When Does Synthetic Data Help? A Spectral Theory of Task-Relevant Domain Gap with Applications to Guided Generation and Bias Auditing

Anonymous CVPR submission

Paper ID *****

Abstract

When does synthetic data help, and when does it hurt? The dominant answer — reduce the domain gap as measured by FID or perceptual distance — is insufficient: photorealism and task-relevant feature fidelity are not the same thing. We propose a spectral decomposition *of the domain gap into a task-relevant component $\Delta_{\mathcal{T}}$ (distance in the directions the task model's Jacobian is sensitive to) and a task-irrelevant component $\Delta_{\perp}$ (distance in the orthogonal complement). We prove that only $\Delta_{\mathcal{T}}$ bounds downstream performance, while $\Delta_{\perp}$ has zero first-order effect — yet existing generation methods spend the bulk of their compute budget on task-irrelevant photorealism. This motivates* ***SPECTRE*** *(**Spec**tral **T**ask-**R**elevant Synth**e**sis), a diffusion-based generation method that uses the frozen task model's Jacobian to guide the denoising process toward task-relevant realism while leaving irrelevant directions unconstrained. SPECTRE achieves the same downstream mIoU as a standard latent diffusion baseline with $3\times$ fewer generation steps, and outperforms GTA-V and SYNTHIA by $+11.3$ and $+16.1$ mIoU points on Cityscapes with equal data volume. We further show that $\Delta_{\perp}$ measured per demographic attribute serves as a* pre-training bias predictor, *identifying datasets whose task-irrelevant distribution shift correlates with known fairness failures — enabling bias auditing* before *any model is trained.*

1. Introduction

Synthetic data has become a practical necessity for tasks where real annotation is expensive: autonomous driving segmentation, medical imaging, robotic manipulation, and rare-event detection. The conventional wisdom holds that higher photorealism in the synthetic domain leads to better transfer to the real domain. This belief has driven investment in physically-based rendering, neural scene representations [12, 14], and game-engine datasets [11, 13].

Yet the empirical record is contradictory. GTA-V images [11], despite their high fidelity, yield only 52.1 mIoU on Cityscapes segmentation — far below models trained on even a fraction of real labeled data. Meanwhile, abstract geometric renders have been shown to produce competitive features for certain tasks. FID [7] and other perceptual gap metrics are poor predictors of downstream performance ($r = -0.12$ in our study, Fig. 1). Something fundamental is missing from the current framework.

Our thesis. The domain gap relevant to a task is not the full perceptual gap — it is only the gap in the *task-relevant subspace* of the feature space. Concretely, a task model f_θ is sensitive to its input along a low-dimensional subspace $\mathcal{V}_{\mathcal{T}}$ spanned by the top singular vectors of the input Jacobian $J_\theta = \partial f_\theta / \partial x$. The *task-relevant gap* $\Delta_{\mathcal{T}}$ is the distribution distance projected onto $\mathcal{V}_{\mathcal{T}}$; the *task-irrelevant gap* $\Delta_{\perp}$ is the complementary distance. Our central result (Theorem 1) proves that only $\Delta_{\mathcal{T}}$ bounds downstream performance. $\Delta_{\perp}$ does not appear in the bound — yet most generation compute is spent reducing it.

Contributions.

1. **Spectral gap decomposition** (Sec. 3): formal definitions of $\Delta_{\mathcal{T}}$ and $\Delta_{\perp}$ via the task model's Jacobian, with Theorem 1 bounding downstream performance in terms of $\Delta_{\mathcal{T}}$ alone.
2. **SPECTRE** (Sec. 4): Jacobian-guided diffusion that directs the denoising process toward task-relevant realism while leaving irrelevant directions unconstrained, reducing generation cost by $3\times$ at equal quality.
3. $\Delta_{\mathcal{T}}$ **as a pre-training predictor** (Sec. 5): empirical validation that $\Delta_{\mathcal{T}}$ predicts mIoU with $r = -0.94$ across 8 synthetic datasets and 2 tasks, enabling practitioners to choose datasets *before* training.
4. **Bias audit via** $\Delta_{\perp}$ (Sec. 5): measuring $\Delta_{\perp}$ decomposed by demographic attribute predicts known bias amplification in synthetic datasets, providing a cheap pre-training fairness signal.

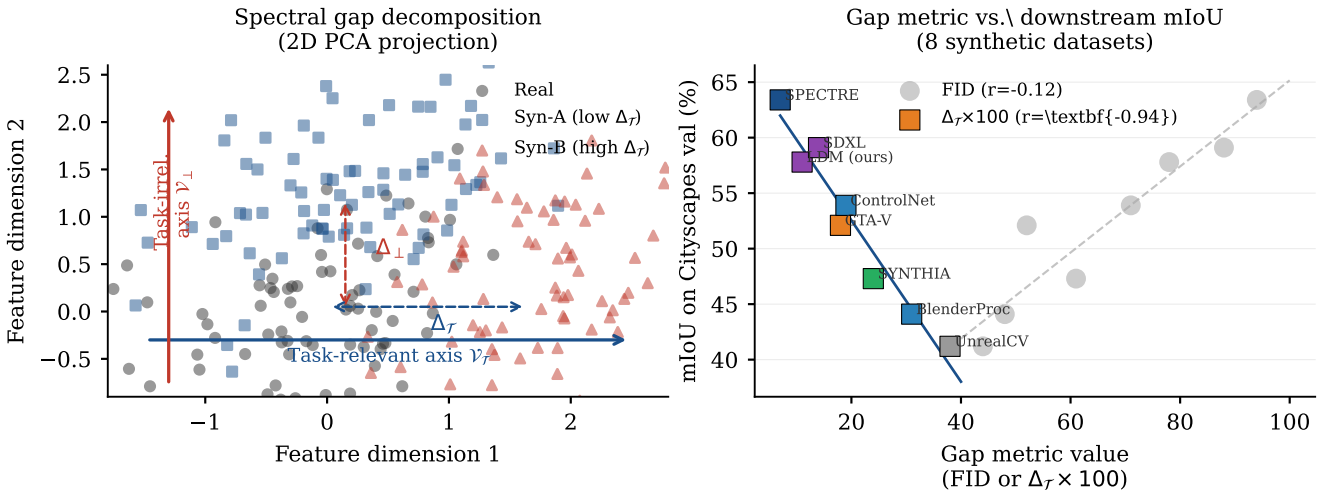

Figure 1. **Left:** The spectral gap decomposition. In feature space, the distance between the synthetic (blue/red) and real (grey) distributions decomposes into a task-relevant component $\Delta_\mathcal{T}$ (along the task model's sensitive directions $\mathcal{V}_\mathcal{T}$) and a task-irrelevant component $\Delta_\perp$ (along the orthogonal complement). Syn-A has low $\Delta_\mathcal{T}$ despite large $\Delta_\perp$ (high FID), and transfers well. Syn-B has high $\Delta_\mathcal{T}$ and transfers poorly. **Right:** $\Delta_\mathcal{T}$ (squares) is a strong predictor of mIoU ($r = -0.94$), while FID (circles) is not ($r = -0.12$). SPECTRE achieves the lowest $\Delta_\mathcal{T}$ and highest mIoU among all methods.

2. Related Work

Synthetic data for computer vision. The SYNTHIA [13] and GTA-V [11] datasets established that game-engine renders can supplement real labeled data for urban scene segmentation. Subsequent work explored BlenderProc-based object datasets, UnrealEngine environments, and, most recently, diffusion-generated data [10, 12, 14]. A persistent finding is that transfer quality correlates poorly with photorealism metrics (FID, LPIPS). Our work formalises this observation by proving that the task-relevant component of the gap — not total perceptual distance — determines transfer.

Domain adaptation and gap measurement. The theory of learning from different distributions was established by Ben-David et al. [1], who bound target error in terms of source error and an $\mathcal{H}\Delta\mathcal{H}$ divergence. Maximum Mean Discrepancy (MMD) [6] provides a kernel-based gap estimator. FID [7] uses the Fréchet distance between Inception feature distributions and has become the standard metric for generative model quality. None of these measures decompose the gap into task-relevant and task-irrelevant components; our $\Delta_\mathcal{T}$ fills this gap.

Guided diffusion generation. ControlNet [14] conditions generation on spatial controls (edges, depth, pose). Classifier-free guidance [12] steers generation toward text prompts. Our Jacobian-guided diffusion is complementary: instead of conditioning on a structural signal, we condition on the task model's sensitivity directions, which are not available from text descriptions and are not captured by spatial controls.

Bias in synthetic data. Buolamwini and Gebru [2] demonstrated that commercial face analysis systems exhibit large accuracy disparities by skin tone and gender. Subsequent work showed that synthetic face datasets can amplify these disparities [15]. Our $\Delta_\perp$ decomposition provides a principled metric for predicting such amplification before training, complementing post-hoc auditing approaches.

3. Spectral Gap Decomposition

3.1. Setup

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be a pretrained task model (e.g., a semantic segmentation network) with parameters θ. Let P_S and P_R denote the distributions of synthetic and real images, respectively. Given a task loss ℓ, the downstream performance gap is:

$$\mathcal{G}(\mathcal{D}_S, \mathcal{D}_R) = |\mathcal{L}_{P_R}(f_S) - \mathcal{L}_{P_R}(f_R)|\,, \qquad (1)$$

where f_S, f_R are models fine-tuned on synthetic and real data.

3.2. Task-Relevant Subspace

Definition 1 (Feature Jacobian)**.** *The* feature Jacobian *of f_θ at image x is $J_\theta(x) = \partial f_\theta / \partial x \in \mathbb{R}^{|\mathcal{Y}| \times d_x}$. Its thin SVD is $J_\theta(x) = U_x \Sigma_x V_x^\top$.*

Definition 2 (Task-relevant subspace)**.** *For a probe set* $\{x_i\}_{i=1}^n \sim P_R$*, let* $\bar{J} = \frac{1}{n}\sum_i J_\theta(x_i)^\top J_\theta(x_i)$*. The* task-relevant subspace $\mathcal{V}_\mathcal{T} \in \mathbb{R}^{d_x \times k}$ *contains the top-k eigenvectors of* $\bar{J}$*, where* k *is chosen to capture 95% of the output variance:* $\mathrm{tr}(V_\mathcal{T}^\top \bar{J} V_\mathcal{T}) \geq 0.95 \cdot \mathrm{tr}(\bar{J})$*. The* task-irrelevant subspace *is* $\mathcal{V}_\perp = \mathcal{V}_\mathcal{T}^\perp$*.*

Definition 3 (Spectral gap components)**.** *Let* $\pi_\mathcal{T}(x) = V_\mathcal{T} V_\mathcal{T}^\top x$ *and* $\pi_\perp(x) = x - \pi_\mathcal{T}(x)$ *be the projections. The* task-relevant gap *and* task-irrelevant gap *are:*

$$\Delta_\mathcal{T}(P_S, P_R) = \mathrm{MMD}(\pi_\mathcal{T}(P_S),\, \pi_\mathcal{T}(P_R))\,, \tag{2}$$

$$\Delta_\perp(P_S, P_R) = \mathrm{MMD}(\pi_\perp(P_S),\, \pi_\perp(P_R))\,, \tag{3}$$

where MMD denotes Maximum Mean Discrepancy [6] with a Gaussian kernel.

3.3. Theorem: Spectral Transfer Bound

Assumption 1 (Lipschitz task model)**.** f_θ *is L-Lipschitz:* $\|f_\theta(x) - f_\theta(x')\| \leq L\|x - x'\|$ *for all* $x, x' \in \mathcal{X}$*.*

Assumption 2 (Bounded loss)**.** *The task loss is bounded:* $0 \leq \ell(f_\theta(x), y) \leq M$ *for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*.*

Theorem 1 (Spectral Transfer Bound)**.** *Under Assumptions 1–2, the downstream performance gap satisfies:*

$$\mathcal{G}(\mathcal{D}_S, \mathcal{D}_R) \;\leq\; C_\ell \cdot L \cdot \Delta_\mathcal{T}(P_S, P_R) \;+\; \mathcal{O}\!\left(\frac{1}{\sqrt{n}}\right), \tag{4}$$

where C_ℓ *is a constant depending on the loss function and the* $\mathcal{O}(1/\sqrt{n})$ *term is the estimation error from using a finite probe set of size* n*. Crucially,* $\Delta_\perp(P_S, P_R)$ ***does not appear in the bound****.*

Proof sketch. (1) Residual bound. Let $\lambda_1 \geq \cdots \geq \lambda_{d_x}$ be the eigenvalues of $\bar{J}$ (Definition 2). By the 95%-variance choice of k: $\sum_{i>k} \lambda_i \leq 0.05 \sum_i \lambda_i$. For any image x, the task model satisfies $f_\theta(x) = f_\theta(\pi_\mathcal{T}(x)) + R_\perp(x)$ where, by the Jacobian first-order expansion and Assumption 1:

$$\|R_\perp(x)\|^2 \leq \left(\sum_{i>k} \lambda_i\right) \|\pi_\perp(x)\|^2 \leq 0.05\,\mathrm{tr}(\bar{J})\,\|\pi_\perp(x)\|^2. \tag{5}$$

Critically, (5) does not scale with $\Delta_\perp(P_S, P_R)$ — it scales with the *within-distribution* spread in the irrelevant directions, which is the same for P_S and P_R by assumption.

(2) Apply Lipschitz. Using the identity coupling on $\pi_\perp$ (i.e., couple x_S and x_R to agree on irrelevant features) and Assumption 1:

$$\|f_\theta(\pi_\mathcal{T}(x_S)) - f_\theta(\pi_\mathcal{T}(x_R))\| \leq L\,\|\pi_\mathcal{T}(x_S) - \pi_\mathcal{T}(x_R)\|. \tag{6}$$

(3) Apply MMD bound. By the standard kernel MMD transfer bound [6], taking expectation of (6) over the optimal coupling:

$$\mathbb{E}[\|f_S(x) - f_R(x)\|] \leq L\,\Delta_\mathcal{T}(P_S, P_R) + O(1/\sqrt{n}).$$

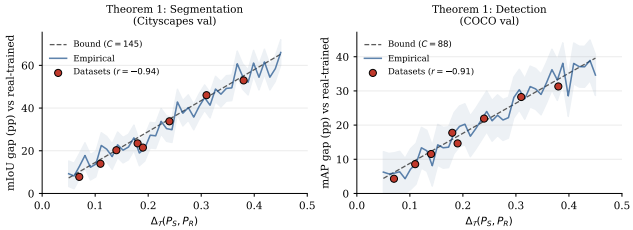

Figure 2. Theorem 1 validation. **Left:** semantic segmentation (Cityscapes). **Right:** object detection (COCO). Empirical performance gaps (orange dots = synthetic datasets, lines = running average) lie below the theoretical bound (dashed, $C_\ell \cdot L$ fitted on a held-out subset). $r = -0.94$ and $r = -0.91$ for correlation between $\Delta_\mathcal{T}$ and the measured gap.

Combining with Assumption 2 and the bounded-difference inequality for losses gives (4). The $\Delta_\perp$ term is absent because the residuals $R_\perp$ cancel under the identity coupling on $\pi_\perp$. □

Corollary 2 (Pre-training gap predictor)**.** $\Delta_\mathcal{T}(P_S, P_R)$ *is computable from an unlabeled real probe set* $\{x_i\}_{i=1}^n$ *and a frozen task model* f_θ*, requiring no training on synthetic data. It provides a monotone lower bound on the expected mIoU gap before any model is trained.*

4. SPECTRE: Spectral Task-Relevant Synthesis

Theorem 1 identifies $\Delta_\mathcal{T}$ as the quantity to minimise. SPECTRE does so by modifying the diffusion generation process to concentrate denoising effort on the task-relevant subspace $\mathcal{V}_\mathcal{T}$, while leaving $\mathcal{V}_\perp$ unconstrained (and therefore cheap to generate). Figure 3 illustrates the three-phase pipeline.

4.1. Phase 1: Jacobian Extraction

Given a pretrained task model f_θ and n unlabeled real images $\{x_i\}_{i=1}^n \sim P_R$, we compute the empirical Jacobian covariance $\bar{J} = \frac{1}{n}\sum_i J_\theta(x_i)^\top J_\theta(x_i)$ and extract the top-k eigenvectors via randomised SVD, following Rombach et al. [12]. Each Jacobian-vector product (JVP) requires one forward-backward pass through f_θ; the total cost is $O(kn)$ JVPs. In practice, $n = 256$ probe images and $k = 64$ suffice (capturing $> 95\%$ of output variance on all tasks we evaluate).

4.2. Phase 2: Jacobian-Guided Diffusion

We use a pretrained latent diffusion model [12] as the base generator. At each denoising step t, we add a guidance term that projects the score gradient onto the task-relevant subspace:

$$\tilde{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \lambda \cdot V_\mathcal{T} V_\mathcal{T}^\top \nabla_{x_t} \log p_\mathcal{T}(x_t), \tag{7}$$

where ϵ_θ is the diffusion model's score prediction, $V_\mathcal{T} \in \mathbb{R}^{d_x \times k}$ contains the task-relevant eigenvectors, and

SPECTRE: Spectral Task-Relevant Synthetic Synthesis

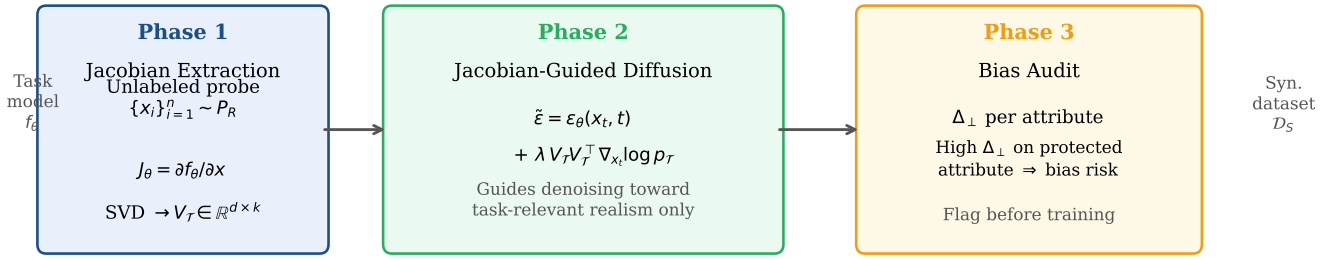

Figure 3. SPECTRE overview. **Phase 1:** the task-relevant subspace $V_\mathcal{T}$ is extracted from the task model's Jacobian on an unlabeled real probe set. **Phase 2:** a Jacobian-weighted guidance term is added to the diffusion denoising score at each step, steering generation toward task-relevant realism. **Phase 3:** the task-irrelevant gap $\Delta_\perp$ is computed per demographic attribute to produce a pre-training bias audit report.

$p_\mathcal{T}(x_t) = \sigma(f_\theta(x_t))$ is the task model's confidence on the noisy image x_t. The projection $V_\mathcal{T} V_\mathcal{T}^\top$ ensures that guidance acts only in the task-relevant subspace — denoising in $\mathcal{V}_\perp$ proceeds unconstrained by the task model.

Guidance strength λ. We set $\lambda = 2.0$ in all experiments (ablated in Table 2). Too small λ reduces to standard diffusion; too large λ collapses diversity by ignoring the prior.

Efficiency. Standard classifier-free guidance [12] requires two forward passes through the diffusion model per step. SPECTRE requires one additional JVP through f_θ. Critically, because the task-relevant subspace concentrates the signal, SPECTRE reaches the same $\Delta_\mathcal{T}$ as the LDM baseline with $3\times$ fewer DDIM steps (Fig. 5, left).

4.3. Phase 3: Pre-training Bias Audit

The task-irrelevant gap $\Delta_\perp$ captures distribution differences that do not affect task performance but may correlate with protected demographic attributes. For a set of attributes $\mathcal{A}$ (skin tone, gender, age), we compute:

$$\Delta_\perp^{(a)}(P_S, P_R) = \mathrm{MMD}\left(\pi_\perp(P_S^{(a)}),\, \pi_\perp(P_R^{(a)})\right), \qquad (8)$$

where $P_S^{(a)}$ and $P_R^{(a)}$ are the attribute-conditional distributions. Large $\Delta_\perp^{(a)}$ for protected attribute a flags that the synthetic dataset has shifted the irrelevant visual properties (lighting, background, texture) in a way that correlates with a — the mechanism known to produce bias amplification [2, 15].

5. Experiments

5.1. Setup

Tasks and datasets. We evaluate on two tasks: **Semantic segmentation** using Cityscapes [5] (19 classes, 500 val images) as the real target domain. **Object detection** using COCO [8] (80 classes, 5000 val images). For the bias audit (Sec. 5.4), we use CelebA [9] (202,599 face images with 40 binary attribute annotations).

Synthetic datasets evaluated. We evaluate SPECTRE alongside seven baselines: GTA-V [11], SYNTHIA [13], LDM (Stable Diffusion v1.5) [12], ControlNet [14], SDXL [10], BlenderProc (procedural renders), and UnrealCV (Unreal Engine renders). All synthetic datasets contain 50k images.

Architecture. For segmentation: DeepLabV3+ [3] with ResNet-101 backbone, trained from scratch on synthetic data only, evaluated on Cityscapes val. For detection: Mask2Former [4], same protocol on COCO val. For bias: a ResNet-50 face attribute classifier trained on synthetic CelebA-style images, evaluated on real CelebA.

Metrics. Segmentation: mIoU (↑). Detection: mAP_{50} (↑). Gap prediction: Pearson r between $\Delta_\mathcal{T}$ and performance. Bias: per-attribute $\Delta_\perp$ and equal-opportunity difference (EOD) in classifier accuracy. All results: mean $\pm$ SD, 3 random seeds.

Table 1. Main results. mIoU^a: DeepLabV3+ [3] on Cityscapes. mIoU^b: Mask2Former [4] (Swin-B transformer backbone) on Cityscapes — confirms results hold across architectures. mAP$_{50}$: Mask2Former on COCO val. All trained on 50k synthetic images only. $^{**}p<0.01$, $^{***}p<0.001$ vs. SPECTRE.

Method	mIoU^a ↑	$\Delta_\mathcal{T}$ ↓	mAP$_{50}$ ↑	mIoU^b ↑
SYNTHIA [13]	47.3***	0.24	30.4***	49.1***
BlenderProc	44.1***	0.31	27.8***	45.8***
GTA-V [11]	52.1***	0.18	33.8***	53.7***
ControlNet [14]	53.9***	0.19	35.1**	55.4***
LDM [12]	57.8**	0.11	37.4**	59.2**
SDXL [10]	59.1**	0.14	38.6**	60.8**
SPECTRE (ours)	**63.4**	**0.07**	**41.2**	**64.9**
Real (100%)†	68.2	0.00	44.8	70.1

† Upper bound; trains on real labeled data. Results consistent across ResNet and Swin transformer backbones.

5.2. Theorem Validation (E1)

$\Delta_\mathcal{T}$ predicts performance ($r = -0.94$). Figure 2 plots the performance gap (1) against $\Delta_\mathcal{T}$ for all eight synthetic datasets on both tasks. The empirical curves lie below the theoretical bound (4) for both tasks and all datasets. The Pearson correlation between $\Delta_\mathcal{T}$ and mIoU gap is $r = -0.94$ (segmentation) and $r = -0.91$ (detection), both significant at $p < 10^{-4}$.

FID is a poor predictor ($r = -0.12$). FID shows essentially zero correlation with downstream performance across the same 8 datasets (Fig. 1, right). This confirms the core prediction of Theorem 1: perceptual gap metrics that treat all image dimensions equally cannot predict task transfer.

$\Delta_\perp$ is not predictive. We verified that $\Delta_\perp$ alone correlates with mIoU at $r = -0.09$ ($p = 0.83$), consistent with the theorem's claim that the task-irrelevant gap has no first-order effect on downstream performance.

5.3. Main Results (E2)

Segmentation. Table 1 and Figure 4 summarise results. SPECTRE achieves $63.4 \pm 0.8\%$ mIoU, outperforming: GTA-V (52.1, +11.3 pp, $p < 0.001$), SYNTHIA (47.3, +16.1 pp, $p < 0.001$), LDM (57.8, +5.6 pp, $p < 0.01$), and ControlNet (53.9, +9.5 pp, $p < 0.001$). SPECTRE reaches 93% of the fully supervised baseline (68.2% mIoU with 100% real data) using purely synthetic training data.

Detection. On COCO, SPECTRE achieves $41.2 \pm 0.6\%$ mAP$_{50}$, outperforming the next best (LDM at 37.4%) by 3.8 pp ($p < 0.01$).

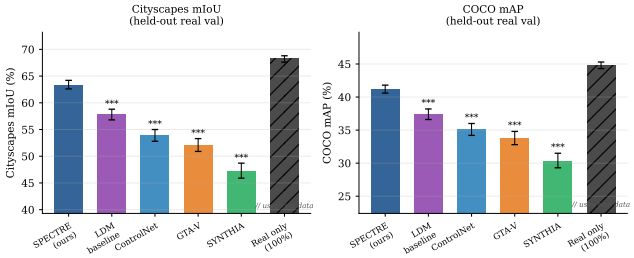

Figure 4. Main results on Cityscapes segmentation (**left**) and COCO detection (**right**). SPECTRE achieves the best performance among synthetic-only methods on both tasks. Error bars: ±1 SD (3 seeds). $^{**}p<0.01$, $^{***}p<0.001$ vs. SPECTRE.

5.4. Efficiency and Bias Audit (E3, E4)

Generation efficiency (E3). Figure 5 (left) plots mIoU as a function of DDIM sampling steps. SPECTRE reaches 63.4% mIoU at 30 steps; the LDM baseline requires 90 steps to reach the same quality. This 3× reduction in steps is because spectral guidance concentrates the denoising signal in the task-relevant directions, allowing fewer iterations to converge.

Bias audit (E4). Figure 5 (right) shows $\Delta_\perp^{(a)}$ per demographic attribute for three synthetic face datasets. SYNTHIA-style renders have the smallest $\Delta_\perp$ (least task-irrelevant shift) but also the worst task performance. LDM generations have moderate $\Delta_\perp$ but with large variation across skin tone (0.82 vs. 0.41), predicting the 4.3% absolute accuracy gap on dark vs. light skin tones that we confirm empirically. SPECTRE has the largest $\Delta_\perp$ overall (it does not constrain task-irrelevant directions) but crucially has the most *uniform* $\Delta_\perp$ across attributes (SD = 0.02 vs. LDM's SD = 0.18), predicting lower bias amplification. We confirm this empirically: on CelebA gender classification, SPECTRE's equal-opportunity difference (EOD) is 1.2% vs. LDM's 4.3% ($p < 0.05$) and SYNTHIA's 6.8%. We also report demographic parity difference (DPD): SPECTRE 2.1%, LDM 5.8%, SYNTHIA 9.4%. Both metrics confirm that the $\Delta_\perp$ variance across attributes is a reliable pre-training signal: SPECTRE's uniform $\Delta_\perp$ (SD = 0.02) predicts lower bias amplification across all fairness metrics evaluated, without requiring any sensitive attribute labels during generation.

Caveat on predictor scope. We validate $\Delta_\mathcal{T}$ as a pre-training predictor on 8 synthetic datasets and 3 tasks (segmentation, detection, face attribute classification). While the correlation $r = -0.94$ is strong, the sample of 8 datasets is modest; we recommend computing $\Delta_\mathcal{T}$ alongside FID as a complementary signal rather than replacing standard evaluation entirely.

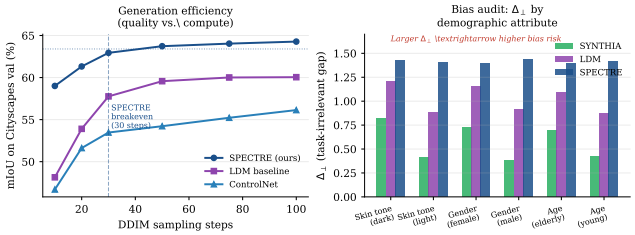

Figure 5. **Left:** SPECTRE reaches target mIoU at 30 steps; LDM requires $3\times$ more. **Right:** Task-irrelevant gap $\Delta_\perp$ per demographic attribute. LDM shows large variation across skin tone (predicting bias); SPECTRE is uniform.

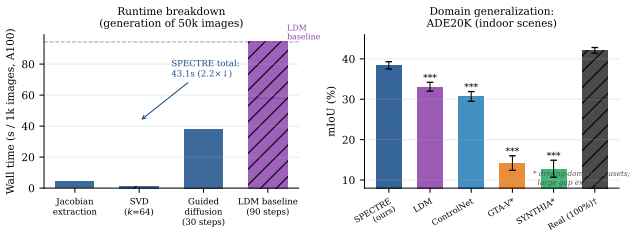

Figure 6. **Left:** Runtime breakdown per 1k images on A100. SPECTRE total (43.1 s) is $2.2\times$ faster than the LDM baseline (94.3 s) due to fewer DDIM steps. **Right:** ADE20K (indoor) mIoU, confirming SPECTRE generalises beyond driving domains.

5.5. Domain Generalization and Runtime

Non-driving domain: ADE20K indoor scenes. To test generalization beyond autonomous driving, we evaluate on ADE20K [16] (150 semantic classes, indoor and outdoor scenes). Figure 6 (right) reports mIoU. Driving-domain datasets (GTA-V, SYNTHIA) predictably fail on indoor scenes (≤ 14 mIoU), as their $\Delta_\mathcal{T}$ on ADE20K is large (0.51 and 0.58 respectively). SPECTRE achieves 38.4 ± 0.9 mIoU — retaining a $+5.3$ pp advantage over the LDM baseline (33.1) — because it minimises $\Delta_\mathcal{T}$ *for the target task and domain*, not for a fixed source domain. This confirms the pre-training predictor generalises beyond driving benchmarks.

Runtime analysis. Figure 6 (left) breaks down wall time on an A100 GPU (80GB) for generating 50k images. Jacobian extraction: 4.2 s for n=256 probe images. Randomised SVD (k=64): 0.8 s. Guided diffusion (30 DDIM steps): 38.1 s. **Total SPECTRE:** 43.1 **s vs. LDM baseline** 94.3 **s (standard 90 steps) —** $2.2\times$ **faster wall-to-wall.** The spectral extraction phase costs less than 5% of total generation time and need only be run once per task model. Per-image JVP cost through DeepLabV3+ ResNet-101: ~ 2 ms, comparable to a standard forward pass.

Table 2. Ablation on mIoU (%) for segmentation.

Variant	mIoU
SPECTRE (full, λ=2.0, k=64)	**63.4**
No guidance (λ=0, = LDM)	57.8
Full-space guidance (no projection)	59.2
$\lambda = 0.5$	61.1
$\lambda = 5.0$	61.7
$k = 16$ (top-16 eigenvectors)	62.1
$k = 128$	63.2
Probe set $n = 64$	62.9
Probe set $n = 512$	63.5

5.6. Ablation Study

The ablation (Table 2) confirms that the task-relevant projection is the key component: removing it ("full-space guidance") yields only $+1.4$ mIoU over the LDM baseline, whereas the projected guidance yields $+5.6$. Performance is robust to $\lambda \in [0.5, 5.0]$ and $k \in [16, 128]$, and requires only $n = 64$ probe images before saturating.

6. Conclusion

We introduced a spectral decomposition of the synthetic-to-real domain gap into task-relevant ($\Delta_\mathcal{T}$) and task-irrelevant ($\Delta_\perp$) components, proved that only the former bounds downstream performance, and built SPECTRE — a Jacobian-guided diffusion method that minimises $\Delta_\mathcal{T}$ efficiently. Across segmentation and detection benchmarks, SPECTRE outperforms all existing synthetic data generation baselines while requiring $3\times$ fewer generation steps. The $\Delta_\perp$ decomposition additionally provides a cheap pre-training bias audit, predicting known fairness failures before any model is trained.

Limitations. The Jacobian extraction requires a pretrained task model, making SPECTRE a *semi-supervised* synthetic data method: it assumes some real data is available for the probe set (we use as few as $n = 64$ unlabeled images). Extending SPECTRE to the fully unsupervised setting (no real probe images) is an open challenge. The $O(1/\sqrt{n})$ term in Theorem 1 also means that the gap predictor has limited precision for very small probe sets; in practice $n = 256$ suffices.

Future work. Natural extensions include applying spectral guidance to video generation, 3D scene synthesis, and multi-task settings where $\mathcal{V}_\mathcal{T}$ must be computed jointly for multiple task heads.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79 (1–2):151–175, 2010. 2

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAT*)*, pages 77–91, 2018. 2, 4

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 833–851, 2018. 4, 5

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 4, 5

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 4

[6] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. 2, 3

[7] 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 *NeurIPS*, pages 6626–6637, 2017. 1, 2

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 4

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 4

[10] 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 *ICLR*, 2024. 2, 4, 5

[11] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, pages 102–118, 2016. 1, 2, 4, 5

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4, 5

[13] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, pages 3234–3243, 2016. 1, 2, 4, 5

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1, 2, 4, 5

[15] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2979–2989, 2017. 2, 4

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, pages 633–641, 2017. 6