. Baseline hyperparameters follow the recipes of their papers, with the adapter choices recorded above.

E.7 Computational cost of selection

Let $K = |C|$ be the number of eligible candidates, n_{fit} the number of observational examples used for eligibility assessment, Q the number of held-out selection queries, $S \geq 2$ the response samples per query, and d the maximum response-feature dimension. The observational screen requires $O(|\mathfrak{M}_{\mathrm{cand}}|n_{\mathrm{fit}})$ factor-prediction evaluations. Scoring the eligible candidates requires $O(KQS)$ candidate responses. Each face response combines factor inference and propagation with the learned scoring renderer and fixed attribute predictor, and the CBMNIST parser reads the resulting factor fields. Anchor responses are cached across candidates, requiring at most $O(QS)$ generated responses for the selection bank. Stable queries reuse factual measurements, so only correction queries require anchor generation.

The linear U-statistic can be evaluated from sufficient sums. For a query with feature dimension $d_q \leq d$ and paired vectors $p_s, a_s \in \mathbb{R}^{d_q}$, writing $v_s = p_s - a_s$ gives

$$\widehat{d}_q^U = \frac{\left\|\sum_{s=1}^S v_s\right\|_2^2 - \sum_{s=1}^S \|v_s\|_2^2}{d_q S(S-1)},$$

which requires $O(Sd)$ arithmetic per query. The fixed-factual-reference and two-sample forms admit the corresponding sufficient-sum evaluations. Thus, score computation adds $O(KQSd)$ arithmetic to response generation. Bootstrap slack estimation and the threshold search operate on the stored scalar query scores and require no further generator evaluations.

These counts describe held-out selection. Candidate training, model loading, and source-conditioned rendering of subsequent edits contribute separate costs. In particular, final face generation uses the pretrained renderer and its backend-specific source encoding or inversion, as described in §E.1.

E.8 Compute resources

Experiments ran on 48-GB NVIDIA RTX 6000 Ada and A6000 GPUs and 24-GB A5000 GPUs, with one GPU per job. The reported runs used about 1,060 GPU-hours in total, measured in RTX 6000 Ada-equivalent time. The SDXL BD-CLS baseline accounts for about 940 of these hours across its two runs. MSMS candidate training, selection, and final rendering took about 35 GPU-hours over the three benchmarks and three runs, and the preservation ablation about 16. These figures exclude the pretrained generators and attribute predictors reused across runs and the development runs.

F Further discussion

This appendix discusses the plausibility of the two conditions of Thm. 2, relates the assumptions of MSMS to prior work, and states the limitations and broader impacts of the method.

F.1 Plausibility of the assumptions

The two conditions of Thm. 2 are selection assumptions rather than identification assumptions on the true SCM. They describe a regime in which the anchor is useful as additional evidence for ranking observationally feasible proxies, and they do not assert that the counterfactual image distribution is

identified from the graph and the data. Their intended reading is therefore practical, and this section states when each can be expected to hold, when it fails, and what the reported experiments do and do not establish about it.

The first condition. The first condition requires that the stable threshold τ not exclude every correct proxy, so that at least one member of $\mathfrak{M}_{\mathrm{cf}}$ satisfies the stable constraint. Operationally, there must exist a proxy that keeps the non-descendants at their factual measurements while giving the correct answer on the editing family. This is most plausible when the intervention is semantically local relative to the image and the anchor already renders the unaffected content well. In portrait editing, for example, one expects a feasible proxy in which changing the smile alters expression-related regions while identity, illumination, and background are preserved. In medical image editing, a similar reading holds when a pretrained model already captures anatomy and scanner style, while the causal semantics of lesion load or treatment response remain underspecified. The condition does not say that the anchor is a correct counterfactual editor. It says only that the stable behavior the anchor renders is compatible with at least one correct causal completion. A restrictive threshold, or a measurement that reads an intended change as a stable violation, can exclude every correct proxy, and a sufficiently expressive candidate family with stable measurements that cover the non-descendant attributes makes the condition plausible.

The second condition. The second condition is the more substantive one and the point at which the method moves beyond observational fit. It controls counterfactual risk through an upper bound determined by the excess correction discrepancy, with the slack β absorbing the anchor's own error on the correction queries. It does not assert that two candidates' true risks are ordered as their discrepancies are. If closeness to the anchor on the correction queries were unrelated to correctness, minimal shift would be a ranking heuristic with no causal guarantee. Prop. 4 in §D.12 isolates one sufficient route by separating the requirement into two parts. First, the correction query family must be rich enough to control the editing risk, which is a query-design condition. Second, the anchor must lie close, on those same queries, to at least one correct proxy in the retained class, which is an anchor-quality condition. The constants α and β then follow from a risk-control premise on the retained class together with declared additive slacks for factor and image recovery.

What the experiments establish. §5.1 tests the second condition on the StyleGAN pools by relating anchor disagreement to the editing error of the candidates. A rank correlation on a finite pool is consistent with the condition, but it does not establish the uniform inequality, its constants, or exact recovery of the counterfactual distribution. On CBMNIST the anchor response is read through the exact factor interface, so the correction score carries no image evidence there, and on SDXL every candidate satisfies the stable constraint, so the constraint does not change the selection. Both conditions are population statements about the retained class, whereas the observational-fit screen of §E.2 establishes eligibility for the finite-pool selector rather than membership in the exact feasible class.

What the conditions assert and how they are instantiated. Semantically the conditions are strong, since they assert the existence of a correct proxy within the retained class and a one-sided risk-control inequality between correction discrepancy and counterfactual risk. Operationally the framework is milder than latent-space approaches that require a globally invertible causal chart, as the anchor is used only through a finite family of queries after a shared measurement. Diffusion models, prompt-based editors, and other black-box generators rarely expose a unique inverse from an image to the latent variables that generated it, and the framework trades this invertibility requirement for the query-response conditions above.

F.2 Relation to prior assumptions

Factor–image invertibility. Identifiability arguments for image-based counterfactual generation differ in what they assume about recovering factors from the image. The augmented SCM underlying ANCM assumes that the observed factors are recoverable from the image, $\boldsymbol{V} = h(I)$ (Pan & Bareinboim, 2024), whereas BD-CLS assumes joint recovery of labeled and unlabeled generative factors, $(\boldsymbol{V}, \boldsymbol{L}) = h(I)$ (Pan & Bareinboim, 2025). These assumptions concern the image-generating SCM. MSMS does not assume an exact image-to-factor inverse. Its fixed predictors provide measurements

Table 5: Representative prior approaches, the assumption each relies on, and the corresponding relaxation in MSMS. Citations are non-exhaustive within each row.

Approach	Key requirement	Relaxation in MSMS
Causal representation learning via nonlinear ICA (Khemakhem et al., 2020; Locatello et al., 2019)	Auxiliary labels and an invertible or weakly identifiable decoder relating latents to observations	Latent invertibility is not required, and models are compared through their responses (Def. 3)
Deep structural causal models (Pawlowski et al., 2020; De Sousa Ribeiro et al., 2023)	A specific parametric generative form absorbing the SCM mechanisms	Form-agnostic at the interface, since any model exposing T_{M} can serve as anchor or proxy
Architecture-specific causal generators (Yang et al., 2021; Komanduri et al., 2024; Sauer & Geiger, 2021)	A decoder tied to a causal VAE, a causal diffusion autoencoder, or independent mechanisms	The same selection rule applies across SDXL, StyleGAN, ColoredBar-DDPM, and NCM proxies
Neural causal models (Xia et al., 2021; 2023)	Expressive for L_3, while observational data leave the counterfactual layer undetermined	NCMs form the proxy class, and a selector ranks them by agreement with the anchor
Interventional or multi-environment representation learning (Ahuja et al., 2023)	Interventional or multi-environment data	A queryable anchor replaces interventional data, with sampled rather than collected queries
Direct counterfactual image editing (Jeanneret et al., 2022; Ben Alaya et al., 2024; Song et al., 2024; Huang et al., 2025)	Single-model edit alignment through target and preservation losses, without a population-level guarantee	Such editors serve as anchors, and the minimal-shift preference becomes a selection rule with a conditional risk guarantee (Thm. 2)
Backbone-specific minimal edits (Pan & Bareinboim, 2025)	A disentangled latent space or edit pathway of one backbone	Backbone-agnostic interface, where the anchor's edits enter only as responses on a shared measurement
Measure-theoretic causal spaces (Park et al., 2023; 2026)	Causality axiomatized on an underlying measure space	Axiomatized on the query-response interface, which covers SCMs and black-box generators uniformly (§C.3)

of an image, and its guarantee is conditional on the selection assumptions stated above. In Def. 1 the factual factor values may remain uncertain after observing the source image, Prop. 1 represents this uncertainty through conditional distributions over factor and rendering noise, and the anchor enters only through its response distributions $\mathsf{T}_{\mathsf{M}_0}(q)$, which absorb its randomness and prompt sensitivity.

Categorized comparison. Tab. 5 groups representative prior approaches by the assumption they rely on and records where MSMS relaxes that assumption.

What MSMS still requires. These relaxations are conditional. MSMS requires a fixed causal graph over the observed variables, on the same footing as every graph-based counterfactual framework above, a proxy family $\Omega_{\boldsymbol{\theta}}(\mathcal{G})$ rich enough to contain a correct model within the retained class, an anchor whose interface supports the stable and correction queries and is informative in the sense of the second condition of Thm. 2, and the graph-induced split of the query domain, which is itself a modeling choice. MSMS therefore shifts the assumption surface from identification, an assertion about the true SCM, to selection, an assertion about the informativeness of the anchor and the richness of the candidate pool on a designed query interface. Its observational-fit screen implements an approximate candidate restriction, and passing this screen establishes eligibility for selection, whereas the counterfactual-risk theorem assumes membership in the exact feasible class.

Minimal change and model selection. Minimal-change criteria also guide backtracking counterfactuals, which vary the exogenous state (von Kügelgen et al., 2023), and natural counterfactuals, which permit limited changes to causal ancestors when needed for feasibility (Hao et al., 2024). MSMS uses the idea differently. It retains ordinary intervention semantics and does not impose a

new ordering over counterfactual worlds. Once the SCM semantics and the observationally feasible class are fixed, minimal shift acts as a model-selection preference that keeps the stable responses at their fixed values and uses the correction queries only to resolve the remaining ambiguity. Causal representation learning likewise clarifies the role of side information, since Locatello et al. (2019) establish nonidentifiability without such information and Khemakhem et al. (2020) and Ahuja et al. (2023) recover representations up to equivalence using auxiliary variables or interventions. Identifying those factors does not identify an individual's counterfactual image, but it illustrates why information beyond an observational distribution is necessary, and MSMS supplies it through the anchor.

Query-response representations and causal distances. Measure-theoretic causal spaces represent interventions through causal kernels over probability spaces (Park et al., 2023). Counterfactual spaces add a representation of information shared across possible worlds (Park et al., 2026). Our queryable-model definition describes the responses that a model exposes on a specified query domain. This interface allows an SCM and a pretrained generator to be compared through the same measurements despite different internal parameterizations. Related distances compare model-induced observational, interventional, and counterfactual distributions (Peyrard & West, 2021), and MSMS uses this behavioral comparison for constrained selection relative to the anchor, with the implemented discrepancy comparing feature means over the chosen query family. The image-risk implication follows from the reference-informativeness condition in Thm. 2.

Backend scope. The interface accommodates different latent parameterizations, including diffusion and GAN backbones, with backend-specific rendering and inversion. The evaluation tests compatibility with both interfaces, and invariance to the choice of anchor remains open. Direct image editors may serve as anchors under the two conditions of Thm. 2.

F.3 Limitations

The supplied graph and fixed measurement functions determine which changes the rule measures, so errors in either can change its preference. The SDXL corpus violates an independence implied by its supplied graph and does not validate the exact-graph theory, and robustness to graph errors and substantially different measurements remains open. Factor-labeled observational data also supervise the proxy and query construction, despite the absence of counterfactual training or selection labels.

The candidate pool bounds what can be selected, and the reported pools contain the iterates of single training trajectories. The measured response further limits the inference, since face scores use a learned scoring renderer, CBMNIST scores read structured factors, and final images use the anchor. The reported correction queries read the targets alone (§E.1). Neither these scores nor the matched face references establish individual counterfactual accuracy, and the study lacks real-image validation with ground-truth counterfactuals.

For the operational quantile threshold, Cor. 2 supplies a simultaneous implication under uniform score accuracy and the selected threshold's margin and slack conditions (§E.3). These conditions, reference informativeness, and factor/renderer recovery accuracy remain unverified for the final pools. The conditional risk theorem additionally requires exact feasibility and a valid surviving model. These are substantive assumptions, rather than consequences of a low empirical score.

Repeated pretrained rendering, model pooling, and held-out scoring add computation. Incomplete accounting and heterogeneous baseline recipes prevent a like-for-like total-cost comparison, so we make no quantitative efficiency claim. Matched perturbations of the query bank, measurement functions, and pretrained reference would assess sensitivity while holding the candidate pool fixed. Source-review procedures and their remaining reporting status appear in §E.5.

F.4 Broader impacts

Counterfactual face editing inherits the risks of the pretrained image generators on which it relies, including identity misuse, misinformation, and non-consensual editing. MSMS selects proxy causal models rather than introducing a new image generator, but selection can still improve the usability of an existing editor. The face experiments use synthetic portraits, and no identity-targeting model is released.

G ADDITIONAL EDITING RESULTS

This appendix supplements the editing results of §5.2 with three measurements on the same evaluation bank. §G.1 reports how the metrics change with the number of renders averaged per case, and §G.2 isolates the effect of holding non-descendants at their factual values. §G.3 reports the distributional image quality behind the FID column of Tab. 1. Mean differences within the reported standard deviations, such as the target success of MSMS against BD-CLS on CBMNIST and against CausalDiffAE on SDXL, are not read as improvements.

G.1 SENSITIVITY TO THE NUMBER OF RENDERS

Each task is rendered eight times under fixed ordered random states, and every render is scored separately. Tab. 6 reports the predeclared nested-prefix analysis derived from each ordered eight-render bank, in which every metric is recomputed from the first one, four, and eight renders of every case. Restricting the scoring to the first render moves TS by at most 0.010 and DR by up to 0.057, so the descendant metric is the more sensitive to using a single render in these banks. This prefix comparison does not by itself establish a variance-reduction rate.

Table 6: Render-count sensitivity. Scores use nested prefixes of the eight ordered renders from the selected model. S is the number of leading renders scored, and the reported protocol uses $S = 8$. Mean±SD over three evaluation runs.

Benchmark	S	TS↑	NDD↓	DR↓
CBMNIST	1	0.990±0.010	0.010±0.005	0.683±0.021
	4	0.983±0.009	0.011±0.003	0.702±0.036
	8	0.980±0.004	0.011±0.003	0.715±0.019
SDXL	1	0.667±0.021	0.076±0.007	0.337±0.031
	4	0.659±0.007	0.073±0.004	0.367±0.011
	8	0.660±0.011	0.075±0.002	0.370±0.006
StyleGAN	1	0.927±0.006	0.097±0.008	0.317±0.090
	4	0.929±0.010	0.095±0.001	0.271±0.021
	8	0.931±0.010	0.098±0.003	0.260±0.007

G.2 HOLDING NON-DESCENDANT FACTORS

Table 7: Preservation at inference. Both arms use the same selected model, evaluation predictor, final tasks and ordered rendering random states, and differ only in whether the non-descendants of the intervened variable are held at their factual values or recomputed. Mean±SD over three runs.

	Non-descendants held at factual values (MSMS)			Non-descendants recomputed		
Benchmark	TS↑	NDD↓	DR↓	TS↑	NDD↓	DR↓
CBMNIST	0.980±0.004	0.011±0.003	0.715±0.019	0.975±0.002	0.412±0.006	0.708±0.021
SDXL	0.660±0.010	0.075±0.002	0.370±0.006	0.636±0.011	0.297±0.005	0.386±0.008
StyleGAN	0.931±0.010	0.098±0.003	0.260±0.008	0.939±0.007	0.529±0.016	0.255±0.009

Tab. 7 compares the same model, evaluation predictor, final tasks, and ordered rendering random states with non-descendant factors held or recomputed, and recomputing them changes TS and DR much less than NDD. StyleGAN's binary non-descendant attributes drift at approximately chance level, 0.5, while the ten-way CBMNIST digit changes class in 86–90% of samples, and SDXL retains more of the source, consistent with source-conditioned rendering at strength 0.70. Fig. 5 shows the same source images edited with and without the invariance step, and in its StyleGAN example the age intervention removes glasses when non-descendants are recomputed, although the supplied graph gives no causal path from age to glasses. The comparison thus exposes factor-reconstruction errors that retaining factual values prevents from reaching the renderer. Both arms use the same selected model, so the comparison isolates the inference-time rule from model selection, and additional cases appear in Fig. 16.

G.3 DISTRIBUTIONAL IMAGE QUALITY

Table 8: Image distribution metrics. Each cell's generated images are compared with $2{,}000$ images of the benchmark's training bank, from Inception-v3 pool features at 299×299. Sample counts are matched across methods within each benchmark: $2{,}400$ images over three runs except SDXL, where every cell uses the 800 images of run 0. KID is mean$\pm$SD over 20 subsets of 500. Bold marks the lowest value per benchmark.

Benchmark	Method	FID↓	KID↓ ($\times 10^3$)
CBMNIST	CGN‡	57.1	46.8±1.9
	ANCM	41.9	19.0±1.2
	BD-CLS	20.8	19.3±1.7
	CausalDiffAE	**8.6**	**4.9±0.9**
	MSMS (ours)	13.0	9.8±0.8
SDXL	CGN‡	142.8	121.4±2.0
	ANCM	180.2	175.8±2.2
	BD-CLS	66.2	21.5±1.2
	CausalDiffAE	97.3	87.7±1.7
	MSMS (ours)	**45.6**	**14.6±1.0**
StyleGAN	CGN‡	**15.0**	**1.6±0.5**
	ANCM	158.8	164.0±2.6
	MSMS (ours)	68.4	23.5±1.3

‡Generates a new image from the intervened labels instead of editing the source.

Tab. 8 lists the Inception distances behind the FID column of Tab. 1, together with the Kernel Inception Distance. These distances assess the rendered image distribution rather than the causal response of the selected model or the preservation of a particular source. ANCM's low TID alongside its high FID illustrates why pixel proximity to a matched reference does not itself imply a detailed edit. Its face outputs are smoothed, whereas MSMS reuses pretrained rendering and retains image detail. CGN changes appearance beyond the requested factor, consistent with its label-conditioned synthesis, and FID favors CGN on StyleGAN and CausalDiffAE on CBMNIST. Target success, image similarity, and distributional quality therefore describe different aspects of editing, which Tab. 1 reports side by side.

H ADDITIONAL FIGURES

The following figures extend the main-paper comparisons with benchmark references, additional edits by every method, further invariance cases, repeated renders, and failure cases. All image assets come from the recorded runs used for evaluation. Method examples were selected by inspecting images, whereas repeated-render cases and failure cases follow the explicit rules stated below. These figures illustrate behavior rather than provide a random sample, and the quantitative tables summarize the full evaluation bank (§E.5).

H.1 BENCHMARKS AND REFERENCE IMAGES

Fig. 11 shows the supplied graph of each benchmark and samples from its training set. CBMNIST couples the digit with its color and the bar color with the bar width. An intervention on the digit may therefore change the digit color and an intervention on the bar color may change the bar width, while the other branch consists of non-descendants. The two face benchmarks share the same structure over different attributes, and their remaining attributes are non-descendants of both intervention targets. Fig. 12 shows the descendant references used by DR for two sources per benchmark. CBMNIST references are procedurally rendered from the intervened factor vector, whereas face references are corpus images matched to the requested target and non-descendant values, with descendant tie-breaking from a fitted factor model (§E.5). They need not depict the same person, and a face reference can serve several cases or appear elsewhere as a source, and none is an observed individual counterfactual.

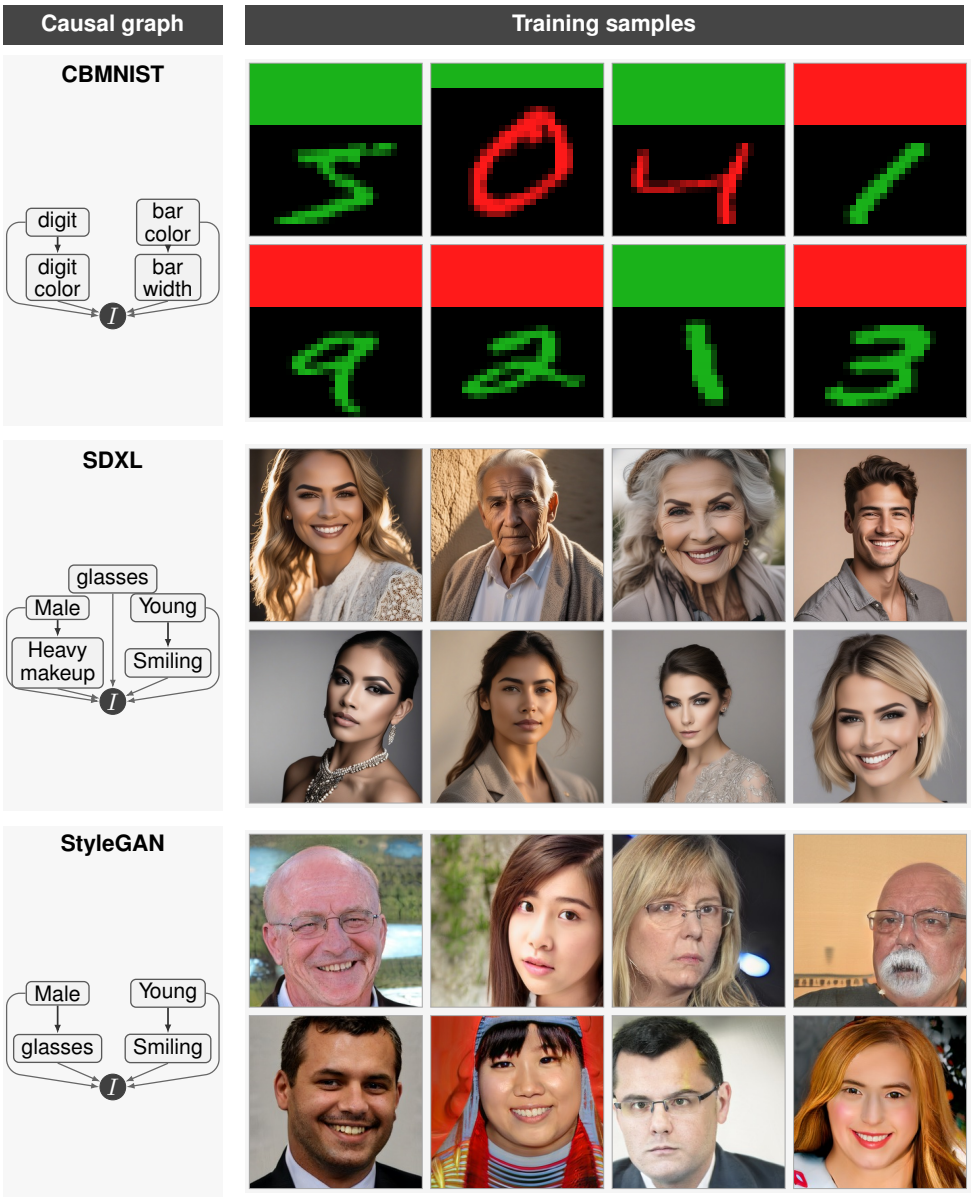


Figure 11: **Benchmark graphs and samples.** Each supplied graph is paired with eight training images. Boxes denote factors, arrows connect each factor to its children and to the image I, and no intervention is marked.

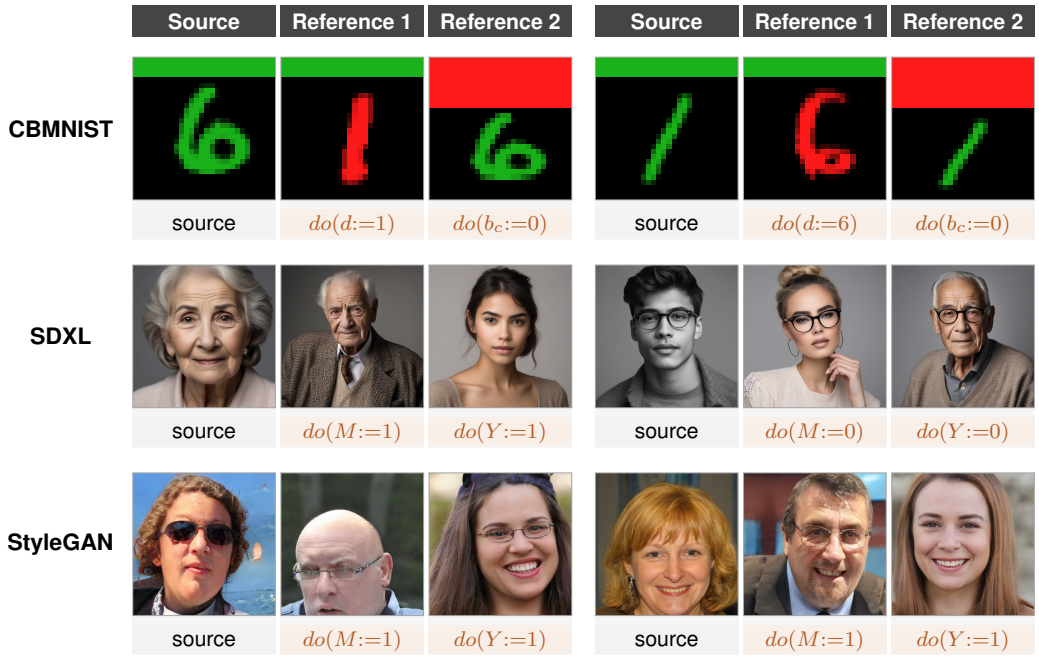


Figure 12: **Descendant reference images.** Each source is followed by the two intervention-matched corpus references used to score descendants, labelled by their intervention with d: digit, b_c: bar color, M: male, and Y: young. These are retrieved images rather than edits, so the face references depict another person.

H.2 Edits by every method

Fig. 6 compares one case per benchmark, and Figs. 13 to 15 expand to four cases per benchmark, two per intervention. All methods receive the same factual image and intervention, with the allocation documented in §E.5. CGN synthesizes a new image from the intervened labels rather than editing the source, which the rows make visible.

H.3 Non-descendant invariance on additional cases

Fig. 16 extends Fig. 5 to four cases per benchmark, and within each pair the model, the intervention, and the rendering random states are identical, and only the invariance step differs. The non-descendants of the intervened variable are held at their factual values in the MSMS column and recomputed from their mechanisms and the inferred noise in the other column. The label above each pair gives the intervention of the case, and the three column groups are the three benchmarks.

H.4 Eight ordered renders of one case

Fig. 17 shows eight ordered renders of one case per benchmark, with per-render target, preservation, and descendant checks. Their means contribute to TS, NDD, and DR in Tab. 1. Each row uses the first case in evaluation order with nonidentical renders. Repeated tiles reflect identical outputs, since the StyleGAN renderer is deterministic given the factors and all eight renders coincide in six of its hundred cases. Across the banks, the renders of one case differ on at least one check in $95/63/82\%$ of cases on CBMNIST, SDXL, and StyleGAN over three runs. They agree on the target in $86/64/89\%$ of cases and on the non-descendant attributes in $84/78/88\%$, but on the descendant in only $9/68/23\%$, which is why DR is the metric most sensitive to the number of renders in §G.1.

H.5 Failure cases

Fig. 18 shows, for each benchmark, the case with the lowest target success and the case with the largest non-descendant drift under the selected model. These are extrema of the recorded TS and

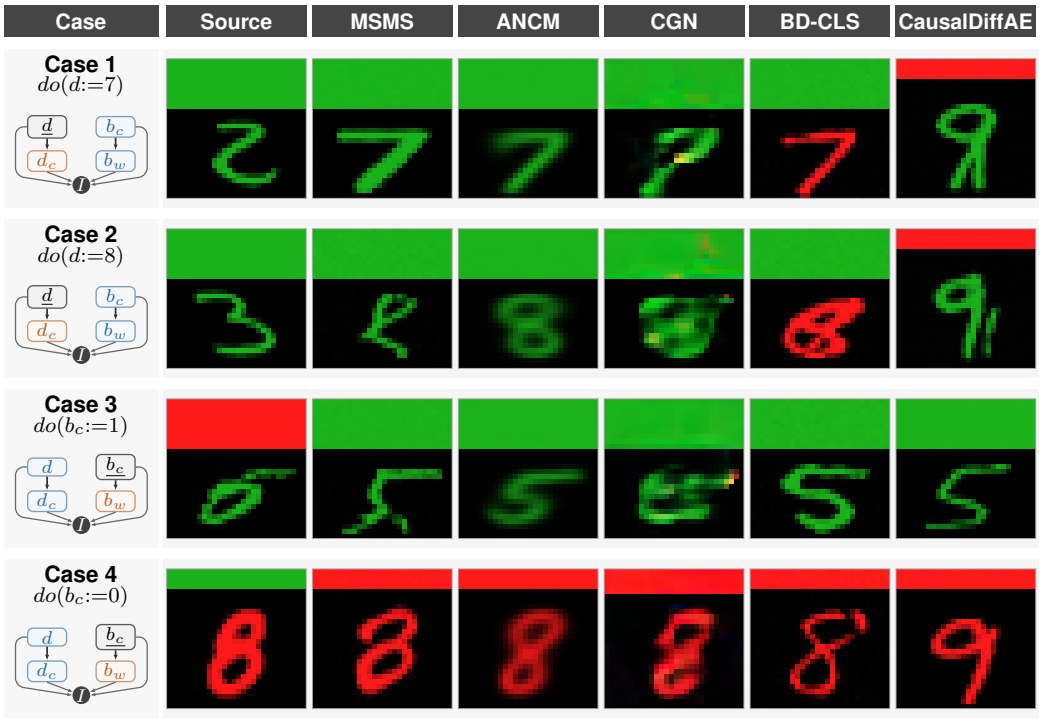


Figure 13: **CBMNIST editing results.** Rows share the source and intervention across methods. In each graph the underlined target is black, descendants are orange, and non-descendants are blue, with d: digit, d_c: digit color, b_c: bar color, b_w: bar width.

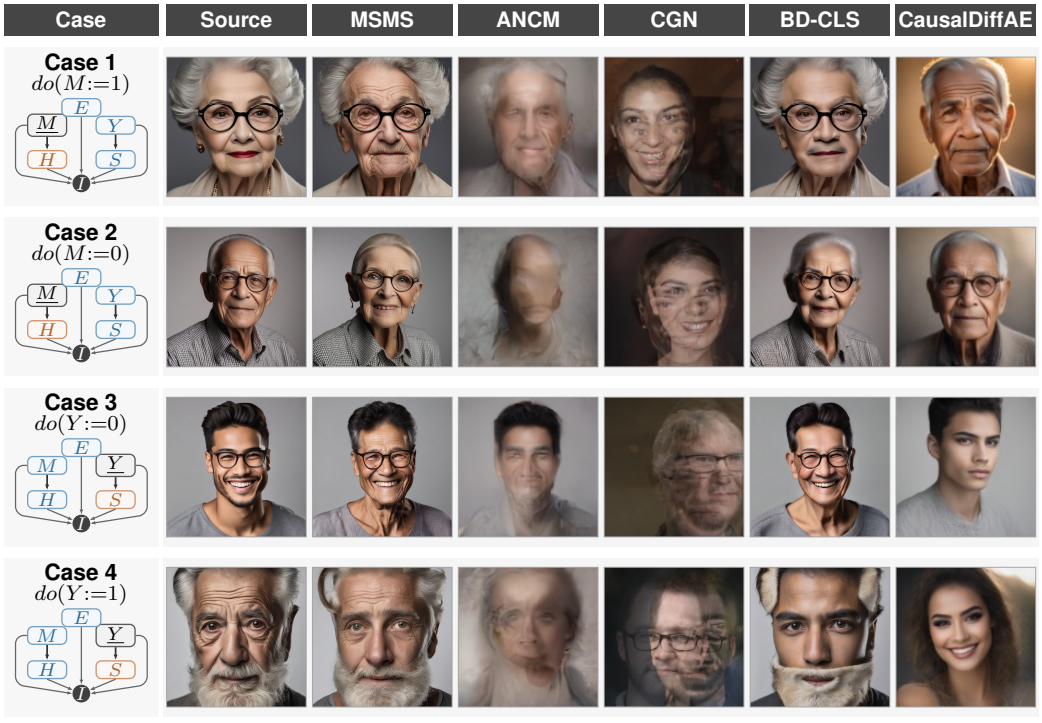


Figure 14: **SDXL editing results.** Rows share the source and intervention across methods. In each graph the underlined target is black, descendants are orange, and non-descendants are blue, with M: male, Y: young, E: eyeglasses, H: heavy makeup, S: smiling.

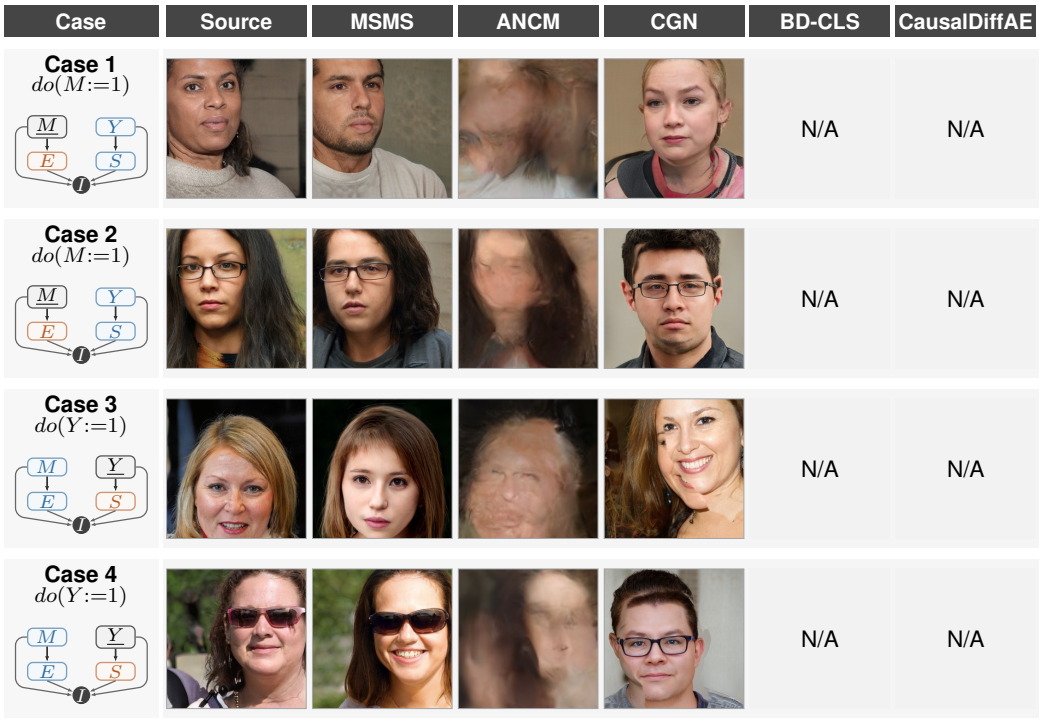


Figure 15: **StyleGAN editing results.** Rows share the source and intervention across methods. In each graph the underlined target is black, descendants are orange, and non-descendants are blue, with M: male, Y: young, E: eyeglasses, S: smiling. N/A denotes an unavailable pairing.

NDD metrics rather than a claim about the visually worst output under every criterion. Each row label names the benchmark, which of the two scores selected the case, and the intervention.

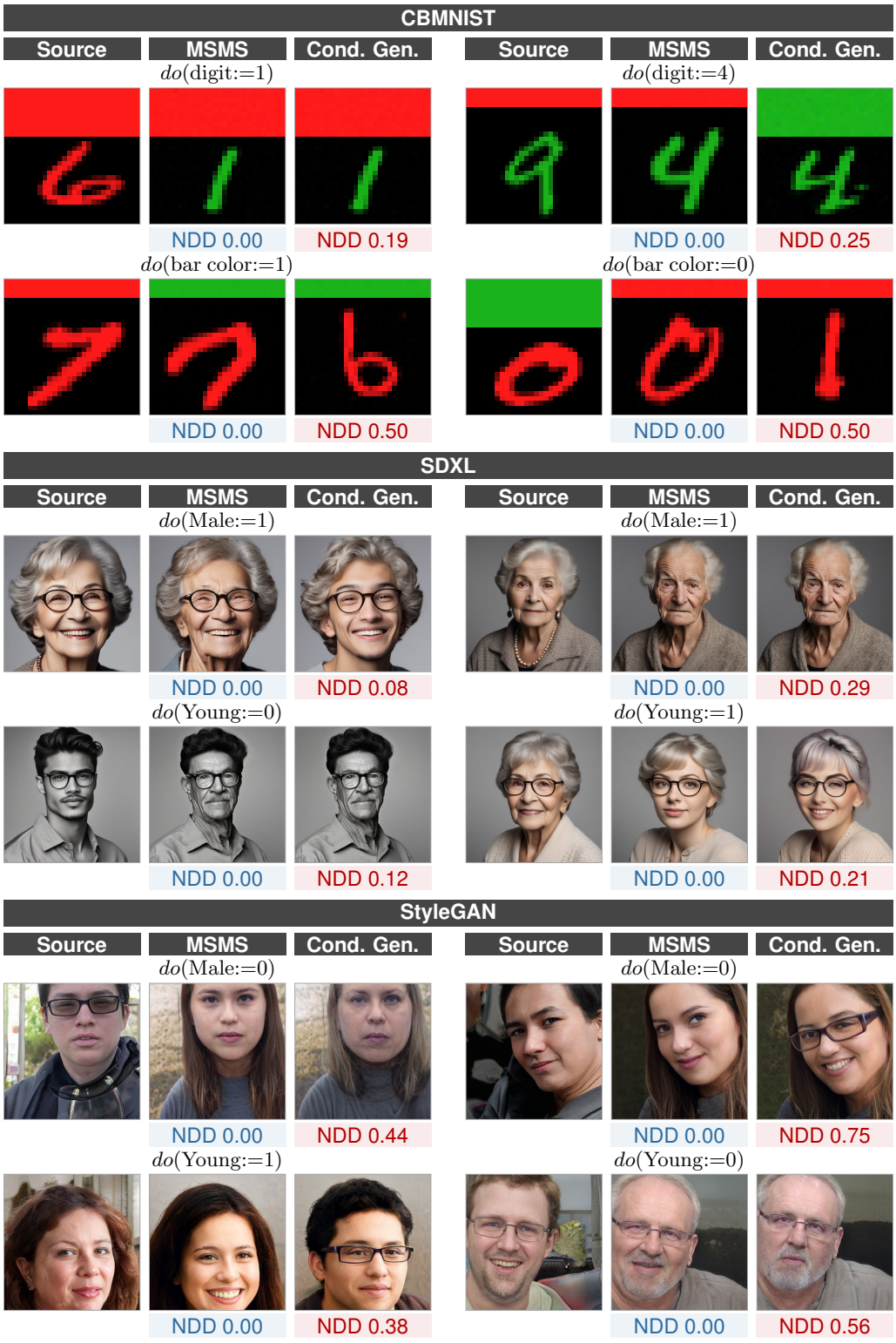


Figure 16: **Additional preservation examples.** Each triple shares the model, intervention and rendering random states. MSMS holds non-descendants at their factual values; conditional generation (Cond. Gen.) recomputes them. Examples are selected by visual inspection: the upper-right CBMNIST triple uses render 3, and the others use render 1. NDD bands report case means over all eight renders.

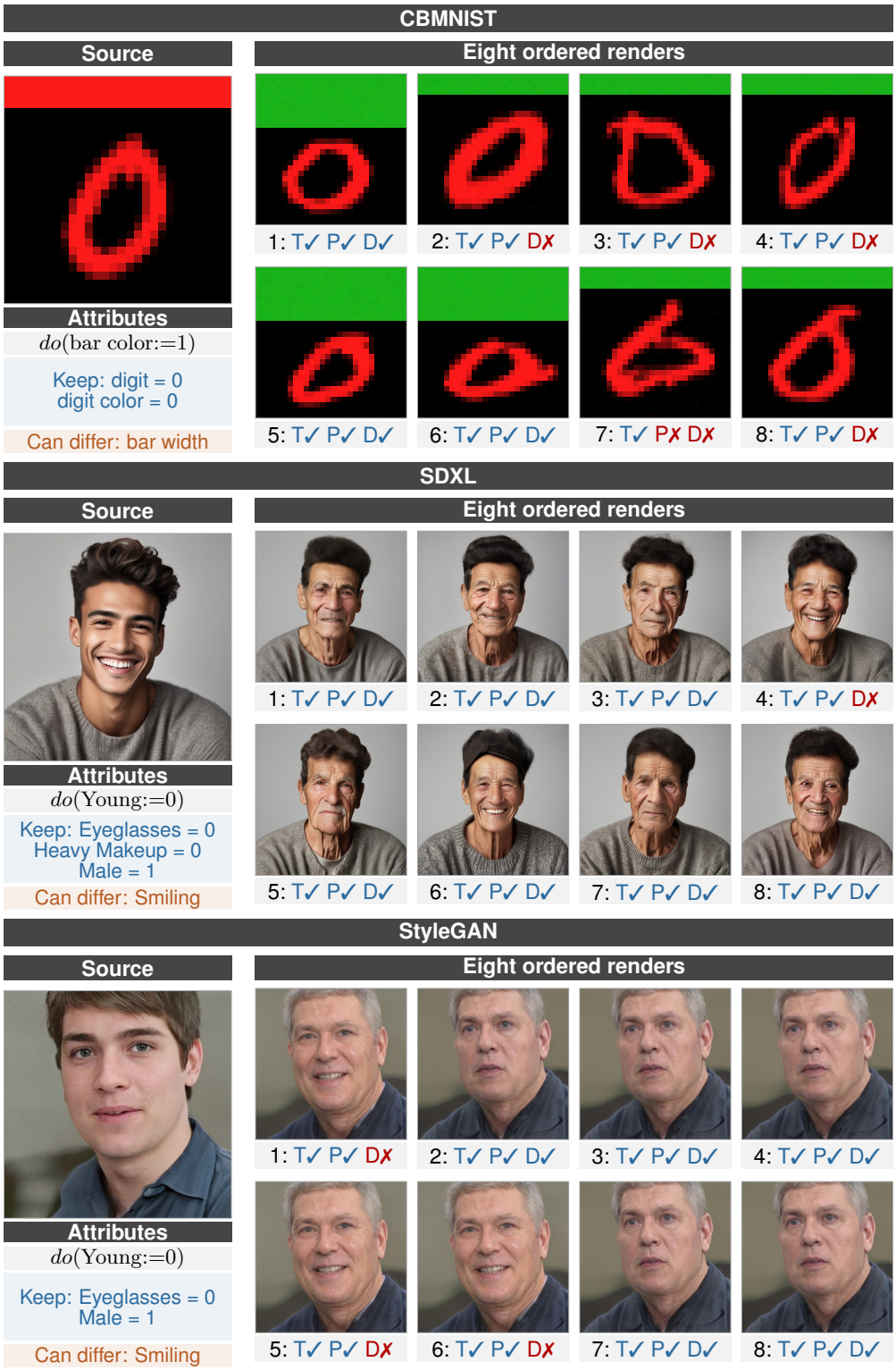


Figure 17: **Repeated renders.** Each block shows a source and eight ordered renders. Gray gives $do(\cdot)$; blue (Keep) lists fixed non-descendants; orange (Can differ) lists descendants allowed to change. T and P check target success and preservation. D checks agreement with the benchmark descendant reference (bar width $= 1$ for CBMNIST; Smiling $= 0$ for faces), not a requirement to keep the source value.

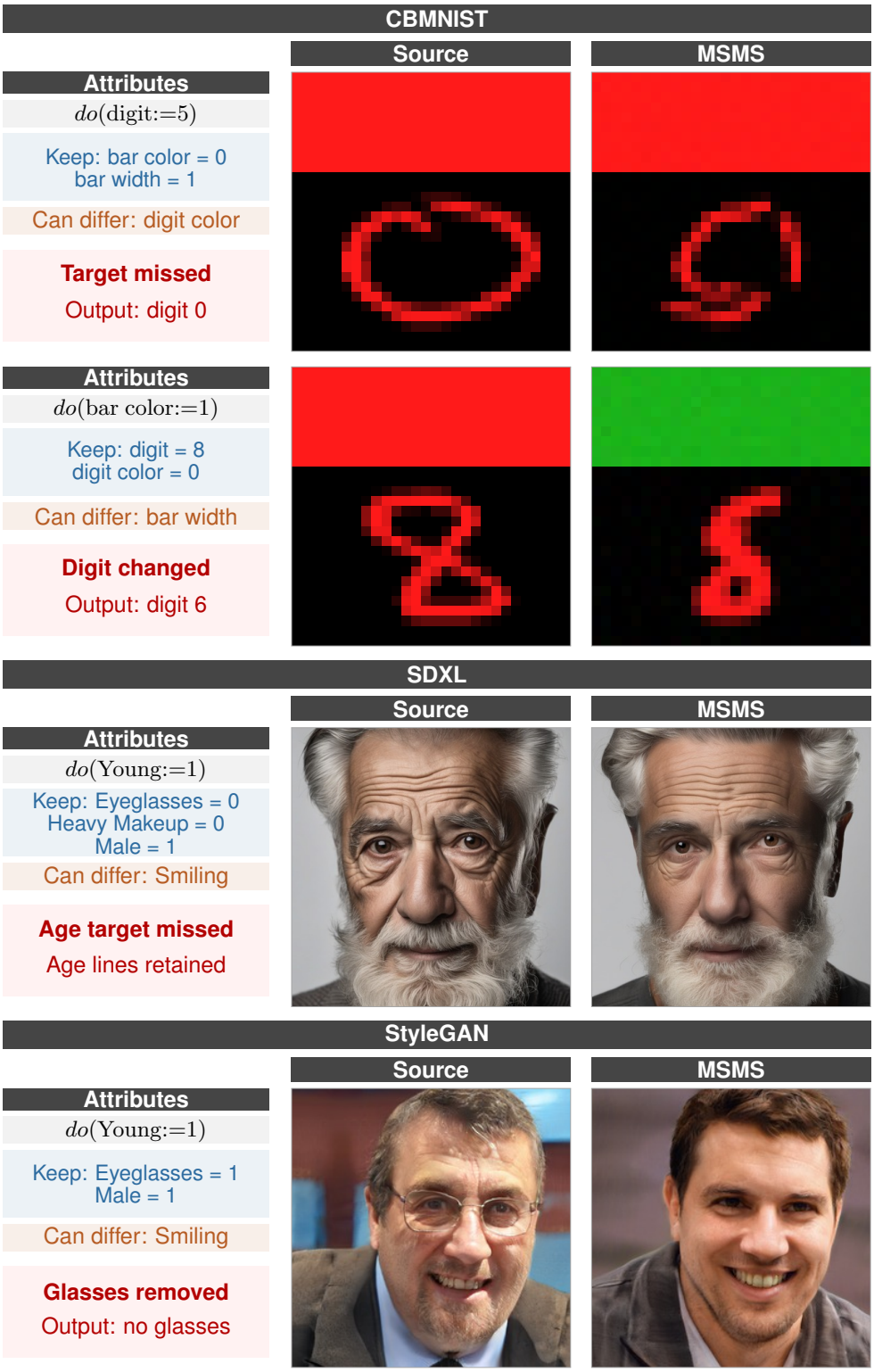


Figure 18: **Visible failure cases.** Individual renders selected by visual inspection from the recorded seed-0 bank. Gray gives $do(\cdot)$; blue (Keep) lists fixed non-descendants; orange (Can differ) lists descendants allowed to change. Red identifies the visible error. Each example also fails its per-render target or preservation check. These examples illustrate specific errors rather than their frequency.