EFFICIENT BLACK-BOX FINGERPRINTING FOR IMAGE GENERATION MODELS WITH AMBIGUOUS IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Black-box attribution of image generation models is difficult because fine-tuning and other modifications can substantially alter their outputs. Text-only fingerprints are fragile because semantic preferences become entangled with unconstrained variation in style, layout, and composition. We present **AmbiPrint**, a passive framework that probes how models resolve controlled semantic ambiguity. AmbiPrint pairs a neutral text prompt with an ambiguous image that guides generation and constrains visual variation without determining the final class. Across 25 open- and closed-source models, AmbiPrint separates source derivatives from independently trained models under fine-tuning, compression, and substantial style and domain shifts. It succeeds with only 100 generations per model where text-only baselines fail. Adaptive attacks can evade attribution by targeting every evaluated category, but doing so substantially reduces output diversity. AmbiPrint provides a practical and robust signal for black-box model attribution.

1 INTRODUCTION

Training high-quality image generation models is computationally expensive (Wang & Du, 2025; Croitoru et al., 2023; Bertazzini et al., 2025), yet released models can be copied, modified, and monetized with little friction. A model owner may therefore need to determine whether a black-box service is derived from a protected model without access to its weights, activations, or generation trajectory. Watermarking embeds a signature before release (Zhao et al., 2023; Zhang et al., 2020; Wen et al., 2023), whereas fingerprinting uses intrinsic model signatures (Nie et al., 2023; Kim et al., 2024; Wen et al., 2023; Teng et al., 2025). Watermarks cannot protect unmarked models, and many fingerprints require white-box access.

CSF (Lee et al., 2026) recently showed that underspecified text prompts can expose semantic preferences through black-box queries. However, text-only ambiguity also leaves object count, layout, viewpoint, background, texture, and style unconstrained. Fine-tuning can alter these visual factors while preserving the source semantic prior, thereby shifting the measured output distribution for reasons unrelated to model origin. More queries reduce sampling noise but cannot remove this systematic shift; commercial APIs also make large query budgets costly.

We present **AmbiPrint**, which preserves semantic ambiguity while controlling the context in which it is resolved. Each probe pairs a neutral text prompt with an image formed by mixing multiple semantic classes. The image provides *semantic guidance* toward the target super-category and a *semantic constraint* on spatial structure, while remaining ambiguous about the final class. Models must therefore resolve the same controlled conflict according to their learned priors, improving invariance to downstream modification without sacrificing separation between independent families.

Across 25 open- and closed-source models spanning Stable Diffusion, Flux, DALL-E, and Nano-Banana, AmbiPrint separates source families under fine-tuning, quantization, pruning, severe style and domain shifts, classifier changes, and inference-parameter variation. It succeeds with as few as 100 generations per model where text-only fingerprints fail. Our contributions are:

- a passive, query-only framework for source-family attribution using controlled ambiguous image-text probes;
- an analysis of why constraining nuisance visual variation stabilizes semantic fingerprints;

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.

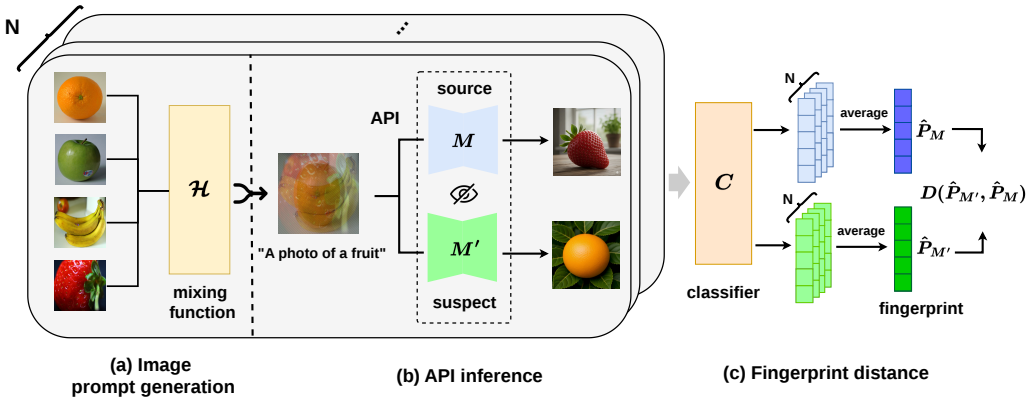


Figure 1: Overview of AmbiPrint. We construct ambiguous images by mixing source images from multiple semantic classes, pair them with neutral text prompts, query the source and suspect models through their black-box APIs, and aggregate the generated outputs in a semantic space to measure the distance between their AmbiPrints.

- a 25-model evaluation showing reliable attribution with only 100 generations per model; and
- adaptive-attack experiments that characterize both robustness and a concrete failure mode.

2 Related Work

Watermarking. Watermarking actively embeds a detectable signal during training (Li et al., 2021) or inference (Kirchenbauer et al., 2023). It is useful when present, but cannot retrospectively protect an unmarked model; it may also affect generation quality or be weakened by removal and adaptation (Zhao et al., 2023). AmbiPrint instead measures behavior already present in the model.

White-Box Fingerprinting. Passive fingerprints can use internal LLM representations (Zhang et al., 2024) or diffusion denoising trajectories (Teng et al., 2025). These signals are discriminative but require parameters, activations, gradients, or latent states unavailable through a black-box API.

Black-Box Fingerprinting. For LLMs, adversarial or ambiguous prompts reveal model-specific confidence and preferences (Maus et al., 2023; Bai et al., 2024); for image generation, output statistics support closed-set model classification (Stalder et al., 2022). Closed-set identity, however, does not establish whether a modified suspect shares a source with a protected model. CSF (Lee et al., 2026) addresses source attribution with compositional text prompts, but leaves structure and style uncontrolled. AmbiPrint adds an ambiguous image to constrain these nuisance factors and tests whether the resulting signature survives downstream modifications and adaptive attacks.

3 Method

AmbiPrint attributes image generation models by measuring how they resolve controlled semantic ambiguity. Figure 1 summarizes probe construction, black-box querying, and source-to-suspect fingerprint comparison.

3.1 Problem Formulation

Threat Model. The defender owns a protected model M; an adversary applies a transformation T (e.g., fine-tuning, quantization, or pruning) and deploys $M' = T(M)$ as a black-box service. A source and its derivatives form a *model family*, while models from different sources are independent families. The defender can submit prompts and observe images, but cannot access weights, activations, gradients, training data, or generation trajectories.

Attribution Objective. Given black-box probes $\mathcal{C} = \{c_i\}_{i=1}^{N}$, we summarize each model's outputs by an empirical fingerprint $\hat{P}$. The suspect is attributed to the source family when

$$D(\hat{P}_{M'}, \hat{P}_M) \leq \tau. \tag{1}$$

Requirements for Robust Attribution. A useful fingerprint must be **invariant** to downstream modification,

$$D(\hat{P}_{T(M)}, \hat{P}_M) \leq \tau. \tag{2}$$

yet **discriminative** between independently trained models,

$$D(\hat{P}_{M_{\text{ind}}}, \hat{P}_M) > \tau. \tag{3}$$

Surface statistics such as texture, color, and style are poor candidates: fine-tuning changes rendering style, while compression perturbs local statistics (Song et al., 2024). Semantic priors learned during pre-training are more persistent because adaptation often reuses existing knowledge (Liu et al., 2026), changing foundational structure can cause forgetting or degradation (Cheng et al., 2025), and targeted unlearning can leave core priors intact (Li et al., 2025). We therefore fingerprint high-level semantic preferences rather than appearance.

Why Text-Only Probes Are Insufficient. An ambiguous text prompt exposes semantic preferences but does not isolate them: the model also chooses object count, layout, viewpoint, background, texture, and style. A modification can shift these factors without changing the source semantic prior. Because the classifier observes rendered images rather than latent decisions, more samples only estimate this shifted distribution more precisely.

3.2 Ambiguous Image-Text Probes

AmbiPrint pairs a neutral super-category prompt with an ambiguous image. Local evidence from every candidate class supplies *semantic guidance* toward the target super-category, while image-to-image generation *constrains* layout, scale, viewpoint, and object placement. Because evidence is distributed across candidates, the image supplies no single correct class. Source and suspect models thus resolve the same constrained semantic context according to their learned priors.

Formally, each probe is defined as

$$c^* = (c^*_{\text{text}}, c^*_{\text{image}}), \tag{4}$$

where c^*_{text} may be "a photo of a fruit" and c^*_{image} combines classes within that super-category. The text defines valid outcomes; the image constrains their realization while leaving the class unresolved.

Pixel-wise Mixing Function ($\mathcal{H}$). We sample K source images $\{x_1, \ldots, x_K\}$ from distinct classes in one super-category and combine them with a pixel-wise function $\mathcal{H}$:

$$c^*_{\text{image}}[u, v] = \mathcal{H}(x_1[u, v], \ldots, x_K[u, v]), \tag{5}$$

This preserves local evidence from all candidates without producing a clean member of any class, while providing a common spatial scaffold for generation.

We test averaging and random, sequential, and spiral discrete mixing. All variants work reliably; spiral mixing gives the strongest separation and is our default (Appendix C).

3.3 Fingerprint Extraction and Attribution

For each model M, we construct N probes $\mathcal{C}^* = \{c^*_1, \ldots, c^*_N\}$ by repeatedly sampling source images and applying $\mathcal{H}$, then generate

$$y_i = M(c^*_i), \qquad i = 1, \ldots, N. \tag{6}$$

For a pretrained classifier or encoder C, the classifier-based fingerprint is the mean softmax response

$$\hat{P}_M = \frac{1}{N} \sum_{i=1}^{N} \sigma(C(y_i)), \tag{7}$$

where σ is softmax. For an embedding model, we instead average features and use an embedding-space distance such as cosine distance.

Attribution Test. Using the same probe distribution, we attribute M' to M when their Jensen–Shannon (JS) divergence satisfies

$$D_{\mathrm{JS}}(\hat{P}_{M'}, \hat{P}_M) \leq \tau. \tag{8}$$

Threshold Calibration. For family-disjoint labeled pairs, let $d_{\max}^{\mathrm{intra}}$ be the largest source–derivative distance and $d_{\min}^{\mathrm{inter}}$ the smallest inter-family distance. When they separate, we set

$$\tau = d_{\max}^{\mathrm{intra}} + \lambda \left(d_{\min}^{\mathrm{inter}} - d_{\max}^{\mathrm{intra}} \right), \qquad 0 < \lambda < 1. \tag{9}$$

We use $\lambda = 0.2$. Each leave-one-family-out split excludes the held-out family from both calibration pairs and the external candidate pool, then applies the threshold unchanged. If no separating threshold exists, we use the calibration-error-minimizing threshold. Sensitivity and the resulting open-set limitation are analyzed in Sec. 4.3.

4 Experiments

4.1 Implementation Details

Baselines. We compare with CSF (Lee et al., 2026), which measures semantic responses to underspecified text, and a text-only AmbiPrint variant. Together they isolate the contribution of image conditioning.

Evaluation Models. Our 25 open- and closed-source models span diffusion (Ho et al., 2020; Saharia et al., 2022), flow matching (Lipman et al., 2023; Liu et al., 2023; Labs et al., 2025), and proprietary APIs (OpenAI, 2023). Five source families have known derivatives; independent models include DALL-E and Gemini Image. The benchmark covers closely related Stable Diffusion (SD) 1.x and 2.x families, artistic adaptation, compression, and a severe medical X-ray domain shift (Appendix B).

Probe Construction and Generation Protocol. We use **Fruits**, **Animals**, **Birds**, and **Insects**, sampling $K = 4$ distinct classes per probe from ImageNet-1K (Deng et al., 2009) and Open Images (Kuznetsova et al., 2020). Unless stated otherwise, spiral mixing and uncontrolled seeds produce $N = 400$ images per model (100 per category), including APIs without seed control. Main results aggregate all four categories.

Fingerprint Computation and Threshold Calibration. Our default projection is an ImageNet-1K-pretrained ResNet-50 (He et al., 2016; Deng et al., 2009); we average its 1K-dimensional probabilities and use JS divergence. ViT (Dosovitskiy et al., 2021) and DINO-v2 (Oquab et al., 2023) are ablated in Sec. 4.5. Strict five-fold leave-one-family-out calibration removes the complete held-out family from both calibration examples and the external candidate pool. The other four source families set the boundaries with $\lambda = 0.2$. Fold thresholds average 0.023 (range 0.019–0.030); DALL-E, Gemini Image, and other independent models without known derivatives are never calibration data.

4.2 Main Results

Figure 2 compares each source or derivative with its source and nearest external family, and each independent model with its nearest evaluated neighbor. Every source–derivative distance is below its fold threshold, whereas every external and independent distance is above the full threshold range. All five held-out-family folds therefore have zero false accepts and zero false rejects at mean $\tau = 0.023$ (range 0.019–0.030).

AmbiPrint separates the closely related but independently trained SD 1.x and 2.x families despite nearly identical architectures and substantially overlapping pretraining data; their principal design difference is the text encoder. It also retains attribution after anime, illustration, comic, and painting adaptation, and even after an SD model is shifted to medical X-ray generation. Every derivative remains closer to its source than to an independent family, indicating a source-specific semantic prior rather than an appearance statistic.

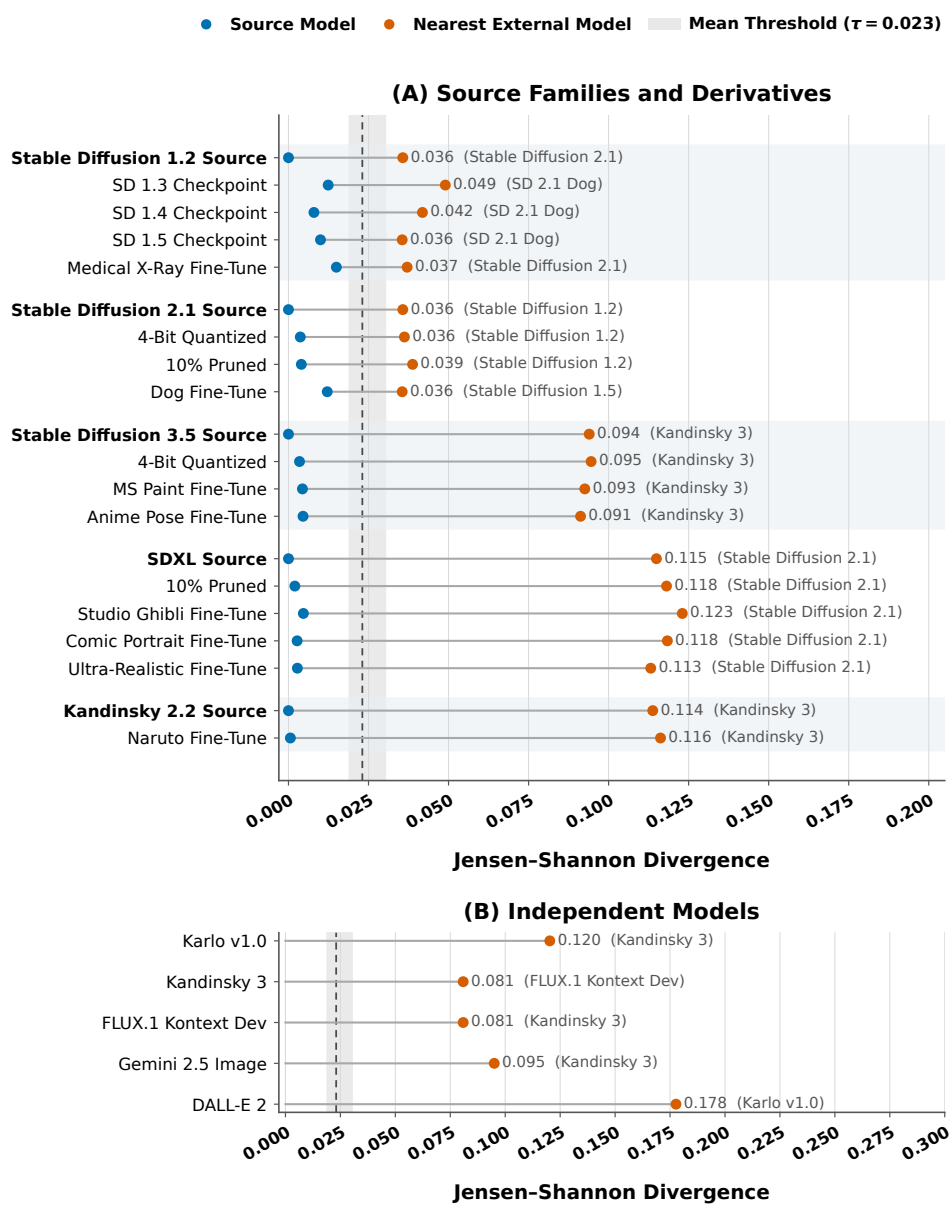


Figure 2: Main attribution results using the mean Jensen-Shannon divergence across fingerprints from fruits, animals, birds, and insects. Lower values indicate greater similarity. In (A), blue points show each model's distance to its source, while orange points show its distance to the nearest model outside its source family. In (B), each independent model is compared with its nearest model among all other evaluated models. The dashed line shows the mean fold threshold $\tau = 0.023$, and the shaded region shows the fold-specific range from 0.019 to 0.030.

Each supercategory independently gives the same error-free outcome with only 100 generations per model; aggregation broadens semantic coverage. Appendix D gives the full category and aggregate matrices.

Model Family Coherence. The pairwise matrices further show that derivatives remain close to one another, not only to the base checkpoint, under style, content, and domain shifts. AmbiPrint thus recovers a coherent family signature rather than a fragile match to one stored checkpoint. Figure 3 gives a qualitative view of the same effect: an identical ambiguous fruit probe elicits related interpretations within the SDXL family and distinct interpretations from independent families.

Table 1: Comparison with text-only baselines under oracle calibration. N denotes generated images per model.

Conditioning	Method	N	FP	FN	Errors
Text	CSF	1260	5	6	11
	AmbiPrint (Text)	100	5	23	28
	AmbiPrint (Text)	400	4	13	17
Text + Image	AmbiPrint	100	0	0	0
	AmbiPrint	400	0	0	0



Figure 3: An identical ambiguous fruit probe produces similar interpretations within the SDXL family and distinct interpretations across independent families.

4.3 Calibration Sensitivity and Open-Set Limitation

We vary $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ under the same strict family holdout. Failures occur only when SD 1.x or SD 2.1 is held out while the other remains in calibration; these families share nearly identical architectures and substantially overlapping pretraining data. The other three folds have no failures at any tested weight, and no fold has false negatives. In the two sensitive folds, $\lambda = 0.1$ and 0.2 give zero false positives; errors first appear at 0.3 and increase as the threshold approaches the inter-family boundary (Figure 4). We therefore use $\lambda = 0.2$, which adds margin above the intra-family boundary without any benchmark error.

This exposes a genuine open-set limitation: a previously unseen independent family can be closer to the protected family than those used for calibration, making the threshold too permissive. Conservative λ reduces the observed risk but cannot guarantee correct attribution. Deployment should use the closest available families—especially those sharing architectures, data, or development history—and update the calibration set as new families emerge.

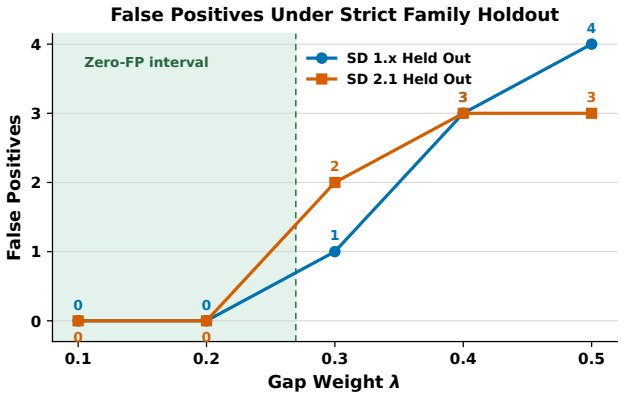


Figure 4: False positives versus gap weight λ under strict family holdout. The SD 1.x and SD 2.1 folds have zero false positives at $\lambda \leq 0.2$; false positives appear from $\lambda = 0.3$. The other three folds have no failures across all evaluated weights, and no fold has false negatives.

Table 2: Probe and fingerprint ablations. Mixing rows use $K = 4$; class-count rows use spiral mixing.

Setting	FP	FN	Err.
Mixing strategy			
Random	0	0	0
Average	0	0	0
Sequential	0	0	0
Spiral	0	0	0
Semantic classes			
$K = 1$	17	0	17
$K = 2$	0	0	0
$K = 4$	0	0	0
$K = 8$	0	0	0

Setting	FP	FN	Err.
Fingerprint distance			
JS divergence	0	0	0
L_1	0	0	0
L_2	0	0	0
Cosine	0	0	0
Semantic projection			
ViT-Base (P=16)	0	0	0
ViT-Base (P=32)	0	0	0
DINO-v2	0	0	0

4.4 Comparison with Text-only Baselines

For an oracle-favorable comparison, both text-only methods select thresholds using the complete pairwise matrix and all family labels. Even then, source–derivative and independent distances overlap, especially after anime, illustration, and medical X-ray adaptation. CSF makes 11 errors with 1260 images per model; text-only AmbiPrint makes 17 with 400. Image-conditioned AmbiPrint makes zero with 100, a 12.6× smaller budget than CSF (Table 1). More text samples cannot repair the systematic loss of invariance.

4.5 Ablation Studies

We ablate probe construction, semantic projection, and fingerprint distance in Table 2; complete matrices are in Appendix D.

Analysis of the Ambiguous Image All mixing functions have zero failures, with spiral giving the strongest separation. Only $K = 1$ fails, confirming that ambiguity is necessary; we default to $K = 4$. All tested distances and projection models also give zero failures. JS is most stable at small sample counts, while DINO-v2 with cosine distance confirms that no specific classifier is required.

4.6 Robustness against Adaptive Adversaries

We give a fully informed adversary the prompts, mixing strategy, categories, and rule, then fine-tune SDXL for 2000 iterations toward one class per targeted category. When only fruits are targeted, that fingerprint crosses the threshold while animals, birds, and insects remain below it (Figure 5); untargeted components still support attribution.

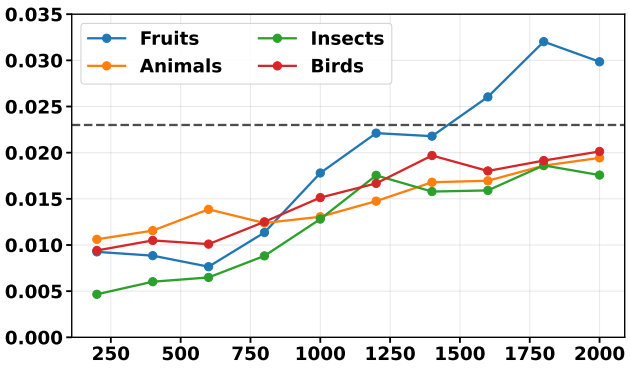


Figure 5: Single-supercategory adaptive fine-tuning on SDXL. Fine-tuning targets fruits for 2000 iterations. The fruit fingerprint progressively diverges from the source and crosses the attribution threshold, while fingerprints from animals, birds, and insects remain below it.

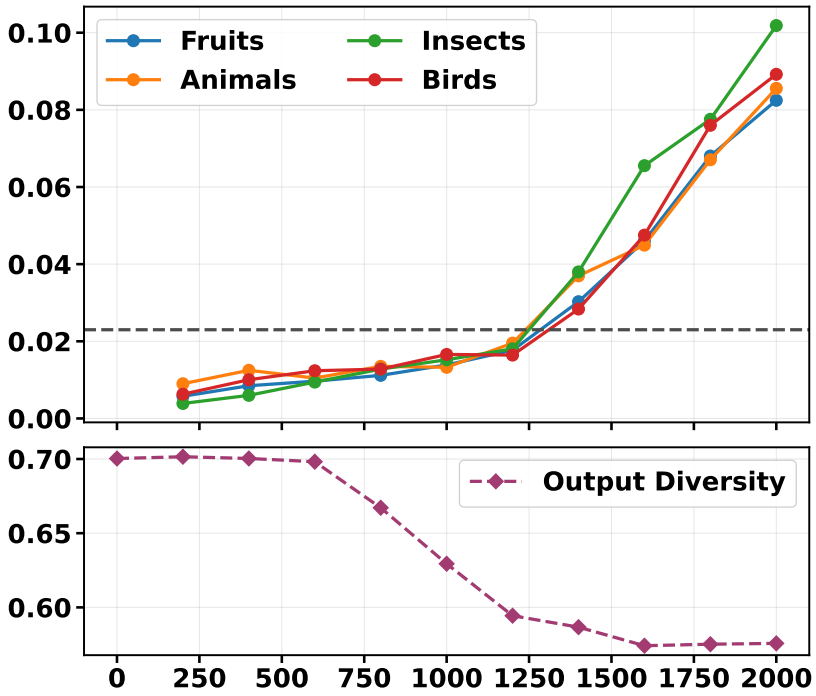


Figure 6: Multi-supercategory adaptive fine-tuning on SDXL. Fine-tuning jointly targets fruits, animals, birds, and insects for 2000 iterations. The fingerprint distance for every supercategory increases beyond the attribution threshold, while output diversity decreases as the fingerprints drift from the source.

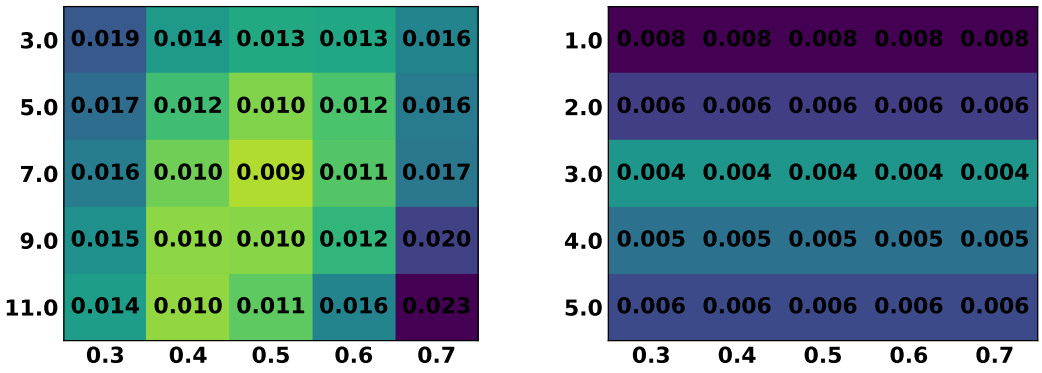
Failure Mode under Multi-Category Fine-tuning When all four categories are targeted jointly, every component eventually crosses the threshold and aggregate attribution fails (Figure 6). This is a concrete failure mode for a fully informed adversary that optimizes against every deployed component.

Evasion also sharply and persistently reduces output diversity. Adding private supercategories expands the semantic regions the adversary must alter, increasing both fine-tuning scope and utility loss. AmbiPrint cannot prevent adaptive evasion, but couples it to broad behavioral change and measurable diversity collapse.

Together, the targeted and multi-category attacks expose the compositional nature of AmbiPrint's robustness. Manipulating one known component does not defeat the aggregate signature, whereas defeating all observed components requires broad semantic drift and coincides with reduced output diversity. Multiple independently chosen supercategories therefore raise the scope of adaptive fine-tuning, although they do not provide a formal guarantee against a fully informed adversary.

4.7 Robustness against Hyperparameters

For SDXL we vary image-to-image strength and guidance scale; for Flux Kontext, which does not expose strength, we vary guidance. Average cross-setting divergences range from 0.009 to 0.023 for SDXL and from 0.004 to 0.008 for Flux Kontext, remaining at or below the mean attribution threshold $\tau = 0.023$ (Figure 7). The fingerprint is therefore stable across the tested inference configurations.



(a) SDXL (b) Flux Kontext

Figure 7: Inference-time hyperparameter sensitivity. Each cell reports the average JS divergence between the AmbiPrint at one setting and those at all other settings. Flux Kontext exposes guidance scale but not image-to-image strength.

5 Conclusion

In this work, we presented AmbiPrint, a novel and efficient black-box fingerprinting framework that uses ambiguous images to stabilize source attribution after model modification. Each image provides semantic guidance and a structural constraint without determining which candidate class dominates. This controlled context preserves the source-specific signal needed to distinguish independent model families, achieving reliable attribution with only 100 generations per model where text-only baselines fail.

Ambiguous images also offer an efficient tool for studying generative models through query access. Future work could adaptively select categories, combine complementary behavioral statistics, and use AmbiPrint as a low-query screening stage before white-box analysis. It could further support model-update monitoring, derivative detection, and model-family mapping.

AI Use Statement

The authors used generative AI to assist with writing, providing overviews and reviews, and preparing tables and related materials.

Ethics Statement

AmbiPrint is intended to support provenance analysis and intellectual-property auditing of image-generation models. Attribution errors could create reputational or legal harm, so a fingerprint match should be treated as supporting technical evidence rather than as a standalone determination of ownership or misuse. Deployments should use representative calibration families, report uncertainty and known failure modes, and respect model licenses, API terms, and applicable privacy and access controls. Our experiments use model outputs and public benchmark images and do not involve human subjects.

Reproducibility Statement

Section 4.1 specifies the probe construction, query budget, semantic projection, distance, and family-disjoint calibration protocol. Appendix A provides the exact prompts, Appendix B lists the evaluated models and configurations, and Appendix C gives complete pseudocode. The supplementary matrices report the pairwise results underlying the aggregate figures and tables.

REFERENCES

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

Giulia Bertazzini, Chiara Albisani, Daniele Baracchi, Dasara Shullani, and Roberto Verdecchia. The hidden cost of an image: Quantifying the energy consumption of ai image generation. *arXiv preprint arXiv:2506.17016*, 2025.

Shihan Cheng, Nilesh Kulkarni, David Hyde, and Dmitriy Smirnov. Less is more: Data-efficient adaptation for controllable text-to-video generation. *arXiv preprint arXiv:2511.17844*, 2025.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10850–10869, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8974–8983, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128:1956–1981, 2020. doi: 10.1007/s11263-020-01316-z. URL https://doi.org/10.1007/s11263-020-01316-z.

Black Forest Labs et al. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742. Version 2 (updated June 23, 2025).

Junhoo Lee, Mijin Koo, and Nojun Kwak. Csf: Black-box fingerprinting via compositional semantics for text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16572–16582, 2026.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

Yue Li, Hongxia Wang, and Mauro Barni. A survey of deep neural network watermarking techniques. *Neurocomputing*, 461:171–193, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Xingchao Liu, Chengyue Zhang, Chujun Yang, Yiming Zhang, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*, 2023.

Yaofang Liu, Yumeng Ren, Aitor Artola, Yuxuan Hu, Xiaodong Cun, Xiaotong Zhao, Alan Zhao, Raymond H. Chan, Suiyun Zhang, Rui Liu, Dandan Tu, and Jean-michel Morel. Pusa v1.0: Unlocking temporal control in pretrained video diffusion models via vectorized timestep adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. Black box adversarial prompting for foundation models. *arXiv preprint arXiv:2302.04237*, 2023.

Guangyu Nie, Changhoon Kim, Yezhou Yang, and Yi Ren. Attributing image generative models using latent fingerprints. In *International Conference on Machine Learning*, pp. 26150–26165. PMLR, 2023.

OpenAI. Dall·e 3 system card. Technical report, OpenAI, 2023.

Maxime Oquab, Timothée Darcet, Theo Mourae, Huy V. Vo, A. Selfridge, ..., L. Bottou, and J. Ponce. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. S. Mahajan, Adam Lamb, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. Manifpt: Defining and analyzing fingerprints of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10791–10801, 2024.

Steven Stalder, Nathanaël Perraudin, Radhakrishna Achanta, Fernando Perez-Cruz, and Michele Volpi. What you see is what you classify: Black box attributions. *Advances in Neural Information Processing Systems*, 35:84–94, 2022.

Huan Teng, Yuhui Quan, Chengyu Wang, Jun Huang, and Hui Ji. Fingerprinting denoising diffusion probabilistic models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28811–28820, 2025.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36: 58047–58063, 2023.

Jie Zhang, Dongdong Chen, Jing Liao, Han Fang, Weiming Zhang, Wenbo Zhou, Hao Cui, and Nenghai Yu. Model watermarking for image processing networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12805–12812, 2020.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. Reef: Representation encoding fingerprints for large language models. *arXiv preprint arXiv:2410.14273*, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

A Input Prompts

In this section, we provide the precise text prompts used to query the generative models across all evaluated semantic categories. As detailed in the main manuscript, we target four distinct super-categories consisting of fruits, animals, insects, and birds. To maintain consistency throughout our experiments, we employ the following prompt templates for each respective category.

- `"A beautiful, high-quality photograph of a fruit, 8k, masterpiece"`
- `"A beautiful, high-quality photograph of an animal, 8k, masterpiece"`
- `"A beautiful, high-quality photograph of an insect, 8k, masterpiece"`
- `"A beautiful, high-quality photograph of a bird, 8k, masterpiece"`

B Model Details

Our work leverages multiple pre-trained generative models. We list the repositories and exact configurations below to ensure reproducibility.

- **Stable Diffusion 1.2** Stable Diffusion 1.2 (https://huggingface.co/CompVis/stable-diffusion-v1-2)
- **Stable Diffusion 1.3** Stable Diffusion 1.3 (https://huggingface.co/CompVis/stable-diffusion-v1-3)
- **Stable Diffusion 1.4** Stable Diffusion 1.4 (https://huggingface.co/CompVis/stable-diffusion-v1-4)
- **Stable Diffusion 1.5** Stable Diffusion 1.5 (https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5)
- **Stable Diffusion 2.1** Stable Diffusion 2.1 Base (https://huggingface.co/Manojb/stable-diffusion-2-1-base). This base model is also used for the 4-bit quantized and 10% unstructured pruned variants.
- **SD 2.1 DreamBooth Dog** SD 2.1 DreamBooth Dog (https://huggingface.co/lora-library/lora-dreambooth-sample-dog)
- **Stable Diffusion 3.5 Medium** Stable Diffusion 3.5 Medium (https://huggingface.co/stabilityai/stable-diffusion-3.5-medium). This base model is also used for the 4-bit quantized variant.
- **SD3.5 MS Paint Style** SD3.5 MS Paint Style (https://huggingface.co/Clybius/mspaint-style-sd3.5medium)
- **SD3.5 Anime Pose** SD3.5 Anime Pose (https://huggingface.co/JINTAIRE/anime_pose)
- **Stable Diffusion XL (Turbo)** Stable Diffusion XL (https://huggingface.co/martyn/sdxl-turbo-mario-merge-top-rated/blob/main/topRatedTurboxlLCM_v10.safetensors). This checkpoint is also used for the 10% unstructured pruned variant.
- **SDXL Studio Ghibli** SDXL Studio Ghibli (https://huggingface.co/ntc-ai/SDXL-LoRA-slider.Studio-Ghibli-style)
- **SDXL Comic Portrait** SDXL Comic Portrait (https://huggingface.co/ntc-ai/SDXL-LoRA-slider.comic-portrait)
- **SDXL Ultra-Realistic Illustration** SDXL Ultra-Realistic Illustration (https://huggingface.co/ntc-ai/SDXL-LoRA-slider.ultra-realistic-illustration)
- **Karlo v1.0** Karlo v1.0 (https://huggingface.co/kakaobrain/karlo-v1-alpha)
- **Kandinsky 2.2** Prior (https://huggingface.co/kandinsky-community/kandinsky-2-2-prior) and Decoder (https://huggingface.co/kandinsky-community/kandinsky-2-2-decoder)

Algorithm 1 AmbiPrint

Input Suspect model M', Samples N, Known fingerprint $\hat{P}_M$, Threshold τ
Output Attribution decision (Derivative or Independent)
for $i = 1$ to N **do do**
    Sample $\{x_1, ..., x_K\}$ from distinct classes ▷ ▷▷ **1. Sample K Source Images**
    Construct ambiguous image c^*_{image} using $\mathcal{H}(x_1, ..., x_K)$ as in Eq. (5) ▷ ▷▷ **2. Get Ambiguous Image**
    Form ambiguous image-text probe $c^* = (c^*_{\text{text}}, c^*_{\text{image}})$
    $y_i \leftarrow C(M'(c^*))$ ▷ ▷▷ **3. Get Model Outputs**
end for
Compute $\hat{P}_{M'}$ by averaging $\{y_1, ..., y_N\}$ as in Eq. (7) ▷ ▷▷ **4. Get Fingerprint**
$d \leftarrow D(\hat{P}_{M'}, \hat{P}_M)$ ▷ ▷▷ **5. Calculate Distance**
if $d \leq \tau$ **then**
    return Derivative
else
    return Independent
end if

Algorithm 2 Spiral Mixing Function ($\mathcal{H}$)

Input K source classes, Input size S_{in}, Output size S_{out}
Output Ambiguous image c^*_{image}
for $k = 1$ to K **do**
    $x_k \leftarrow \text{RandomSample}(\text{Class}_k)$ ▷ ▷▷ **1. Sample K Source Images**
    $x_k \leftarrow \text{Resize}(x_k, (S_{\text{in}}, S_{\text{in}}))$ ▷ ▷▷ **2. Resize Source Images**
end for
Initialize empty output image c^*_{image} of size $(S_{\text{out}}, S_{\text{out}})$
$n \leftarrow 1$
for each pixel (y, x) in an outward-in spiral path **do** ▷ ▷▷ **3. Apply Mixing Algorithm**
    $k \leftarrow ((n-1) \pmod K) + 1$ ▷ Get round-robin index
    $(sx, sy) \leftarrow \text{ScaleCoordinates}((y, x), S_{\text{out}}, S_{\text{in}})$
    $c^*_{\text{image}}[y, x] \leftarrow x_k[sy, sx]$ ▷ Assign pixel from source x_k
    $n \leftarrow n + 1$
end for
return c^*_{image}

- **Kandinsky 2.2 Naruto** Kandinsky 2.2 Naruto (https://huggingface.co/leffff/kandinsky-2.2-naruto-lora)
- **Kandinsky 3** Kandinsky 3 (https://huggingface.co/kandinsky-community/kandinsky-3)
- **FLUX.1 Kontext Dev** FLUX.1 Kontext Dev (https://huggingface.co/black-forest-labs/FLUX.1-Kontext-dev)
- **Gemini 2.5 Flash Image (Nano Banana)** Gemini 2.5 Flash Image (https://aistudio.google.com/models/gemini-2-5-flash-image)
- **DALL-E 2** DALL-E 2 (https://openai.com/index/dall-e-2/)

We encourage checking the respective model cards for specific licenses and further details.

C Algorithm

We provide detailed pseudocode for AmbiPrint in Alg. 1 and the mixing procedure in Alg. 2.

D Detailed Results

This section presents additional qualitative and quantitative results, supplementing the main paper.

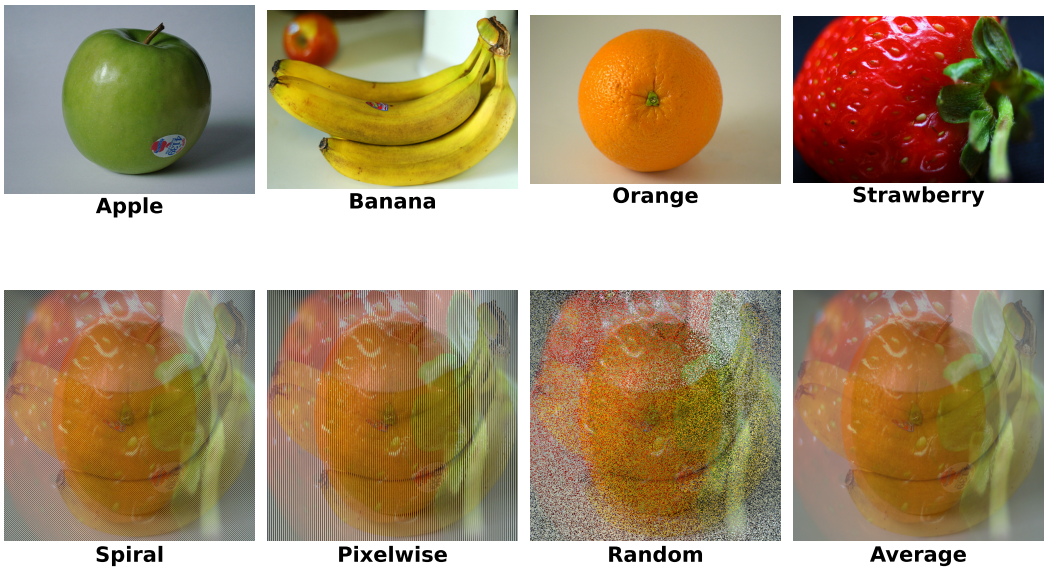


Figure 8: Qualitative demonstration of various mixing algorithms. The top row contains the source images, and the bottom row contains the resulting ambiguous images.

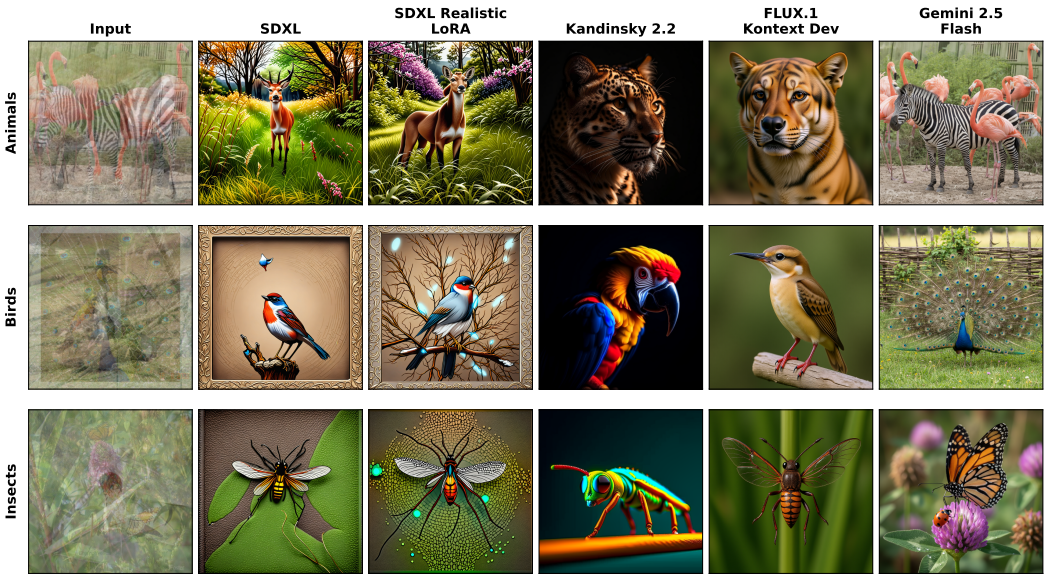


Figure 9: Qualitative demonstration of our method's principle. An identical ambiguous image (left) is passed to models from distinct families.

D.1 Qualitative Results

We provide a qualitative example of how each mixing function operates in Fig. 8. Additionally, Fig. 9 presents the qualitative results of passing the same input to multiple model families.

D.2 Quantitative Results

We provide the full matrices used in all our experiments in Figs. 10–27.

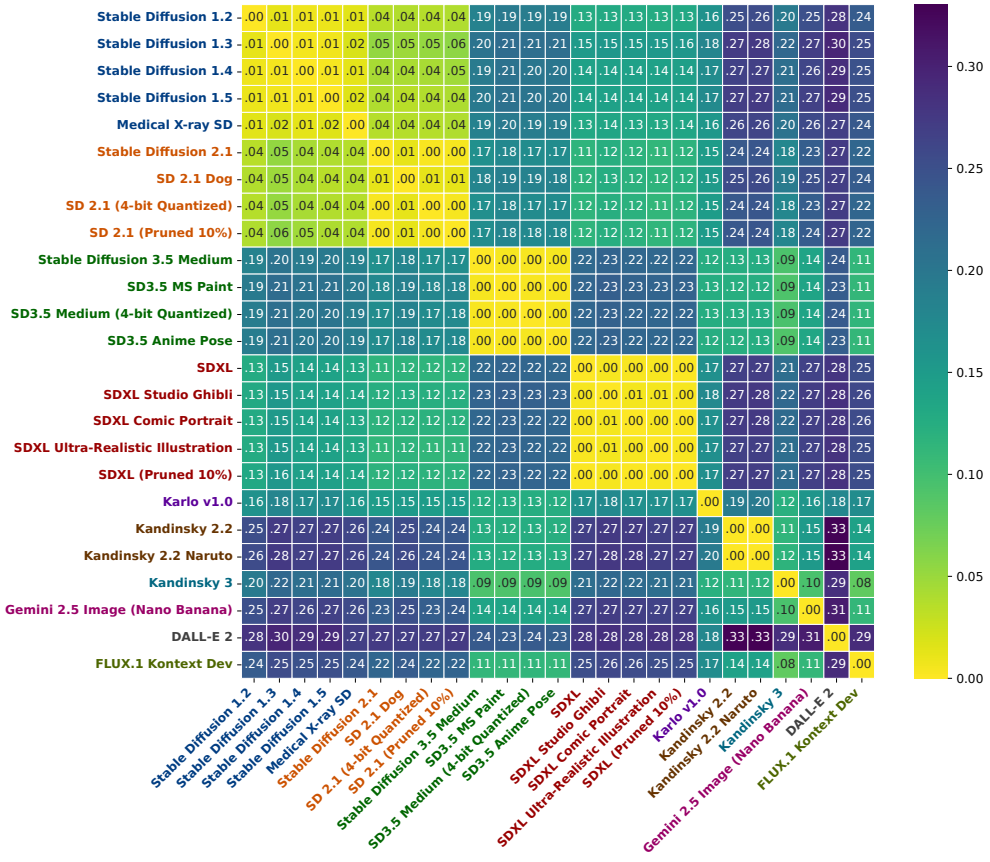


Figure 10: Aggregate AmbiPrint distance

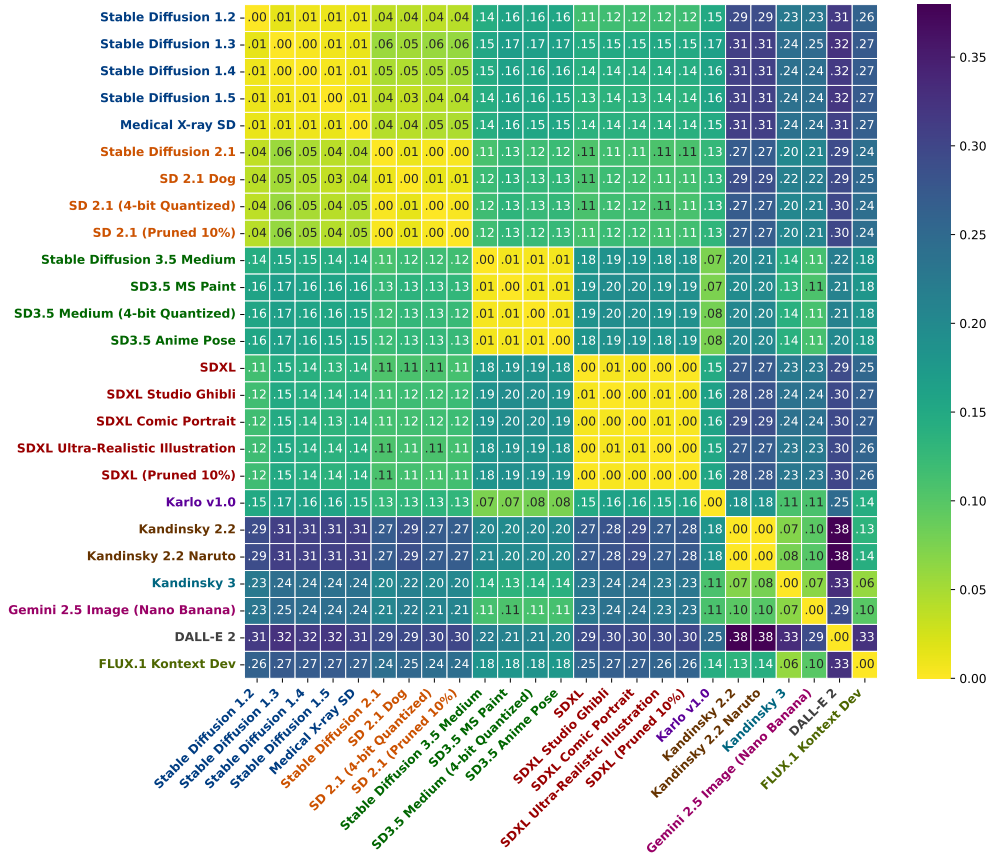


Figure 11: Fruit AmbiPrint distance

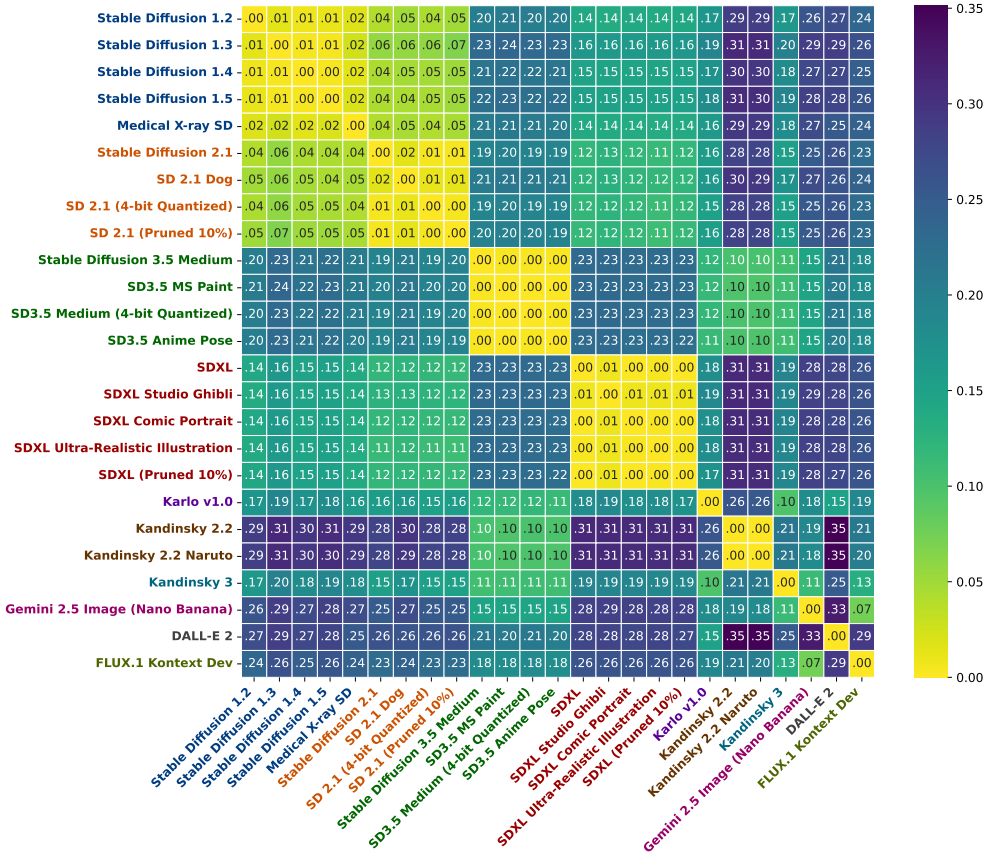


Figure 12: Animal AmbiPrint distance

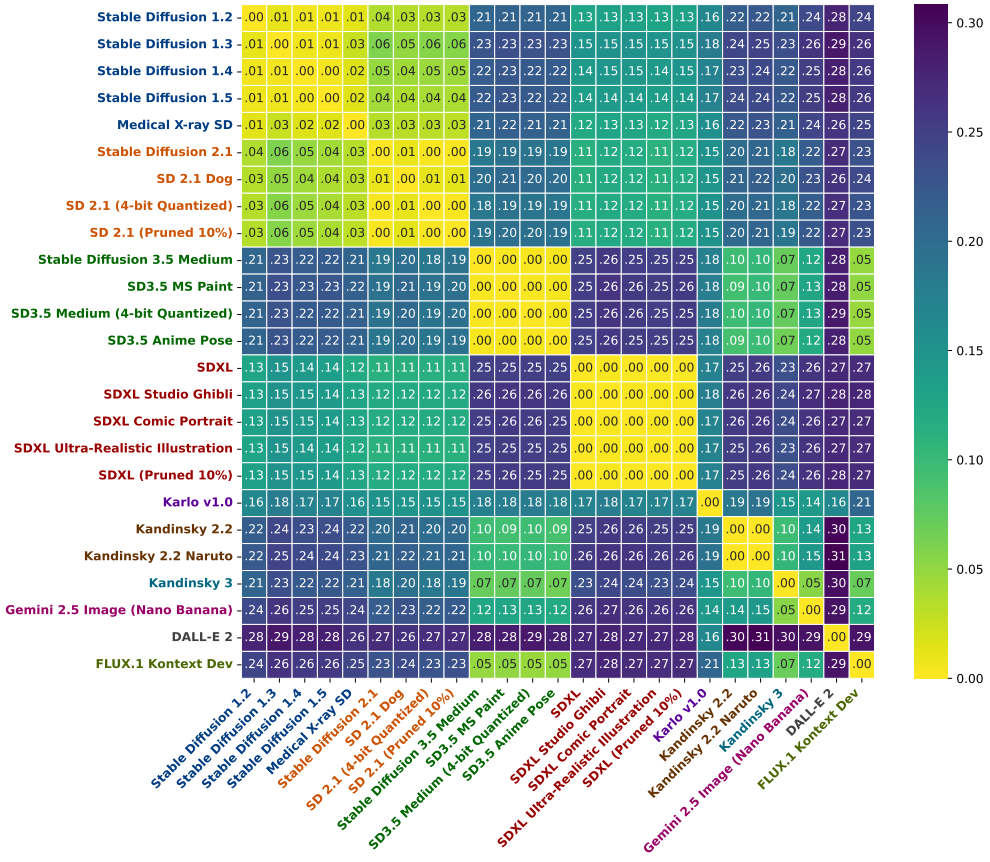


Figure 13: Insect AmbiPrint distance

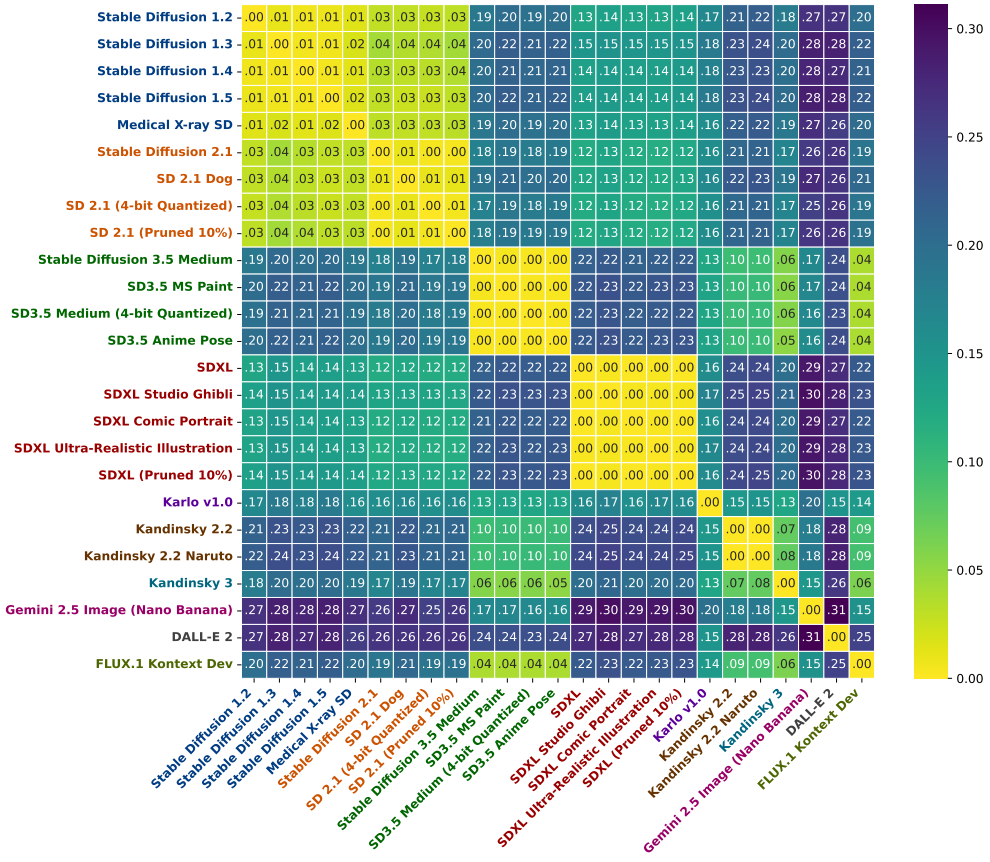


Figure 14: Bird AmbiPrint distance

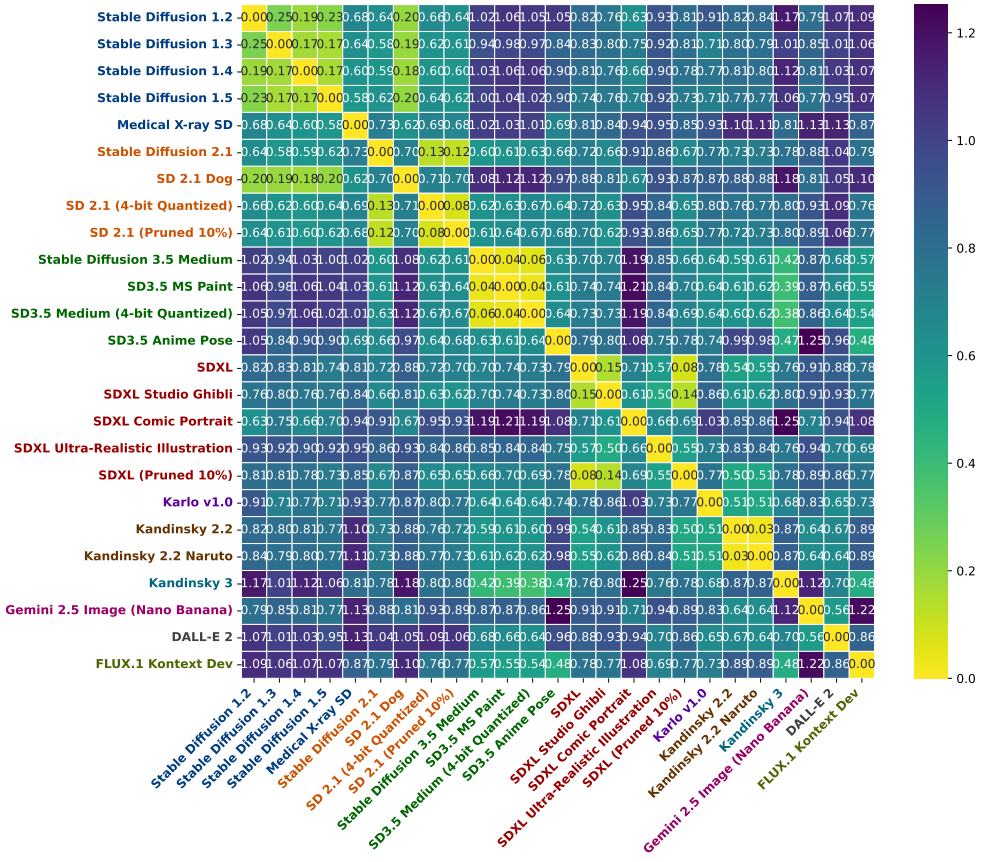


Figure 15: CSF fingerprinting results

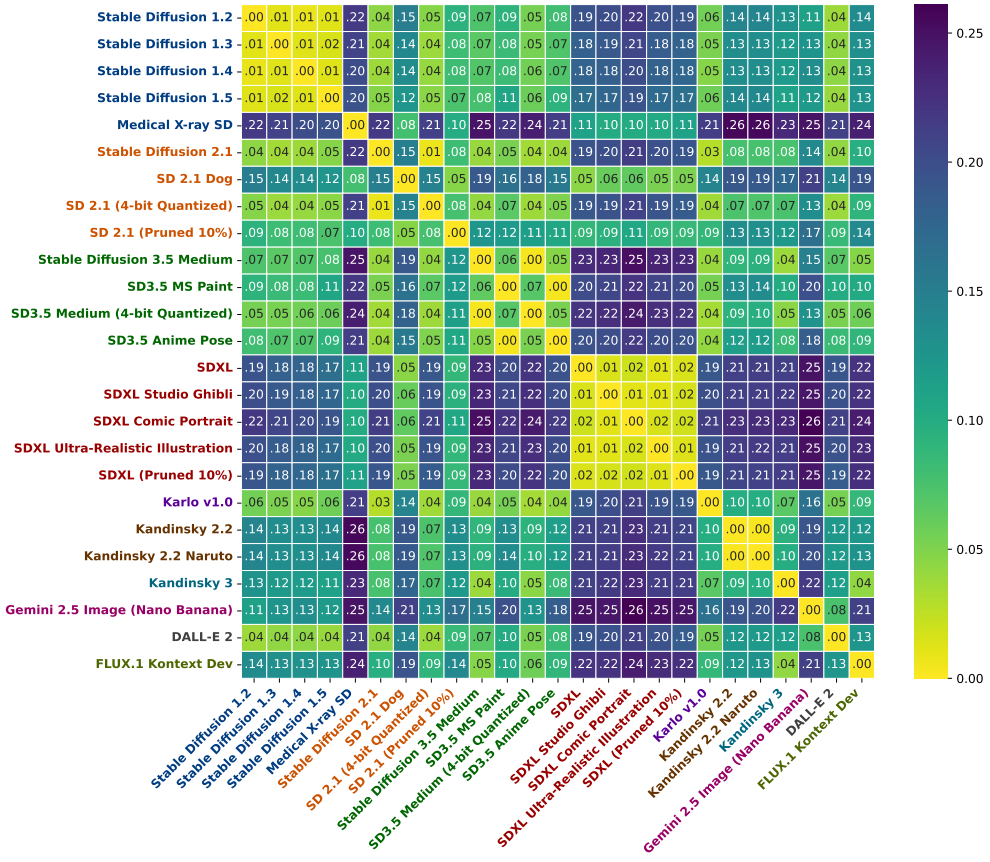


Figure 16: Text-only AmbiPrint distance with $N = 100$

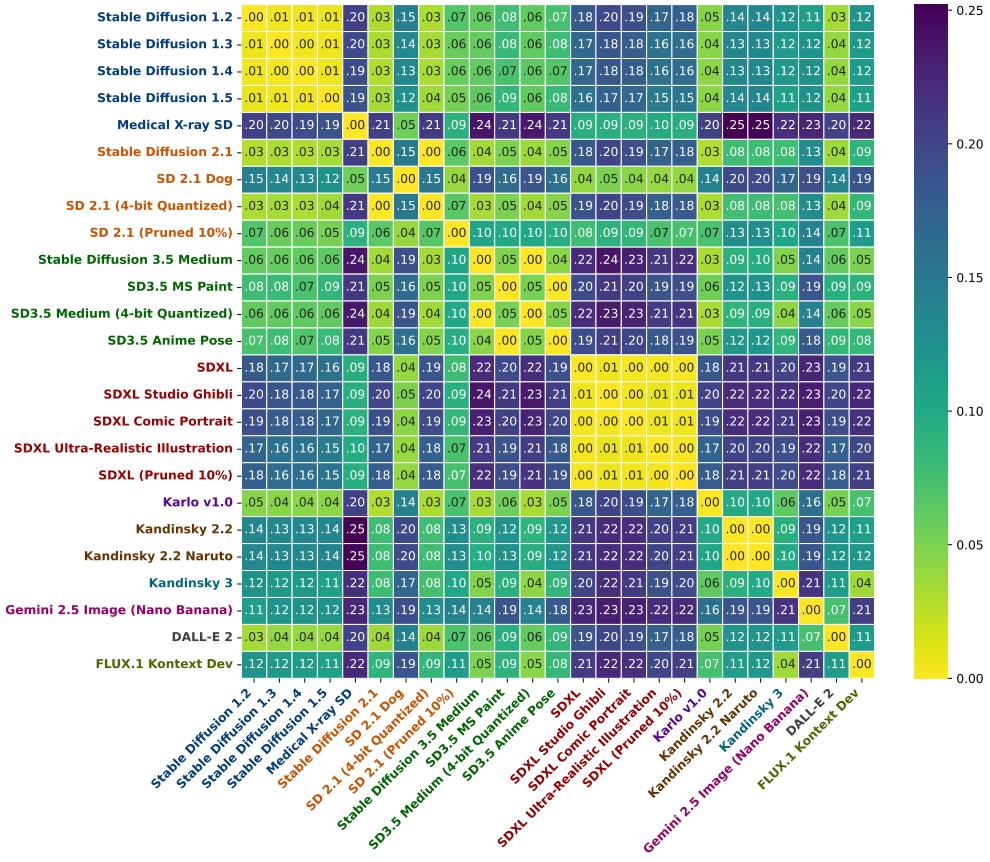


Figure 17: Text-only AmbiPrint distance with $N = 400$

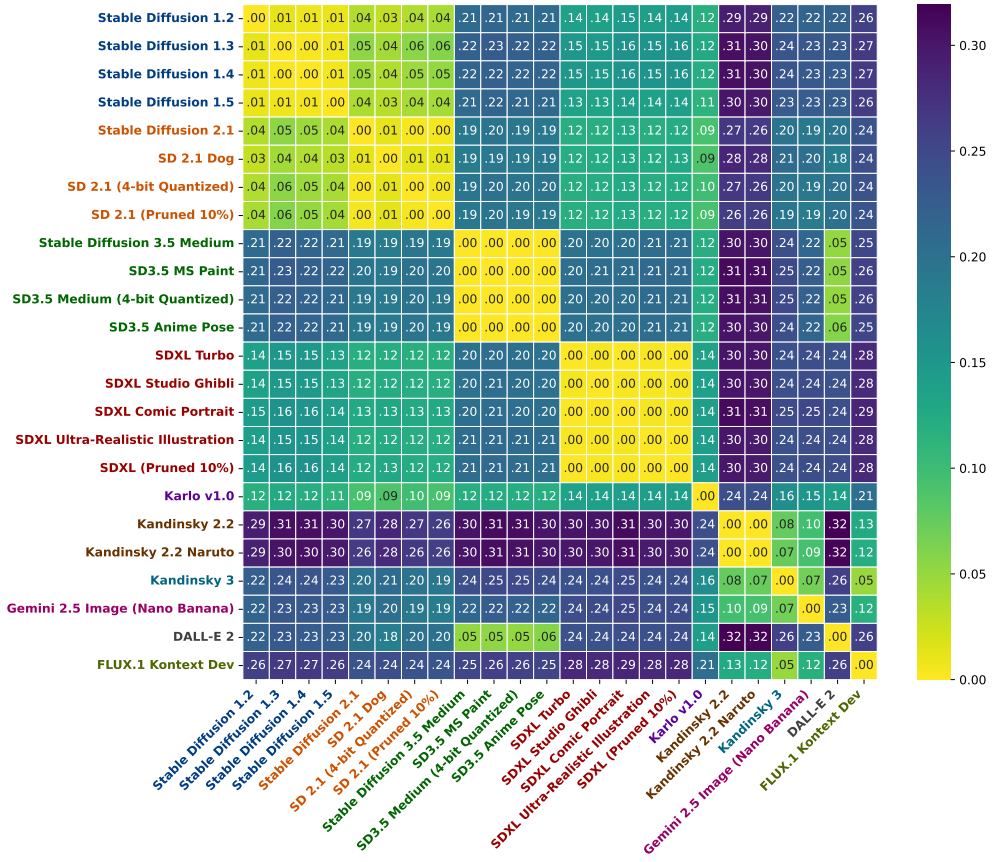


Figure 18: AmbiPrint distance with Seqeuntial-mix Images

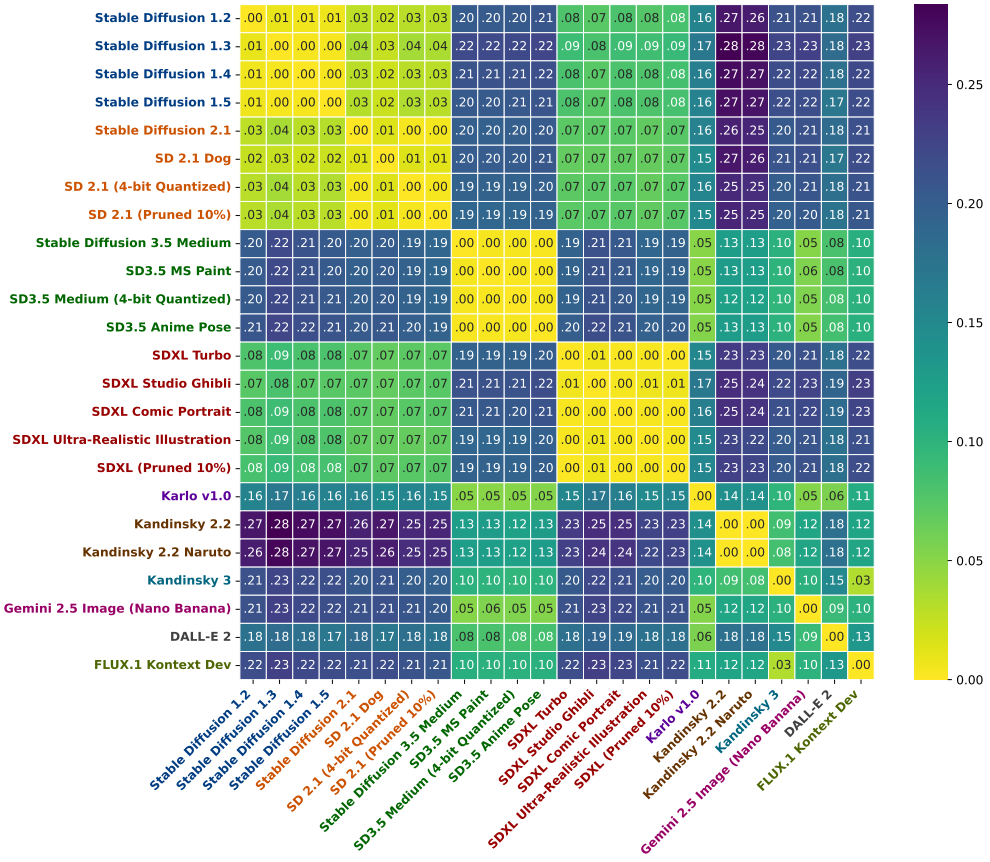


Figure 19: AmbiPrint distance with Averaging-mix Images

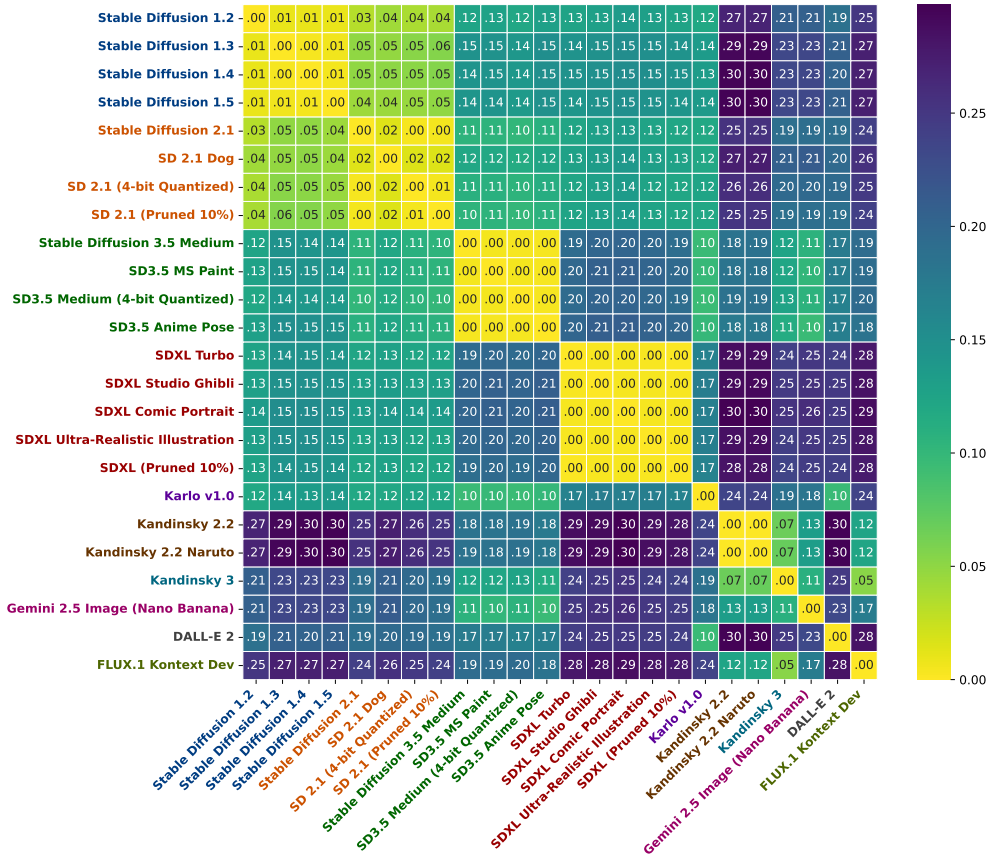


Figure 20: AmbiPrint distance with Random-mix Images

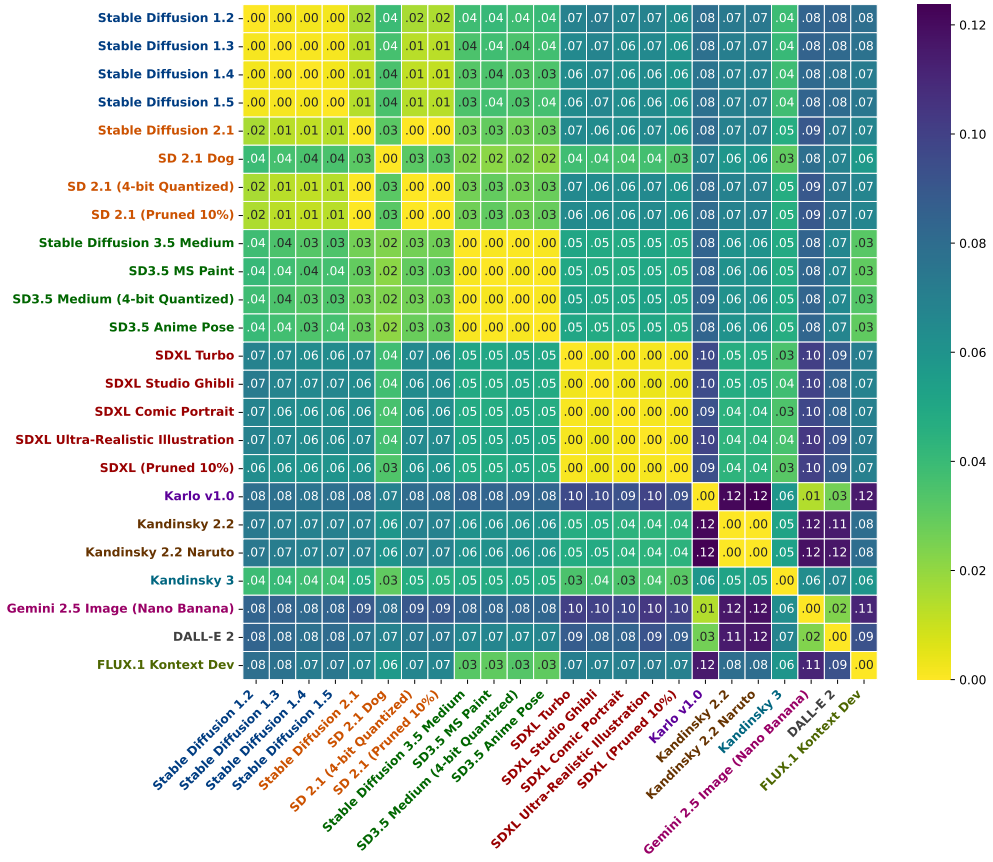


Figure 21: AmbiPrint distance with $K = 1$

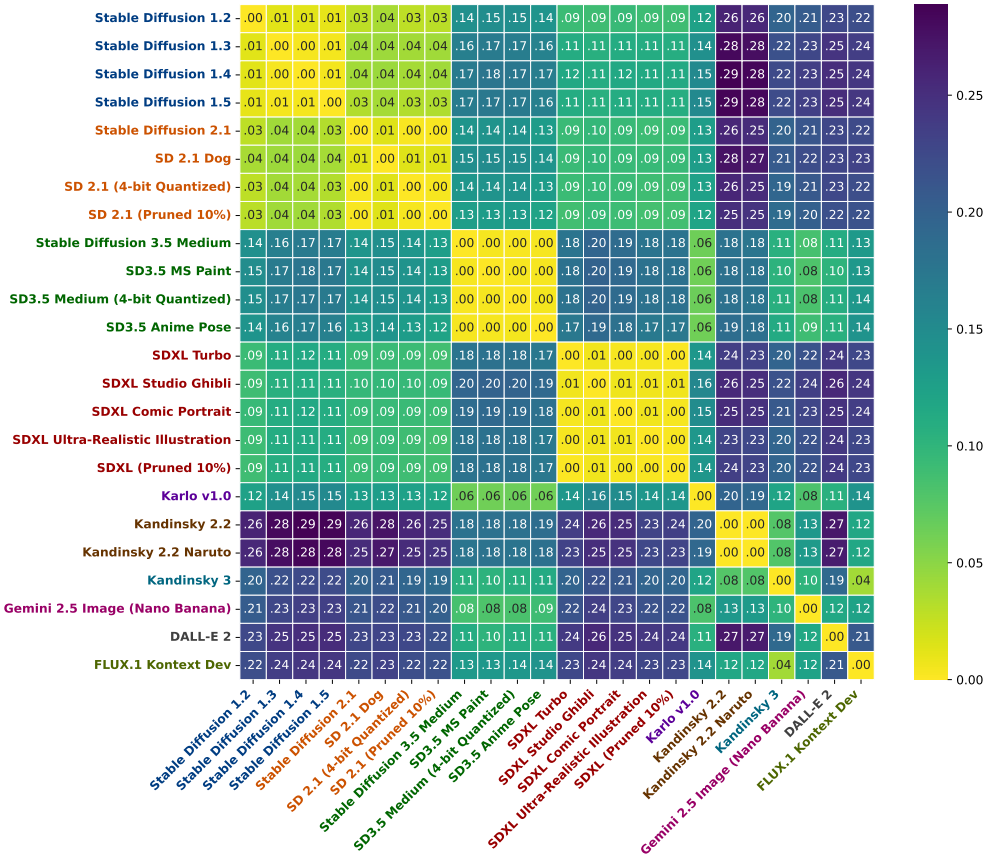


Figure 22: AmbiPrint distance with $K = 2$

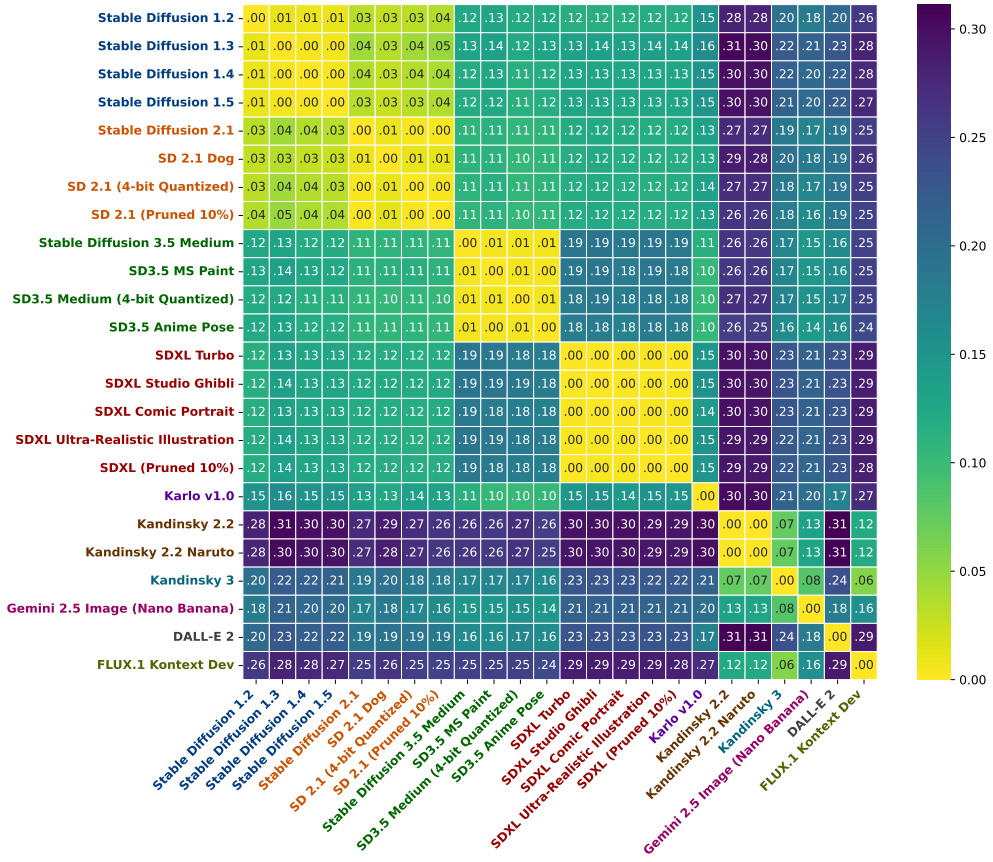


Figure 23: AmbiPrint distance with $K = 8$

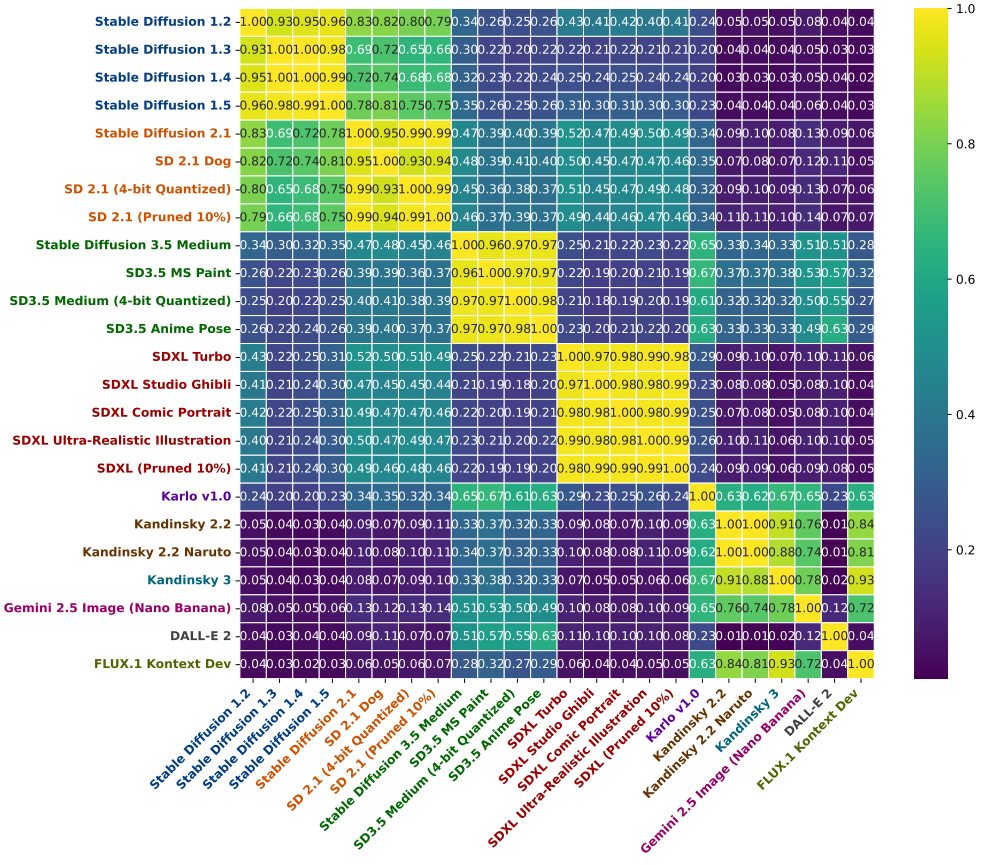


Figure 24: AmbiPrint distance using cosine similarity

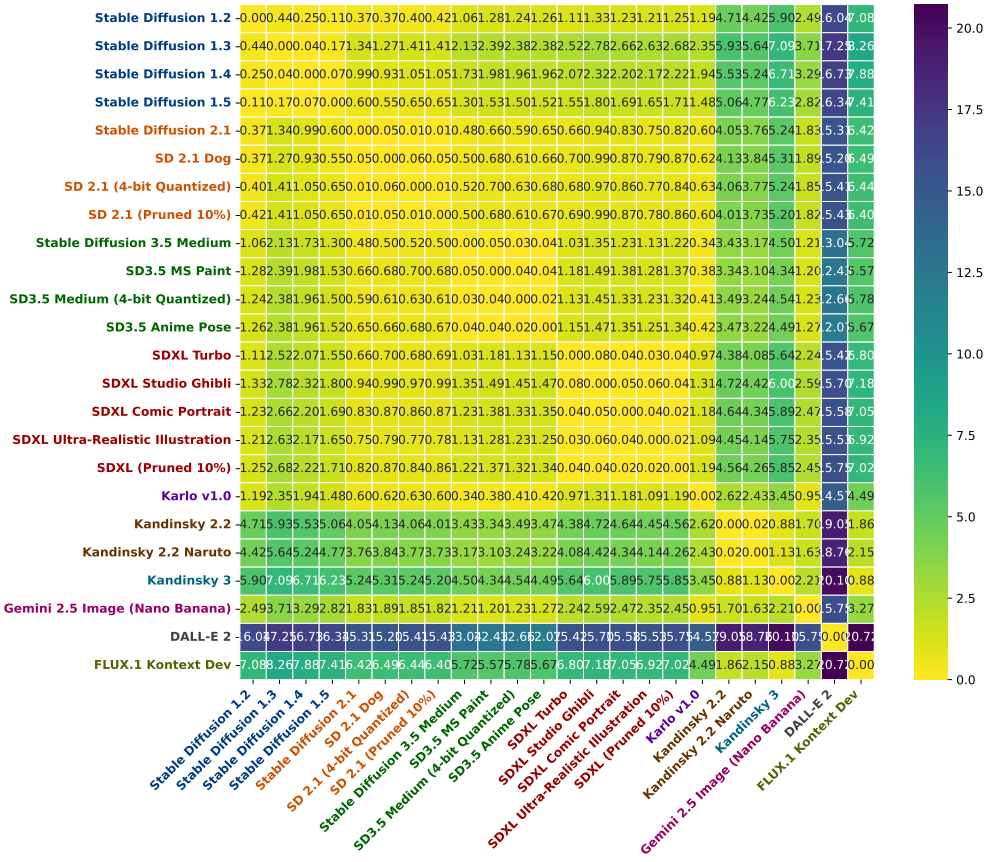


Figure 25: AmbiPrint distance using MSE

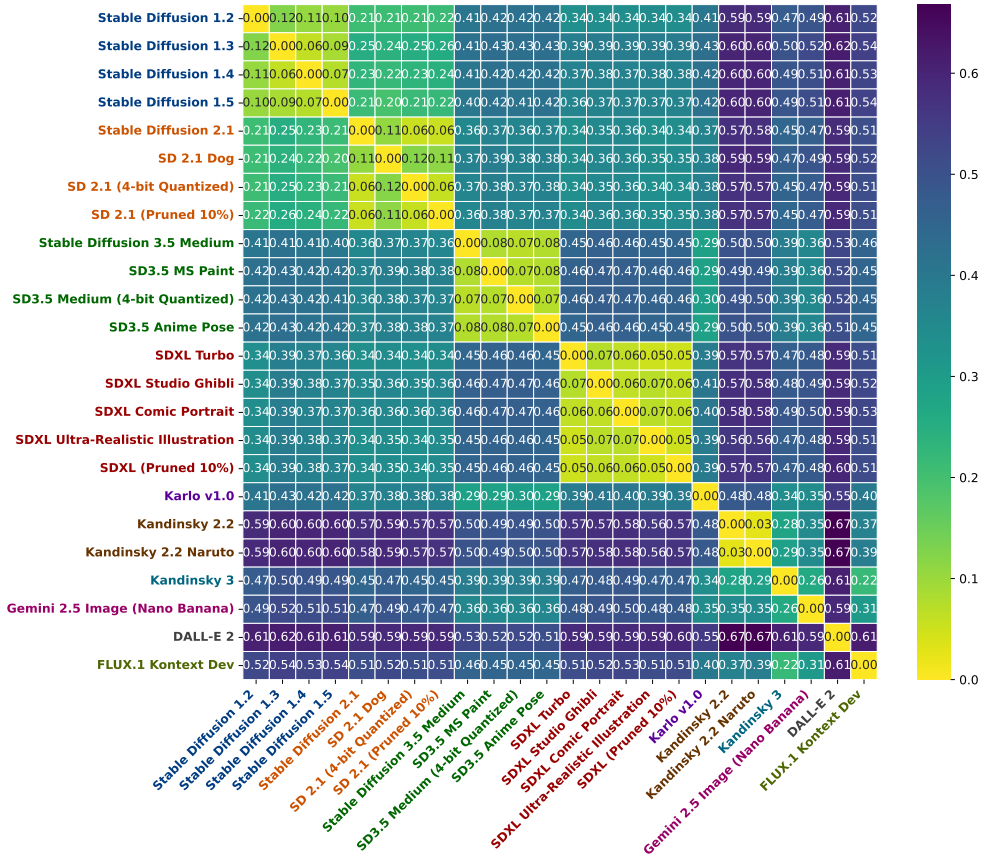


Figure 26: AmbiPrint distance using total variational distance

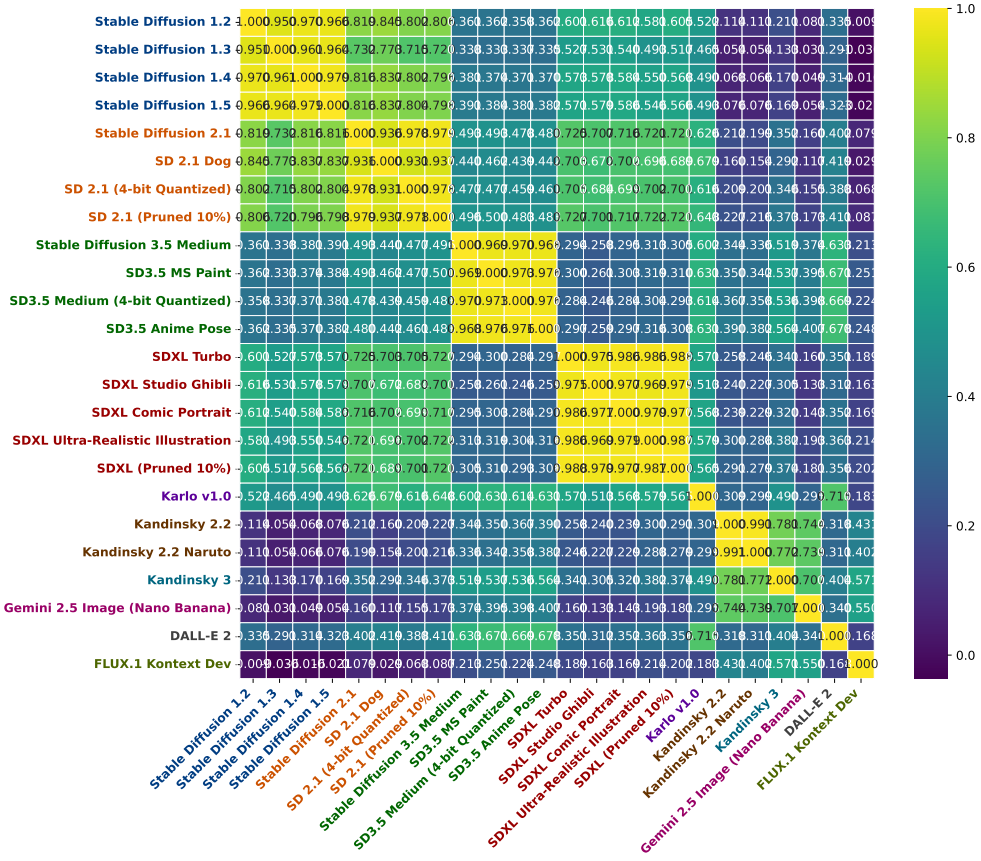


Figure 27: AmbiPrint distance with the DINO embedding model, using cosine similarity