PromptSplit: Revealing Prompt-Level Disagreement in Generative Models

Mehdi Lotfian [1] **Mohammad Jalali** [1] **Farzan Farnia** [1]

Abstract

We study the problem of comparing prompt-conditioned generative models through the lens of kernel-based distribution comparison. Given two models evaluated on prompts, our goal is to identify prompt regions where their conditional output distributions differ, without supervision or repeated sampling for each prompt. We propose *PromptSplit*, an unsupervised spectral kernel method that represents each prompt–response pair through a tensor-product feature map and compares models via the difference of their joint prompt–output kernel covariance matrices. The leading positive eigendirections of this covariance difference provide soft modes of discrepancy, allowing the method to localize prompt categories associated with systematic differences in model behavior. We show that the required eigenspectrum can be computed through an equivalent block kernel matrix involving Hadamard products of prompt and output kernels, and introduce a random-projection implementation that reduces the computation to $O(r^2 \cdot \max\{n, r\})$ for projection dimension r. We further prove that the projected formulation provides a controlled approximation to the full spectral comparison. Our numerical evaluation on controlled and real prompt-guided generation tasks, including text-to-image and text-to-text models, shows that PromptSplit recovers known prompt-dependent differences and reveals explainable modes of difference between generative models.

1. Introduction

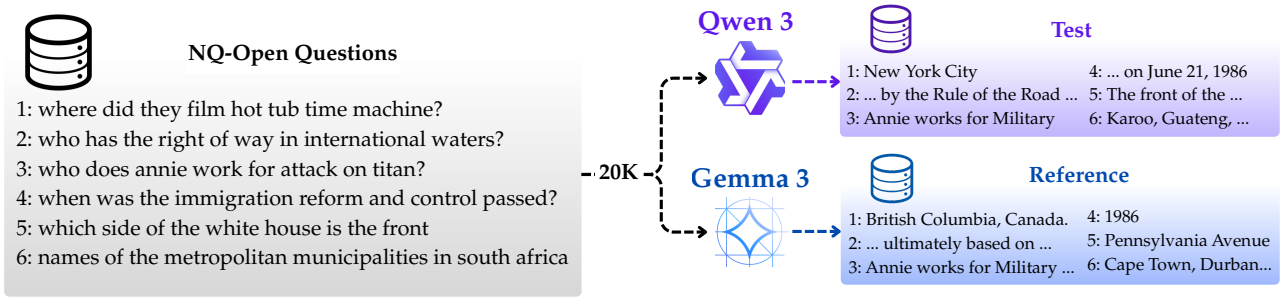
Prompt-guided generative models have advanced rapidly across vision and language, enabling high-fidelity synthesis from input text prompts. In computer vision, diffusion-based text-to-image and text-to-video generation systems have demonstrated impressive realism and prompt alignment (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Singer et al., 2022; Ho et al., 2022). In parallel, large language models (LLMs) have scaled text generation and often serve as language backbones within multimodal pipelines (Brown et al., 2020; Touvron et al., 2023; Gemini-Team, 2023; Bi et al., 2024; Qwen-Team, 2024; Minaee et al., 2024). These advances have produced a rich ecosystem of prompt-guided AI models differing in training data, model architectures, and conditioning strategies.

With such diversity, a central question arises: *how can we systematically compare prompt-guided generative models and determine when their responses differ for an input prompt?* Beyond qualitative demonstrations, quantitative assessment has become a key research focus. While fidelity evaluation metrics such as FID (Heusel et al., 2017), Inception Score (Salimans et al., 2016), and CLIP-Score (Hessel et al., 2021) provide useful aggregate indicators of fidelity and alignment, they compress model behavior into single quality-based scores and can obscure prompt-dependent discrepancies across models that may not necessarily stem from the quality of model outputs. For example, a text-to-image model $\mathcal{G}_A$ model may generate a female-individual whenever the prompt contains the word "person", while text-to-image model $\mathcal{G}_B$ always generates the picture of a male individual. While such a difference may not influence the quality of image generation for a single prompt, it may lead to different model responses for certain categories of input prompts. A more informative comparison should reveal where and how two models behave differently *as a function of the prompt* and specifically identify the prompt categories that lead to divergent output responses by the two models.

Recent work has explored distributional and spectral methods to compare the output spaces of generative models from their samples (Zhang et al., 2025; 2024). However, existing formulations generally operate in a prompt-free setting, effectively marginalizing over prompts. Applying such methods directly to prompt-conditioned generation ignores the input prompt and blurs prompt-aware differences into an aggregate distribution. A naive workaround is to analyze each prompt separately by generating many outputs per prompt and running a per-prompt spectral comparison. However,

[1]The Chinese University of Hong Kong. Correspondence to: Mehdi Lotfian <lotfian25@cse.cuhk.edu.hk>, Mohammad Jalali <mjalali24@cse.cuhk.edu.hk>, Farzan Farnia <farnia@cse.cuhk.edu.hk>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.



Top Distinct Modes Detected By Prompt Split

	Prompts	Test: Qwen 3	Reference: Gemma 3
Identified Mode #1 ($\lambda = 1.95 \times 10^{-2}$)	1: **who played** kelly on save by the bell	1: Morgan Freeman	1: Teri Hatcher
	2: **who plays** jessie in saved by the bell	2: Morgan Freeman	2: Lisa Marie Taylor
	3: **who plays** the voice of gaston in beauty and the beast	3: Morgan Freeman	3: Rex Harrison
	4: **who plays** nell jones on ncis los angeles	4: Morgan Freeman	4: Annie Parisse
Identified Mode #2 ($\lambda = 6.48 \times 10^{-3}$)	1: who was the first **us president** who was not a military veteran	1: Franklin D. Roosevelt	1: John Adams
	2: which **u.s. president** enacted the federal income tax system	2: Franklin D. Roosevelt	2: John Quincy Adams
	3: who was the **us president** when uncle sam got his nickname	3: Franklin D. Roosevelt.	3: Abraham Lincoln
	4: who was our second **president** of the **united states**	4: George Washington.	4: John Adams

Figure 1. Overview of Method for discovering different types of (prompt, answer) between two models. (a) From NQ-Open questions, we generate outputs from the test model (Qwen3) and reference model (Gemma3). (b) Two high-scoring modes found by PromptSplit: for each mode we show representative prompts and the corresponding model outputs.

such an approach will be computationally expensive as it requires a significant number of output generations per prompt, which may be unnecessary for the goal of identifying only the prompt categories of the different behavior.

To address this task, we introduce *PromptSplit* (Figure 1), a prompt-aware unsupervised spectral framework to detect prompt categories leading to different model behaviors, which couples prompts and outputs within a joint representation to attain the goal. To apply PromptSplit, for every generative model, we construct tensor-product embeddings of the prompt and output features (e.g., text and image embeddings in text-to-image models) and then compute the joint kernel covariance matrix of every model. For a pair of models with kernel covariance matrices $C_{X\otimes T}$ and $C_{Y\otimes T}$, we analyze their weighted difference as follows

$$\Lambda_{X,Y|T} \;:=\; C_{X\otimes T} - C_{Y\otimes T}. \tag{1}$$

We highlight that the principal eigenvalues and eigenvectors of $\Lambda_{X,Y|T}$ can reveal the prompt clusters resulting in different output structures by the models. To compute the eigenvectors of the above matrix, we apply kernel trick and show that the eigendirections of the above matrix is in one-to-one correspondence to those of the following block kernel matrix:

$$\mathbf{K}_{X,Y|T} \;=\; \begin{bmatrix} K_{XX} \odot K_{TT} & K_{XY} \odot K_{TT} \\ -K_{YX} \odot K_{TT} & -K_{YY} \odot K_{TT} \end{bmatrix},$$

where $\odot$ denotes the elementwise matrix product and each $K_{TT'}, K_{XX}, K_{XY}$ represents kernel similarities between prompts or outputs across the two models. This joint structure enables spectral comparison of model responses while explicitly accounting for prompt information, allowing PromptSplit to identify prompt categories that differentiate the two generation models.

Subsequently, we discuss that a direct computation of the eigenvectors of $\mathbf{K}_{\Delta}$ will lead to a cubically growing complexity $\mathcal{O}\big((m+n)^3\big)$ in the number of samples, limiting the application of the algorithm to m, n values of at most a few tens of thousands. To apply PromptSplit for significantly larger sample sizes, which would be necessary to ensure its proper convergence, we introduce a random-projection reduction of the joint (tensor) features to a target dimension r, reducing per-sample cost to $\mathcal{O}(r(d_t + d_x))$ for prompt and output embedding dimensions d_t, d_X, and overall spectral decomposition complexity to $\mathcal{O}\big((m+n)r^2 + r^3\big)$. We further show that this approximation achieves an expected eigenspace deviation bounded by $\mathcal{O}(1/\sqrt{r})$, enabling efficient spectral analysis with bounded random projection dimension r values.

We evaluate PromptSplit across text-to-image models and text-to-text (LLM) comparisons. In synthetic settings with known prompt-wise differences, PromptSplit accurately recovers the responsible prompt categories. On real systems, including latent-diffusion and diffusion-transformer models such as Stable Diffusion, Kandinsky, and PixArt—PromptSplit reveals prompt families where responses diverge in style, composition, and alignment, com-

plementing aggregate metrics with interpretable prompt-level disagreement maps (Rombach et al., 2022; Arkhipkin et al., 2024; Chen et al., 2023).In summary, this work (i) formulates prompt-level model comparison as a joint prompt–output spectral problem, (ii) introduces PromptSplit for analyzing the eigenspectrum of joint kernel covariance differences, (iii) provides a scalable random-projection approximation with $\mathcal{O}(1/\sqrt{r})$ theoretical accuracy given projection dimension r, and (iv) numerically demonstrates the analysis of prompt-dependent disagreement across synthetic and real prompt-guided generative models.

2. Related Work

Evaluation of generative models. Classical metrics for evaluating generative models include IS (Salimans et al., 2016), FID (Heusel et al., 2017), and KID (Binkowski et al., 2018), while precision–recall metrics (Kynkäänniemi et al., 2019) and density/coverage (Naeem et al., 2020) separately measure fidelity and diversity. Spectral metrics such as Vendi (Friedman et al., 2022), RKE (Jalali et al., 2023), and KEN (Zhang et al., 2024) incorporate eigenvalue distributions of kernel similarities to capture global diversity and novelty. CLIPScore (Hessel et al., 2021) evaluates text–image alignment using CLIP embeddings (Radford et al., 2021), and recent analyses (Stein et al., 2023) show that DINOv2 features (Oquab et al., 2024) often yield more reliable evaluations than Inception features. These metrics provide scalar fidelity/diversity summaries, whereas PromptSplit uses embeddings only to build joint kernels and performs prompt-conditioned spectral comparison.

Comparisons of generative models. Spectral and kernel-based approaches compare generative models by analyzing eigenspectra or kernel embeddings of their outputs. The spectral methods in (Zhang et al., 2024; 2025; Ospanov et al., 2024) detects novelty relative to a reference distribution. Beyond generative models, dataset-level comparison frameworks explain how two datasets differ, either via interpretable prototype and influential-example explanations (Babbar et al., 2025) or via transport maps and counterfactuals for image-based distribution shifts (Kulinski & Inouye, 2022), and survey work provides a taxonomy of dataset similarity measures including distance-, kernel-, and embedding-based criteria (Stolte et al., 2024). Compared to these unconditional or dataset-level methods, PromptSplit directly targets *prompt-conditioned* comparison by joint prompt–output embeddings and performing spectral analysis on prompt-dependent covariance differences.

Interpretability for generative models and embeddings. Network Dissection (Bau et al., 2017) and GAN Dissection (Bau et al., 2019) analyze internal units of CNNs and GANs to identify semantic concepts, while GANSpace (Härkönen et al., 2020) uncovers interpretable latent directions through PCA in latent space. Concept activation vectors (TCAV) (Kim et al., 2018) assign user-defined concepts to model directions. These methods explain latent or hidden representations *within* a single model. PromptSplit differs by analyzing disagreement between *two models* through eigenspaces of prompt–output kernel covariance differences, without accessing internal activations.

Random projection of feature embeddings. Random features (Rahimi & Recht, 2007) approximate shift-invariant kernels with low-dimensional embeddings, enabling scalable kernel learning. Approximate kernel k-means (Chitta et al., 2011) and explicit polynomial feature maps (Pham & Pagh, 2013) accelerate clustering and large-scale learning, and Fourier features (Tancik et al., 2020) are widely used to capture high-frequency structure in neural fields. PromptSplit uses a model-agnostic linear projection of joint tensor-product embeddings to reduce the cost of eigendecomposition while preserving disagreement directions.

3. Preliminaries

3.1. Kernel Matrices and Covariance Operators

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined to be a symmetric positive semi-definite (PSD) function that admits a feature map $\phi : \mathcal{X} \to \mathcal{H}$ into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ such that

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

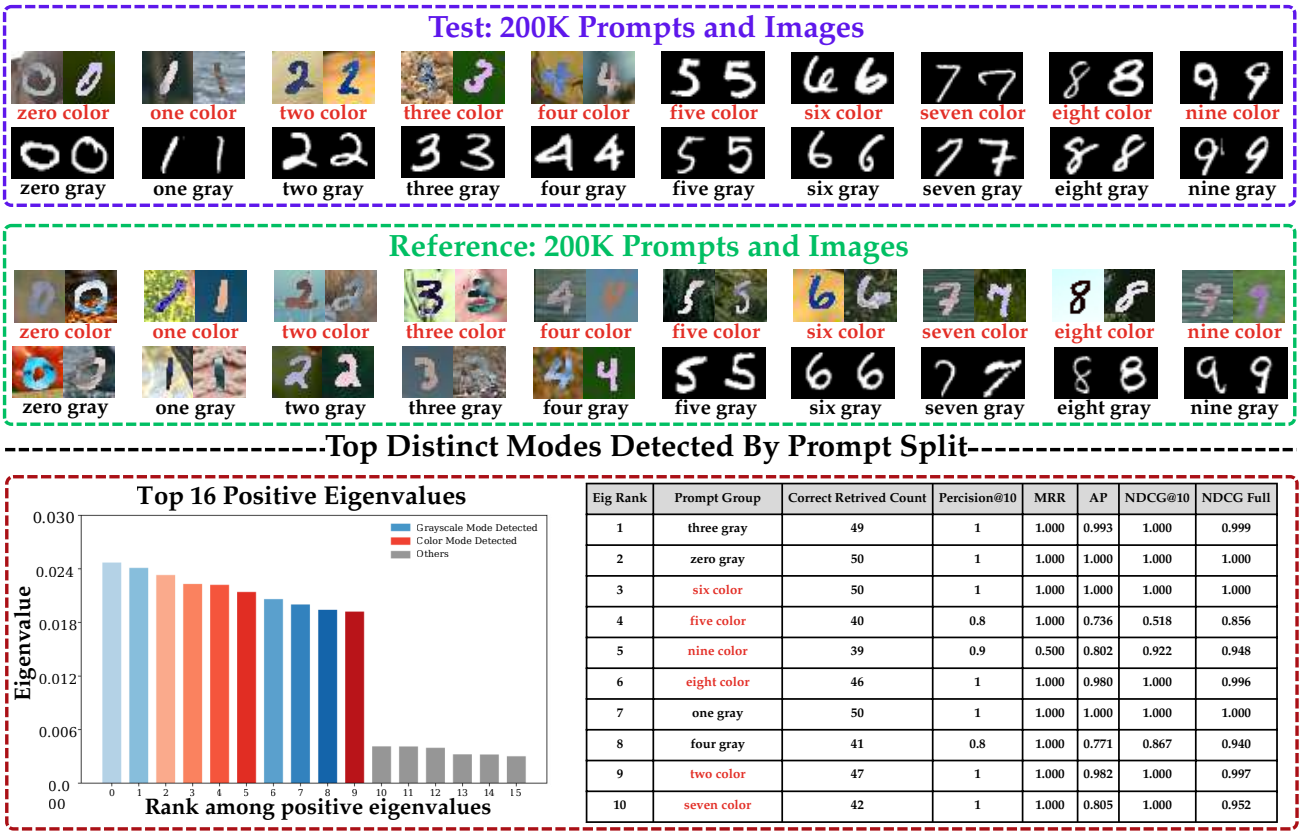
Given samples $\{x_i\}_{i=1}^n$, the empirical kernel matrix is defined as $K_{XX} = \big[k(x_i, x_j)\big]_{i,j=1}^n \in \mathbb{R}^{n\times n}$, and for two sets $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$, the cross-kernel matrix is $K_{XY} = \big[k(x_i, y_j)\big]_{i,j}$.

The kernel covariance operator associated with a random variable $X \sim P_X$ is defined by

$$C_X = \mathbb{E}_{X\sim P_X}\big[\phi(X)\phi(X)^\top\big].$$

We denote the empirical kernel covariance operator with $\widehat{C}_X = \frac{1}{n}\sum_{i=1}^n \phi(x_i)\phi(x_i)^\top$. Note that the nonzero eigenvalues of $\widehat{C}_X$ coincide with those of $\frac{1}{n}K_{XX}$, since $\frac{1}{n}K_{XX} = \frac{1}{n}\Phi_X^\top\Phi_X$ and $\widehat{C}_X = \frac{1}{n}\Phi_X\Phi_X^\top$, for matrix $\Phi_X = \big[\phi(x_1); \cdots ; \phi(x_n)\big]^\top \in \mathbb{R}^{n\times d}$, where the matrix multiplication order is flipped, preserving the non-zero eigenvalues. Assuming a normalized kernel satisfying $k(x, x) = 1$ for every $x \in \mathcal{X}$, we note that the eigenvalues of the normalized kernel matrix $\frac{1}{n}K_{XX}$ will be all non-negative and they sum up to one as $\mathrm{Tr}\big(\frac{1}{n}K_{XX}\big) = 1$.

We call a kernel function k shift invariant if there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that for every $x, y \in \mathbb{R}^d$: $k(x, y) = \kappa(x - y)$. For shift-invariant kernels of the form $k(x, y) = \kappa(x - y)$ on $\mathbb{R}^d$, Bochner's theorem proves that



Eig Rank	Prompt Group	Correct Retrived Count	Percision@10	MRR	AP	NDCG@10	NDCG Full
1	three gray	49	1	1.000	0.993	1.000	0.999
2	zero gray	50	1	1.000	1.000	1.000	1.000
3	six color	50	1	1.000	1.000	1.000	1.000
4	five color	40	0.8	1.000	0.736	0.518	0.856
5	nine color	39	0.9	0.500	0.802	0.922	0.948
6	eight color	46	1	1.000	0.980	1.000	0.996
7	one gray	50	1	1.000	1.000	1.000	1.000
8	four gray	41	0.8	1.000	0.771	0.867	0.940
9	two color	47	1	1.000	0.982	1.000	0.997
10	seven color	42	1	1.000	0.805	1.000	0.952

Figure 2. PromptSplit identified top clusters of prompts with distinct images. Top: 20 clusters of prompts with sample images for test and reference dataset. Bottom: Top 16 eigenvalues showing top 10 disagreement causing prompts and the retrieval metrics accuracy for each mode.

the Fourier transform $\widehat{\kappa} : \mathcal{R}^d \to \mathbb{R}$ is a valid probability density function (PDF), where the Fourier transform is defined as

$$\widehat{\kappa}(\omega) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \kappa(x) \exp\big(-i\langle \omega, x\rangle\big) \mathrm{d}x,$$

then we have $k(x, y) = \mathbb{E}_{\omega\sim\widehat{\kappa}}\big[\exp(i\langle \omega, x - y\rangle)\big]$. The Random Fourier Features (RFF) (Rahimi & Recht, 2007; Sutherland & Schneider, 2015) approximate such kernels via Monte Carlo sampling. Drawing $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \overset{\text{i.i.d.}}{\sim} \widehat{\kappa}$, we define the following proxy RFF feature map:

$$\varphi_r(x) = \frac{1}{\sqrt{r}} \big[\cos(\boldsymbol{\omega}_\ell^\top x), \sin(\boldsymbol{\omega}_\ell^\top x)\big]_{\ell=1}^r,$$

which satisfies $\mathbb{E}[\varphi_r(x)^\top \varphi_r(y)] = k(x, y)$. Replacing $\phi(x)$ with $\varphi_r(x)$ yields a low-dimensional approximation

$$\widehat{C}_X \approx \frac{1}{n} \sum_{i=1}^n \varphi_r(x_i)\varphi_r(x_i)^\top \in \mathbb{R}^{2r\times 2r}.$$

3.2. Tensor-Product Kernels and Hadamard Joint Kernel Matrices

Let T and X denote the random variables of an input prompt and generated output, equipped with feature maps $\phi_T : \mathcal{T} \to \mathbb{R}^{d_t}$ and $\phi_X : \mathcal{X} \to \mathbb{R}^{d_x}$. We define the joint tensor-product feature map as

$$\phi_\otimes(t, x) = \phi_T(t) \otimes \phi_X(x) \in \mathbb{R}^{d_t} \otimes \mathbb{R}^{d_x},$$

where $\otimes$ denotes the tensor product of Hilbert spaces, which is for vectors $t = [t^{(1)}, \ldots, t^{(d_t)}] \in \mathbb{R}^{d_t}$ and $x = [x^{(1)}, \ldots, x^{(d_x)}] \in \mathbb{R}^{d_x}$, is defined as:

$$t\otimes x = \big[t^{(1)}x^{(1)}, .., t^{(1)}x^{(d_x)}, .., t^{(d_t)}x^{(1)}, .., t^{(d_t)}x^{(d_x)}\big] \in \mathbb{R}^{d_x d_t}$$

This representation captures multiplicative interactions between prompt and output embeddings and induces the product kernel

$$\begin{aligned} k_\otimes\big([t, x], [t', x']\big) &= \big\langle \phi_\otimes(t, x), \phi_\otimes(t', x')\big\rangle \\ &= k_T(t, t') \cdot k_X(x, x'). \end{aligned}$$

Given empirical samples $\{(t_i, x_i)\}_{i=1}^n$, the empirical joint kernel covariance is

$$C_{X\otimes T} = \frac{1}{n} \sum_{i=1}^n \phi_\otimes(t_i, x_i)\phi_\otimes(t_i, x_i)^\top.$$

Algorithm 1 PromptSplit: Kernel-based Formulation

Require: Datasets $\mathcal{D}_X = \{(t_i, x_i)\}_{i=1}^n$, $\mathcal{D}_Y = \{(t'_j, y_j)\}_{j=1}^m$; kernels k_T, k_X; parameter $\eta > 0$; eigenpair number R.

1: Build kernel blocks: K_{TT}, K_{XX} on $\mathcal{D}_X$; $K_{T'T'}, K_{YY}$ on $\mathcal{D}_Y$; and cross-blocks $K_{TT'}, K_{XY}$.
2: Form $K_{X,\eta Y|T}$ as in (3).
3: Compute the top R_+ positive eigenpairs $\{(\lambda_r, u_r)\}$ with $u_r = [u_{1:n}^{(r)}; u_{(n+1):(n+m)}(r)]$.
4: Construct eigenfunctions $v_r = \sum_{i=1}^n u_i^{(r)} \phi_X(t_i, x_i) + \sum_{j=1}^m u_{n+j}^{(r)} \phi_X(t'_j, y_j)$.
5: **Return** Positive eigvalues $\{\lambda_r\}$, eigenfunctions $\{v_r\}$.

Algorithm 2 Random Projection PromptSplit

Require: Datasets $\mathcal{D}_X, \mathcal{D}_Y$; explicit features ϕ_T, ϕ_X (or RFF maps); RP size r; parameter $\eta > 0$; number of eigenpairs R.

1: Draw Gaussian matrices $R_T \sim \mathcal{N}(0,1)^{d_t \times r}$ and $R_X \sim \mathcal{N}(0,1)^{d_x \times r}$; set $R = \frac{1}{r}(R_T \otimes R_X)$.
2: Compute sketched joint features: $\widetilde{\phi}_r(t_i, x_i) = R_T \phi_T(t_i) \odot R_X \phi_X(x_i)$ and $\widetilde{\phi}_r(t'_i, y_i) = R_T \phi_T(t'_i) \odot R_X \phi_X(y_i)$.
3: Form $\widehat{\Lambda}^{(r)} = \frac{1}{n}\sum_i \widetilde{\Phi}_{r,X} \widetilde{\Phi}_{r,X}^\top - \frac{\eta}{m}\sum_j \widetilde{\Phi}_{r,Y} \widetilde{\Phi}_{r,Y}^\top$.
4: Compute positive eigenpairs $(\widehat{\lambda}_r, \widehat{w}_r)$ of $\widehat{\Lambda}^{(r)}$.
5: **Return** Reduced-dimensional eigenpairs $\{(\widehat{\lambda}_r, \widehat{w}_r)\}$.

4. Method

We are given two prompt–response datasets produced by two generative systems,

$$\mathcal{D}_X = \{(t_i, x_i)\}_{i=1}^n, \qquad \mathcal{D}_Y = \{(t'_j, y_j)\}_{j=1}^m.$$

Each prompt t and output $z \in \{x, y\}$ is embedded and mapped to RKHSs via feature maps $\phi_T : \mathcal{T} \to \mathcal{H}_T$ and $\phi_X, \phi_Y : \mathcal{X} \to \mathcal{H}_X$ with normalized kernels $k_\mathcal{T} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ and $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying $k_T(t,t) = k_X(x,x) = 1$ for every $t \in \mathcal{T}$, $x \in \mathcal{X}$. Our goal is to identify and interpret where, in the joint prompt–output space, the two systems differ. When prompt distributions between the two datasets match $P_T = P_{T'}$, the analysis from (Zhang et al., 2024)suggests that the kernel matrices of the collected data can reveal the differences between the two conditional distributions $P_{X|T}$ and $P_{Y|T}$.

4.1. PromptSplit via Joint Kernel Covariance Difference

Consider the joint tensor feature $\phi_\otimes(t,x) = \phi_T(t) \otimes \phi_X(x)$. Then, the empirical joint covariances of the two datasets $\mathcal{D}_X$ and $\mathcal{D}_Y$ are defined as

$$\widehat{C}_{T\otimes X} = \frac{1}{n}\sum_{i=1}^n \phi_\otimes(t_i, x_i)\phi_\otimes(t_i, x_i)^\top,$$

$$\widehat{C}_{T\otimes Y} = \frac{1}{m}\sum_{j=1}^m \phi_\otimes(t'_j, y_j)\phi_\otimes(t'_j, y_j)^\top.$$

Definition 4.1. The covariance–difference operator between $\mathcal{D}_X$ and $\mathcal{D}_Y$ with hyperparameter $\eta > 0$ is defined as

$$\widehat{\Lambda}_{X,Y|T} = \widehat{C}_{T\otimes X} - \eta\, \widehat{C}_{T\otimes Y}, \tag{2}$$

As demonstrated in (Zhang et al., 2024), the eigendirections of the above matrix can reveal the differently expressed modes between the distributions $P_{T,X}$ and $P_{T',Y}$. We highlight that when the prompt marginal distributions P_T and P'_T of the two datasets are the same, the differences between the (prompt,output) joint distributions will reveal the differences between the conditional models $P_{X|T}$ and $P_{Y|T}$, that is precisely the aim of our comparative analysis. Therefore, in our analysis, we aim to efficiently compute the principal eigendirections of $\widehat{\Lambda}_{X,Y|T}$ for a sufficiently large dataset size n.

First, we observe that for the prompt and output feature dimensions d_t, d_x, the dimension of matrix $\widehat{C}_{X\otimes T}$ will be $d_t d_x \times d_t d_x$. As the standard embedding dimensions for input prompt and output visual data are usually lower bounded by several hundreds, e.g. 512 for CLIP embeddings, performing the eigendecomposition of $\widehat{\Lambda}_{X,Y|T}$ with even a simple linear kernel, would require the eigendecomposition of a matrix with at least a few hundreds of thousands rows, which would be infeasible on standard CPU and GPU processors.

To address the computatioanl challenge, we first apply the kernel trick and formulate a kernel matrix of size $2n \times 2n$ for n samples that share the eigenspectrum with $\widehat{\Lambda}_{X,Y|T}$. To do this, let K_{TT}, K_{XX} be prompt/output kernels on $\mathcal{D}_X$, and $K_{T'T'}, K_{YY}$ prompt/output kernels on $\mathcal{D}_Y$, and $K_{TT'}, K_{XY}$ be the cross-kernels. Considering the product kernel $k_\odot\big([t,x],[t',x']\big) = k_\mathcal{T}(t,t') \cdot k_\mathcal{X}(x,x')$, we define the following kernel matric $K_{X,\eta Y|T}$

$$K_{X,\eta Y|T} = \begin{bmatrix} \frac{1}{n} K_{TT} \odot K_{XX} & \frac{1}{\sqrt{nm}} K_{TT'} \odot K_{XY} \\ -\frac{\eta}{\sqrt{nm}} K_{TT'}^\top \odot K_{XY}^\top & -\frac{\eta}{m} K_{T'T'} \odot K_{YY} \end{bmatrix}. \tag{3}$$

Proposition 4.2. *The matrices $\widehat{\Lambda}_{X,Y|T}$ and $K_{X,\eta Y|T}$ share the same non-zero eigenvalues. Also, for every $K_{X,\eta Y|T}$'s eigenvector $u = [u_{1:n}; u_{(n+1):(n+m)}]$ of $K_{X,\eta Y|T}$ with non-zero eigenvalue λ, then the following v is the eigenvector of $\widehat{\Lambda}_{X,Y|T}$ for the same eigenvalue λ:*

$$v = \sum_{i=1}^n v_i \phi_\otimes(t_i, x_i) + \sum_{j=1}^m v_{n+j} \phi_\otimes(t'_j, y_j)$$

Proof. We defer the proof to the Appendix. □

The above proposition shows that given n, m samples in the two datasets, the cost of the eigendecomposition of $\widehat{\Lambda}_{X,Y|T}$ will grow at most as $\mathcal{O}\big((m+n)^3\big)$, where the upper-bound is independent of the kernel feature map size, as long as the kernel function can be efficiently evaluated.

4.2. Random Projection for Scalable PromptSplit

As discussed earlier, we can perform the eigendecomposition of $K_{X,\eta Y|T}$, with compute cost of $\mathcal{O}\big((n+m)^3\big)$ to compute the eigenspace of the target kernel covariance operator difference $\widehat{\Lambda}_{X,Y|T}$. However, when the dataset sizes grow beyong few tens of thousands, this approach becomes computationally infeasible.

To reduce the computational costs, in this section we propose a joint Gaussian random projection that approximately preserves the kernel covariance difference geometry while reducing the dimensionality of the feature map. Let $R_T \in \mathbb{R}^{d_T \times r}$, $R_X \in \mathbb{R}^{d_X \times r}$ with i.i.d. $\mathcal{N}(0,1)$ entries and then define

$$\widetilde{\phi}_r(x,t) = \frac{r}{\sqrt{d_T d_X}} R_T \phi_{\mathcal{T}}(t) \odot R_X \phi_{\mathcal{X}}(x) \in \mathbb{R}^r$$

Note that the computational cost of computing $\widetilde{\phi}_r(x,t)$ will be $\mathcal{O}\big(r(d_x+d_t)\big)$. Then, we propose to consider the kernel covariance difference operator with the joint feature map $\widetilde{\phi}_r$ to define:

$$\widetilde{\Lambda}_{r\,X,Y|T} = \widehat{C}_{\widetilde{\phi}_r(X,T)} - \eta\, \widehat{C}_{\widetilde{\phi}_r(Y,T)} \in \mathbb{R}^{r\times r}. \tag{4}$$

Note that the total computational cost of computing and performing the eigen decomposition of $\widetilde{\Lambda}_{r\,X,Y|T}$ wil be $\mathcal{O}\big((m+n)r(d_x+d_t)+r^3\big)$, which will grow linearly with $(m+n)(d_x+d_t)$ for a properly bounded random projection size r.

Furthermore, we note that in the case of a shift-invariant kernel, the above random projection can be unified with the mapping to the random Fourier feature space with the following definition of $\widetilde{\phi}_r(t,x)$ (for an even integer r) given the Fourier features $\omega_{x,1},\ldots,\omega_{x,r/2} \sim \widehat{\kappa}_x$ and $\omega_{t,1},\ldots,\omega_{t,r/2} \sim \widehat{\kappa}_t$ for the output and prompt kernels:

$$\widetilde{\phi}_r(t,x) = \big[\cos(\omega_{t,i}^\top t + \omega_{x,i}^\top x), \sin(\omega_{t,i}^\top t + \omega_{x,i}^\top x)\big]_{i=1}^{r/2} \in \mathbb{R}^r.$$

In the following, we prove that the eigendirections of the random projection $\widetilde{\Lambda}_{r\,X,Y|T}$ will lead to an $O(\sqrt{r})$-accurate approximation of the eigendirections of the matrix $\Lambda_{X,Y|T}$.

Theorem 4.3. *Consider the kernel covariance difference matrix $K_{X,\eta Y|T}$ and the proxy kernel covariance difference matrix $\widetilde{K}_{r\,X,\eta Y|T}$ of the random projection approach with dimension r. Then, for every $\delta > 0$, the following holds with probability at least $1-\delta$,*

$$\big\|\boldsymbol{\lambda}(K_{X,\eta Y|T}) - \boldsymbol{\lambda}(\widetilde{K}_{r\,X,\eta Y|T})\big\|_2 \leq \sqrt{\frac{8+8\eta^2}{r}}\big(1+\sqrt{2\log\tfrac{1}{\delta}}\big). \tag{5}$$

Proof. We defer the proof to the Appendix. □

5. PromptSplit Guidance for Text-Guided Diffusion Models

As discussed in the previous section, PromptSplit can identify differences in text-guided generative models. One application of this framework is to use it for guiding conditional diffusion models to align with a reference dataset, where we specifically focus on text-conditioned LDMs (Rombach et al., 2022).

Let $\mathrm{PromptSplit}(x_{1:n}, x_{1:n}|t_{1:n})$ denote the $\|\widehat{\Lambda}_{X,Y|T}\|_F^2$ in the latent space $\mathcal{Z}$ and $\text{Joint-Diversity}(x_{1:n}, t_{1:n})$ denote $\|C_{X\otimes T}\|_F^2$ which promotes diversity between samples with correlated prompts. At step τ, we augment the classifier-free update (Ho & Salimans, 2022) with an ascent step where $\eta_\tau > 0$ is the guidance scale at iteration τ, and ρ is the guidance scale of the diversity term:

$$\begin{aligned} \boldsymbol{z}_{\tau-1}^{(n)} \leftarrow\ & \mathrm{Sampler}\big(\boldsymbol{z}_\tau^{(n)},\, \hat{\epsilon}_\theta(\boldsymbol{z}_\tau^{(n)}, \tau, t_n)\big) \\ & - \eta_\tau \Big(\nabla_{\boldsymbol{z}^{(n)}} \mathrm{PromptSplit}(x_{1:n}, y_{1:n}|t_{1:n}) \\ & + \rho\, \nabla_{\boldsymbol{z}^{(n)}} \text{Joint-Diversity}(x_{1:n}, t_{1:n})\Big) \end{aligned} \tag{6}$$

6. Numerical Results

We evaluate PromptSplit across controlled and real prompt-conditioned generation tasks spanning text-to-image, text-to-text, and image captioning. Unless noted otherwise, images are embedded with DINOv2-giant (Oquab et al., 2024) and texts with Sentence-BERT (Reimers & Gurevych, 2019).

Models. For text-to-image we use SDXL (Podell et al., 2023), PixArt-Σ (Chen et al., 2024), and Kandinsky (Arkhipkin et al., 2024). For text-to-text we use Llama 3.2 (Aaron Grattafiori et al., 2024), Gemma 3 (Gemma-Team, 2025), DeepSeek-R1 (DeepSeek-AI et al., 2025), and Qwen 3 (Qwen-Team, 2025). For image captioning we use BLIP-2 (Li et al., 2023) and GPT-4o mini (team, 2024). We use Mistral 7B (Jiang et al., 2023) for automatic categorization.

Datasets. We use MNIST (LeCun et al., 2010) and MNIST-M (Ganin et al., 2016) for controlled validation, MS-COCO (Lin et al., 2015) for T2I comparisons, Ima-

Table 1. Per-mode distribution matching across tasks. We report $\text{MMD}^2 \times 10^2$ (RBF) for the top-3 PromptSplit modes together with the k-means baseline. Lower is more similar.

Task	Dataset	Test	Ref.	k-means	M1	M2	M3	M4
Text-to-Image	MS-COCO	SDXL	PixArt-Σ	0.82 ± 0.01	2.01	0.82	1.66	3.73
Text-to-Text	NQ-Open	Qwen 3	Gemma 3	1.24 ± 0.01	25.34	29.37	8.02	8.02
Image Captioning	ImageNet	BLIP-2	GPT-4o	1.43 ± 0.01	5.63	7.74	4.43	–

Table 2. Generalization of PromptSplit modes to held-out MS-COCO prompts for SDXL vs. PixArt-Σ. Entries report kernel similarity mean.

Split	k-means	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
Validation	0.47 ± 0.02	0.40 ± 0.02	0.40 ± 0.01	0.41 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	0.41 ± 0.01
Held-out	0.47 ± 0.02	0.41 ± 0.02	0.39 ± 0.01	0.40 ± 0.01	0.38 ± 0.01	0.41 ± 0.02	0.39 ± 0.01

Table 3. Runtime comparison (in seconds) of different PromptSplit algorithms on the MS-COCO.

Sample Size	Kernel Method	$r = 500$	$r = 1000$	$r = 2000$	$r = 3000$
1000	0.856	2.950	4.186	43.674	88.156
5000	111.89	3.13	5.39	40.29	162.83
10000	902.82	3.21	7.83	64.16	158.45
20000	—	3.57	12.03	77.39	188.62
30000	—	3.81	17.76	72.27	206.99

geNet (Russakovsky et al., 2014) subclasses for image captioning, and NQ-Open (Kwiatkowski et al., 2019) for LLMs.

Experimental Settings. We approximate the Gaussian kernel using $r = 3000$ random Fourier features and select bandwidth σ via the eigenvalue-gap heuristic of (Ospanov et al., 2024). Full algorithm details appear in Algorithms 1 and 2. Experiments are run on four RTX A5000 and two RTX 4090 GPUs.

6.1. Validation of PromptSplit in Settings with Known Ground Truth.

MNIST-M Color Disagreement. We construct a controlled benchmark using 20 prompts with 1000 images per prompt, shifting grayscale digits to colored prompts 5–9 in the test set and colored digits to grayscale prompts 0–4 in the reference set (Figure 2). The random-projection variant of PromptSplit recovers all ten planted disagreement prompts as the top-ranked modes, with near-perfect retrieval metrics in the table in Figure 2. Additional controlled validation for text-to-image models and LLMs appears in the appendix (Figures 7, 8, 9). In both settings, PromptSplit captures the shifted modes while assigning lower importance to prompt clusters where the two systems are designed to agree.

6.2. Application of PromptSplit in Real World Settings Across Tasks.

NQ-Open Experiment. We generate answers for 20K NQ-Open validation questions with Qwen 3 as test and Gemma 3 as reference. PromptSplit identifies high-eigenvalue modes centered on questions about actors and U.S. presidents, where the two models diverge most, and a low-eigenvalue mode where their answers are relatively similar (Figure 1). Additional LLM pair comparisons appear in appendix.

ImageNet. We apply PromptSplit to image captioning on three visually distinct ImageNet subclasses, comparing BLIP-2 (test) against GPT-4o mini (reference). The discovered modes are semantically coherent at the class level and reveal systematic captioning style differences. (Figure 17).

MSCOCO. We apply PromptSplit to 30K MS-COCO captions using SDXL, PixArt-Σ, and Kandinsky. The leading modes reveal prompt families where models diverge in style, composition, object placement, and prompt alignment. Additional model pair comparisons appear in Figures 18–21.

To further show the disagreement in top detected modes numerically, we calculated per cluster MMD^2 in table 1. Top identified PromptSplit modes show more disagreement than mean of top distinct modes identified by a heuristic K-Means. Sample wise kernel similarities are reported in appendix.

6.3. Generalization to Held-Out Prompts

We test whether discovered modes transfer beyond the split on which they are found. In the SDXL–PixArt-Σ setting, modes discovered on the validation split retain nearly identical kernel-similarity statistics when re-evaluated on 30K held-out training prompts (Table 2), confirming that PromptSplit captures reusable disagreement patterns rather than split-specific artifacts.

6.4. Automatic Categorization

To interpret discovered modes, we extract TF-IDF cluster keywords and use Mistral 7B (Jiang et al., 2023) to generate short semantic summaries of the highest-ranked prompt

Figure 3. Qualitative comparison of reference set and PS-guided image generation with SDXL.

groups. On both NQ-Open (tables 7, 8) and MS-COCO (tables 9, 10), the generated summaries are consistent with the retrieved prompts, turning identified modes into human-readable disagreement categories without manual inspection.

6.5. PromptSplit Guidance for Distribution Matching in LDMs

We use PromptSplit to guide SDXL in the latent diffusion process (Section 5), using 100 reference samples per painting style. As shown in Figure 3, PS-guided generation improves alignment with the reference distribution both qualitatively and quantitatively. Additional results appear in the appendix.

7. Conclusion

In this work, we introduced PromptSplit, a prompt-aware spectral framework for comparing prompt-guided generative models through the eigenspectrum of joint prompt–output kernel covariance differences. By coupling prompt and output representations via tensor-product embeddings, PromptSplit enables a structured analysis of where and how model behaviors diverge as a function of the input prompt, moving beyond aggregate quality metrics toward prompt-conditioned comparison. We proposed a kernel formulation that admits an efficient spectral implementation, provided theoretical guarantees for a scalable random-projection approximation, and demonstrated the effectiveness of the approach on both synthetic settings and real text-to-image and text-to-text generative models.

PromptSplit is designed as an embedding-based, second-order spectral method, and its analysis reflects the information captured by the chosen prompt and output representations. While this design enables scalability and interpretability, it does not aim to characterize all higher-order aspects of conditional generative behavior. The random-projection scheme introduces a standard accuracy–efficiency trade-off that is controlled by the projection dimension and can be tuned in practice. The current formulation focuses on pairwise model comparison under matched prompt distributions; extending the framework to multi-model comparisons, adaptive projection strategies, or evolving model families represents natural directions for future work.

References

Aaron Grattafiori, A. D. et al. The llama 3 herd of models, 2024.

Arkhipkin, V. et al. Kandinsky 3: Text-to-image synthesis for multifunctional and high-quality generation. *arXiv preprint arXiv:2410.21061*, 2024.

Babbar, V., Guo, Z., and Rudin, C. What is different between these datasets? a framework for explaining data distribu-

tion shifts. *Journal of Machine Learning Research*, 26 (180):1–64, 2025.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017.

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. GAN dissection: Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Bi, X., Chen, D., Chen, G., et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1877–1901, 2020.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., and Li, Z. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024.

Chitta, R., Jin, R., Havens, T. C., and Jain, A. K. Approximate kernel k-means: Solution to large scale kernel clustering. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 895–903, 2011.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J.-M., et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081): 633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z.

Friedman, D., Blei, D. M., and Dieng, A. B. Be more diverse with the most diverse: Vendi score for diversity evaluation of text. *Transactions of the Association for Computational Linguistics*, 10:1155–1172, 2022.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016.

Gemini-Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma-Team. Gemma 3 technical report, 2025.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. GANSpace: Discovering interpretable GAN controls. *Advances in Neural Information Processing Systems*, 33: 9841–9850, 2020.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 7514–7528, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., and Salimans, T. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Jalali, M., Li, C. T., and Farnia, F. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9931–9943, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018.

Kulinski, S. and Inouye, D. I. Towards explaining image-based distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4788–4792, June 2022.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

LeCun, Y., Cortes, C., and Burges, C. J. C. Mnist handwritten digit database. *ATT Labs [Online].*, 2010.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Naeem, M. F., Chung, S., Lim, J. H., Mo, S., and Moon, I.-C. Reliable fidelity and diversity metrics for generative models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13464–13474, 2020.

Oquab, M., Darcet, T., Moutakanni, T., Fedi, J., Szafraniec, M., Stock, P., Joulin, A., Bojanowski, P., Douze, M., and Massa, F. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Special Issue on Foundation Models.

Ospanov, A., Zhang, J., Jalali, M., Cao, X., Bogdanov, A., and Farnia, F. Towards a scalable reference-free evaluation of generative models. In *Advances in Neural Information Processing Systems*, 2024.

Pham, N. D. and Pagh, R. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 239–247, 2013.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

Qwen-Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Qwen-Team. Qwen3 technical report, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, nov 2019. URL https://arxiv.org/abs/1908.10084.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., and Taigman, Y. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Stolte, M., Kappenberg, F., Rahnenführer, J., and Bommert, A. Methods for quantifying dataset similarity: a review, taxonomy and comparison. *Statistics Surveys*, 18:1–118, 2024.

Sutherland, D. J. and Schneider, J. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015.

Sutherland, D. J., Strathmann, H., Arbel, M., and Gretton, A. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pp. 652–660. PMLR, 2018.

Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

team, O. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Zhang, J., Li, C. T., and Farnia, F. An interpretable evaluation of entropy-based novelty of generative models. *Proceedings of Machine Learning Research (ICML 2024)*, 2024. Also available as arXiv:2402.17287.

Zhang, J., Jalali, M., Li, C. T., and Farnia, F. Unveiling differences in generative models: A scalable differential clustering approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

A. Proofs

A.1. Proof of Proposition 4.2

We start by writing the empirical covariance–difference operator explicitly. By definition,

$$\widehat{\Lambda}_{X,Y|T} = \frac{1}{n}\sum_{i=1}^{n} \phi_{\otimes}(t_i, x_i)\phi_{\otimes}(t_i, x_i)^{\top} - \frac{\eta}{m}\sum_{j=1}^{m} \phi_{\otimes}(t'_j, y_j)\phi_{\otimes}(t'_j, y_j)^{\top}.$$

Consider the following stacked feature matrices:

$$\Psi_X = \begin{bmatrix} \phi_{\otimes}(t_1, x_1)^{\top} \\ \vdots \\ \phi_{\otimes}(t_n, x_n)^{\top} \end{bmatrix} \in \mathbb{R}^{n\times D}, \qquad \Psi_Y = \begin{bmatrix} \phi_{\otimes}(t'_1, y_1)^{\top} \\ \vdots \\ \phi_{\otimes}(t'_m, y_m)^{\top} \end{bmatrix} \in \mathbb{R}^{m\times D},$$

where $D = d_T d_X$. Then, we can write

$$\widehat{\Lambda}_{X,Y|T} = \Psi^{\top} \begin{bmatrix} \frac{1}{n} I_n & 0 \\ 0 & -\frac{\eta}{m} I_m \end{bmatrix} \Psi, \qquad \Psi = \begin{bmatrix} \Psi_X \\ \Psi_Y \end{bmatrix}.$$

Next, consider the corresponding kernel matrix

$$K_{X,\eta Y|T} = \begin{bmatrix} \frac{1}{n} K_{TT} \odot K_{XX} & \frac{1}{\sqrt{nm}} K_{TT'} \odot K_{XY} \\ -\frac{\eta}{\sqrt{nm}} K_{TT'}^{\top} \odot K_{XY}^{\top} & -\frac{\eta}{m} K_{T'T'} \odot K_{YY} \end{bmatrix}.$$

Using the tensor-product identity $\langle \phi_T(t) \otimes \phi_X(x),\, \phi_T(t') \otimes \phi_X(x') \rangle = k_T(t, t')\, k_X(x, x')$, we observe that

$$K_{X,\eta Y|T} = \begin{bmatrix} \Psi_X \\ \Psi_Y \end{bmatrix} \begin{bmatrix} \frac{1}{n} I_n & 0 \\ 0 & -\frac{\eta}{m} I_m \end{bmatrix} \begin{bmatrix} \Psi_X \\ \Psi_Y \end{bmatrix}^{\top} = \Psi \begin{bmatrix} \frac{1}{n} I_n & 0 \\ 0 & -\frac{\eta}{m} I_m \end{bmatrix} \Psi^{\top}.$$

Hence, the following equations hold

$$\widehat{\Lambda}_{X,Y|T} = \Psi^{\top} D \Psi, \qquad K_{X,\eta Y|T} = \Psi D \Psi^{\top}, \quad D = \operatorname{diag}\big(\tfrac{1}{n} I_n, -\tfrac{\eta}{m} I_m\big).$$

It is a standard linear-algebra fact that the matrices $\Psi^{\top} D \Psi$ and $\Psi D \Psi^{\top}$ have the same nonzero eigenvalues (due to flipped multiplication order). Therefore, $\widehat{\Lambda}_{X,Y|T}$ and $K_{X,\eta Y|T}$ share the same nonzero spectrum. Finally, let $u = [u_{1:n}; u_{(n+1):(n+m)}]$ be an eigenvector of $K_{X,\eta Y|T}$ with eigenvalue $\lambda \neq 0$. Define vector v as follows:

$$v = \Psi^{\top} u = \sum_{i=1}^{n} u_i\, \phi_{\otimes}(t_i, x_i) + \sum_{j=1}^{m} u_{n+j}\, \phi_{\otimes}(t'_j, y_j).$$

Then, we have the following

$$\widehat{\Lambda}_{X,Y|T} v = \Psi^{\top} D \Psi \Psi^{\top} u = \Psi^{\top} D (K_{X,\eta Y|T} u) = \lambda\, \Psi^{\top} u = \lambda v,$$

which proves that v is an eigenvector of $\widehat{\Lambda}_{X,Y|T}$ associated with the same eigenvalue λ. This completes the proof.

A.2. Proof of Theorem 4.3

First, we show the unbiasedness and boundedness of the random-feature product kernel. By Bochner's theorem, for shift-invariant $k_T(t, t') = \kappa_T(t - t')$ and $k_X(x, x') = \kappa_X(x - x')$, there exist spectral measures μ_T, μ_X such that

$$k_T(t, t') = \mathbb{E}_{\omega_t \sim \mu_T}\big[\cos(\omega_t^{\top}(t - t'))\big], \qquad k_X(x, x') = \mathbb{E}_{\omega_x \sim \mu_X}\big[\cos(\omega_x^{\top}(x - x'))\big].$$

Using independence between $\omega_{t,\ell}$ and $\omega_{x,\ell}$, and the identity $\cos(a)\cos(b) = \frac{1}{2}\cos(a+b)+\frac{1}{2}\cos(a-b)$, we may equivalently use the single cosine feature to represent the product kernel:

$$\begin{aligned}\mathbb{E}\Big[g_\ell\big((t,x),(t',x')\big)\Big] &= \mathbb{E}_{\omega_t,\omega_x}\Big[\cos\big(\omega_t^\top(t-t') + \omega_x^\top(x-x')\big)\Big] \\ &= \mathbb{E}_{\omega_t}\big[\cos(\omega_t^\top(t-t'))\big]\ \mathbb{E}_{\omega_x}\big[\cos(\omega_x^\top(x-x'))\big] \\ &= k_T(t,t')\,k_X(x,x').\end{aligned} \tag{7}$$

Moreover, for every ℓ and every pair of inputs,

$$\big|g_\ell\big((t,x),(t',x')\big)\big| \le 1, \qquad \big|\widetilde{k}_\otimes\big((t,x),(t',x')\big)\big| \le 1. \tag{8}$$

Next, we decompose the matrix as an average of i.i.d. bounded matrices. To do this, define $N := n+m$, and for every $\ell \in [r]$, consider the *single-feature* block matrix $K^{(\ell)}_{X,\eta Y|T} \in \mathbb{R}^{N\times N}$ by

$$K^{(\ell)}_{X,\eta Y|T} = \begin{bmatrix} \frac{1}{n}G^{(\ell)}_{XX} & \frac{1}{\sqrt{nm}}G^{(\ell)}_{XY} \\ -\frac{\eta}{\sqrt{nm}}(G^{(\ell)}_{XY})^\top & -\frac{\eta}{m}G^{(\ell)}_{YY}\end{bmatrix}, \tag{9}$$

where the entries are

$$\begin{aligned}\big(G^{(\ell)}_{XX}\big)_{ij} &:= g_\ell\big((t_i,x_i),(t_j,x_j)\big), && i,j\in[n],\\ \big(G^{(\ell)}_{YY}\big)_{jj'} &:= g_\ell\big((t'_j,y_j),(t'_{j'},y_{j'})\big), && j,j'\in[m],\\ \big(G^{(\ell)}_{XY}\big)_{ij} &:= g_\ell\big((t_i,x_i),(t'_j,y_j)\big), && i\in[n],\ j\in[m].\end{aligned}$$

By construction of $\widetilde{k}_\otimes$, the proxy matrix is the average of these blocks:

$$\widetilde{K}_{r\,X,\eta Y|T} = \frac{1}{r}\sum_{\ell=1}^{r} K^{(\ell)}_{X,\eta Y|T}. \tag{10}$$

Taking expectation and using (7) entrywise shows that

$$\mathbb{E}\Big[K^{(\ell)}_{X,\eta Y|T}\Big] = K_{X,\eta Y|T}. \tag{11}$$

From (8), every entry of $G^{(\ell)}_{XX}, G^{(\ell)}_{YY}, G^{(\ell)}_{XY}$ has magnitude at most 1. Hence,

$$\left\|\frac{1}{n}G^{(\ell)}_{XX}\right\|_F^2 \le \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{n}\right)^2 = 1, \tag{12}$$

$$\left\|\frac{1}{\sqrt{nm}}G^{(\ell)}_{XY}\right\|_F^2 \le \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{1}{\sqrt{nm}}\right)^2 = 1, \tag{13}$$

$$\left\|\frac{\eta}{m}G^{(\ell)}_{YY}\right\|_F^2 \le \sum_{j=1}^{m}\sum_{j'=1}^{m}\left(\frac{\eta}{m}\right)^2 = \eta^2. \tag{14}$$

Using the matrix formulation, we obtain the following

$$\begin{aligned}\big\|K^{(\ell)}_{X,\eta Y|T}\big\|_F^2 &\le \left\|\frac{1}{n}G^{(\ell)}_{XX}\right\|_F^2 + \left\|\frac{1}{\sqrt{nm}}G^{(\ell)}_{XY}\right\|_F^2 + \left\|\frac{\eta}{\sqrt{nm}}(G^{(\ell)}_{XY})^\top\right\|_F^2 + \left\|\frac{\eta}{m}G^{(\ell)}_{YY}\right\|_F^2 \\ &\le 1+1+\eta^2+\eta^2 = 2(1+\eta^2)\end{aligned} \tag{15}$$

Therefore,

$$\big\|K^{(\ell)}_{X,\eta Y|T}\big\|_F \le \sqrt{2+2\eta^2} \tag{16}$$

The same argument applied to the expectation in (11) yields $\|K_{X,\eta Y|T}\|_F \le \sqrt{2+2\eta^2}$, and thus, by the triangle inequality,

$$\left\|K^{(\ell)}_{X,\eta Y|T} - K_{X,\eta Y|T}\right\|_F \le \left\|K^{(\ell)}_{X,\eta Y|T}\right\|_F + \left\|K_{X,\eta Y|T}\right\|_F \le \sqrt{8+8\eta^2} \tag{17}$$

Now, we consider the Hilbert space $(\mathbb{R}^{N\times N}, \langle\cdot,\cdot\rangle_F)$. Consider the centered random matrices

$$X_\ell := K^{(\ell)}_{X,\eta Y|T} - K_{X,\eta Y|T}, \qquad \ell = 1,\dots,r.$$

Then, $X_1,\dots,X_r$ are i.i.d. random vectors satisfying $\mathbb{E}[X_\ell] = 0$, and by (17), $\|X_\ell\|_F \le 4$ almost surely. A Hoeffding inequality for Hilbert-space-valued random vectors ((Sutherland et al., 2018, Lemma 11)) implies that, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\|\frac{1}{r}\sum_{\ell=1}^{r} X_\ell\right\|_F \le \frac{4}{\sqrt{r}}\left(1+\sqrt{2\log\frac{1}{\delta}}\right). \tag{18}$$

Since $\frac{1}{r}\sum_{\ell=1}^{r} X_\ell = \widetilde{K}_{r\,X,\eta Y|T} - K_{X,\eta Y|T}$ by (10) and $\mathbb{E}[K^{(\ell)}] = K$, we obtain

$$\left\|\widetilde{K}_{r\,X,\eta Y|T} - K_{X,\eta Y|T}\right\|_F \le \sqrt{\frac{8+8\eta^2}{r}}\left(1+\sqrt{2\log\tfrac{1}{\delta}}\right). \tag{19}$$

Since both $K_{X,\eta Y|T}$ and $\widetilde{K}_{r\,X,\eta Y|T}$ are symmetric, the Hoffman–Wielandt inequality gives

$$\left\|\boldsymbol{\lambda}(K_{X,\eta Y|T}) - \boldsymbol{\lambda}(\widetilde{K}_{r\,X,\eta Y|T})\right\|_2 \le \left\|K_{X,\eta Y|T} - \widetilde{K}_{r\,X,\eta Y|T}\right\|_F. \tag{20}$$

Combining these two inequalities proves the stated inequality.

Table 4. Quantitative evaluation of PromptSplit-guided generation against unguided SDXL. Lower Fréchet Distance (FD) and Kernel Distance (KD) are better, while higher density and coverage are better.

Model	**FD**	**KD**	**Density** ($\times 10^2$)	**Coverage** ($\times 10^2$)
SDXL (No Guidance)	1723.43	3.73	1.82	28.25
SDXL + PromptSplit (Ours)	1564.35	2.92	1.64	62.92

B. Additional Numerical Results

B.1. Additional Guidance Results.

We additionally evaluate whether the disagreement modes identified by PromptSplit can be used not only for analysis, but also to guide generation toward a reference style. In this experiment, we form small reference sets corresponding to two distinct semantic-stylistic groups: Van Gogh-style paintings of dogs and chalkboard-style monochrome scenes featuring butterflies and cages. We then compare standard SDXL samples against SDXL samples generated with PromptSplit-based guidance toward the selected reference mode.

Appendix Figure 4 provides a qualitative comparison. In both settings, PromptSplit guidance shifts the generated samples toward the target reference distribution: in the Van Gogh-style case, the guided samples better reflect the painterly texture, color palette, and composition of the reference images, while in the chalkboard-style case they more faithfully capture the monochrome aesthetic and recurring butterfly-and-cage motifs. By contrast, unguided SDXL generations often preserve only coarse object identity while drifting away from the reference style.

We also quantify this effect in Table 4. PromptSplit guidance improves both Fréchet Distance and Kernel Distance relative to unguided SDXL, and substantially increases coverage, while density remains comparable. Together, these qualitative and quantitative results suggest that PromptSplit is not merely an evaluation method, but can serve as useful control signals for steering generation toward a chosen reference distribution.

B.2. Low Ranks Disagreement Modes.

The most prominent PromptSplit modes correspond to the largest eigenvalues and therefore capture the strongest disagreement directions between the test and reference models. As a result, the highest-ranked examples often exhibit visually obvious differences. To illustrate that PromptSplit is not limited to these dominant cases, we also inspect lower-ranked modes from the top 100 eigendirections for the SDXL versus PixArt comparison on MS-COCO val2014.

Figure 5 shows representative lower-ranked modes. Although the disagreements are less pronounced than in the leading modes, the selected prompts still organize into coherent semantic clusters, including teddy bears, brown bears, pizza scenes, tennis players, and horses. In these cases, the differences between SDXL and PixArt are often expressed through more subtle factors such as the number of objects, scene composition, background structure, local style, or the relative emphasis placed on different visual elements.

These examples clarify the interpretation of the eigenvalue ranking. Large eigenvalues correspond to the strongest and most visually salient disagreement directions, whereas smaller positive eigenvalues reveal weaker but still structured differences that remain semantically meaningful. This supports the view that PromptSplit recovers a spectrum of disagreement modes, ranging from highly distinctive behaviors to more nuanced and reference-aligned variations.

B.3. Occupation Level Disagreement Identification.

Figure 6 presents a controlled occupation-level comparison between SDXL (reference) and PixArt-Σ (test). We consider nine occupation prompt clusters, each containing several closely related prompt variants, and generate 500 images per cluster for each model. Since the setup is deliberately structured around occupation groups, the leading PromptSplit eigenmodes align closely with these clusters, and the eigenspectrum exhibits a sharp decay beyond the dominant directions. Among the identified modes, Nurse shows the strongest disagreement, while Carpenter and Teacher reveal interpretable differences associated with age or gender presentation; by contrast, Judge exhibits only a comparatively weak discrepancy, consistent with its smaller eigenvalue. These findings are further supported by the t-SNE visualization of image embeddings, which shows coherent occupation-level structure aligned with the disagreement patterns recovered by PromptSplit.

B.4. Image Style Disagreement.

In figure 1 we generated three clusters of prompts and sampled 1000 images per cluster with different styles. For test dataset we used oil painting style for cityscape images, and pop art style for forests which differs from photo realistic style images generated for reference. We also generated photo realistic mountain images for both test and reference datasets. PromptSplit kernel-based algorithm successfully identified the two distinct modes in style with significant larger eigenvalues.

B.5. Answer Distribution Disagreement Detection

We compare Llama 3.2 (Test) and Gemma 3 (Reference) on three prompt clusters (Figure 8 for celebrity, American city, and fast modern fuel car, using 100 prompts per cluster and 10,000 generated answers per cluster (30,000 prompt–answer pairs per model). The per-cluster answer dispersions show a clear difference in answer generation: Gemma collapses strongly onto one or two dominant entities in all clusters, while Llama's answers are more distributed across multiple names. Applying PromptSplit to the aggregated prompt–answer pairs yields a small number of dominant test-dominant disagreement directions, and the leading samples for these modes isolate interpretable, cluster-specific differences (top three modes respectively shows entities in celebrities, modern cars, and cities that are most different from the reference set answers) which aligns with variations in prompt/answer behaviors visible in the dispersions.

B.6. Controlled Answers Disagreement.

To validate that PromptSplit can recover known differences, we construct a controlled text-to-text setting (Figure 9) where both "reference" and "test" outputs are produced by the same DeepSeek-r1 (DeepSeek-AI et al., 2025) model, but with different answer-control constraints. Prompts are partitioned into three clusters, and the control is intentionally applied so that the test condition differs from the reference primarily in the movie-genre and historical-phenomena clusters, while the third cluster is a control where reference and test are intended to match. We apply PromptSplit by embedding prompts and answers, constructing a joint prompt–answer similarity structure, and computing the leading eigen-directions of the covariance-difference operator, where large positive eigenvalues indicate the most prominent test-dominant disagreement modes. The resulting eigenspectrum shows two dominant positive modes, consistent with the two clusters where answers were deliberately shifted, and inspecting the leading samples for these modes demonstrates human-interpretable differences (Horror w.r.t Action and Poleponnesian War w.r.t The Fall of Roman Empire) that align with the intended controlled shifts.

B.7. Ablation Study on Number of Samples.

To investigate the effect of dataset size on the quality and stability of the detected disagreement modes, we evaluate the random-projection variant of PromptSplit across varying numbers of prompt–output pairs. We use the MS-COCO 2014 validation set (30k captions) as the base and create subsampled versions with $n = m \in \{5k, 10k, 20k, 30k, 50k, 90k\}$ pairs per model/dataset. All experiments employ the same hyperparameters (r = 3000, η = 1, DINOv2-giant + Sentence-BERT embeddings, bandwidth selected via eigenvalue-gap heuristic). Figure X (top row) shows the top-10 eigenvalues and the strongest-attributed prompts/images for the pairwise comparison SDXL vs. PixArt-Σ across different sample sizes. We observe that:

- With as few as 5k–10k samples, the method already identifies the dominant disagreement modes (e.g., tennis-court geometry, plate placement, ramp/sideboard composition), with the largest eigenvalues corresponding to semantically consistent prompt clusters, however there are minor flaws in some modes.
- At 30k samples (the full MS-COCO val2014 size), eigenvalue magnitudes stabilize, and the ranking of top modes becomes highly consistent with the full-dataset run. Beyond 30k (up to 90k subsampled with replacement from the same caption pool), further gains are marginal: the top 3–5 modes remain nearly identical in attributed prompts and visual patterns, while smaller modes show minor fluctuations due to sampling variance.
- Beyond 30k (up to 90k generated with different seeds from the same 30K caption), further gains are marginal: the top 3–5 modes remain nearly identical in attributed prompts and visual patterns, while smaller modes show minor fluctuations due to added samples disagreements.

These results indicate that PromptSplit is robust even at moderate sample sizes and 30k pairs are typically sufficient to reliably recover the principal directions of prompt-dependent disagreement in real-world text-to-image settings. Importantly,

Table 5. Sensitivity of PromptSplit to the guidance parameter η on the MS-COCO text-to-image experiment (SDXL as test, PixArt-Σ as reference). We report MMD^2 $\times 10^2$ values for the discovered modes and the k-means baseline. Lower values indicate more similar distributions.

η	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6	Data
1	2.01	0.82	1.66	3.73	1.46	1.55	0.82 ± 0.02
2	1.54	1.34	0.62	2.10	6.60	1.76	0.82 ± 0.02
3	1.74	1.39	1.96	2.15	6.64	1.10	0.82 ± 0.02
4	1.85	1.42	2.05	1.42	6.52	1.97	0.82 ± 0.02
5	1.45	2.12	2.12	1.09	0.97	1.16	0.82 ± 0.02
1000	1.28	1.37	—	—	—	—	0.82 ± 0.02

the kernel-based (non-projected) formulation becomes computationally prohibitive beyond 10K-15K samples ($O((n+m)^3)$ scaling), showing the necessity of the random-projection approximation for large-scale analysis. Runtime measurements (single RTX A5000 GPU) are reported in Table 3.

B.8. Ablation Study on Number of random Fourier features.

We evaluate projection dimension $r \in \{800, 1500, 2000, 3000, 5000\}$ on the full 30k-pair SDXL vs. PixArt-Σ comparison (η = 1, DINOv2-giant + Sentence-BERT embeddings, bandwidth selected via eigenvalue-gap heuristic). At r = 800, primary modes are recovered but eigenvalues are compressed and smaller modes noisier. By r = 1500–2000, eigenvalue magnitudes increase, mode ranking stabilizes, and prompt attribution sharpens. At r = 3000 (default), top modes match those at r = 5000 almost identically in spectrum shape, prompt semantics, and visual patterns.

B.9. Ablation Study on η.

The parameter η controls the relative contribution of the reference covariance in the PromptSplit operator, and therefore directly affects the eigenspectrum and the number of positive disagreement modes. In the main experiments, we use the unbiased choice $\eta = n/m$, which reduces to $\eta = 1$ when the test and reference sets have the same size. To assess the sensitivity of the method to this choice, we perform an ablation on the MS-COCO text-to-image setting with SDXL as the test model and PixArt-Σ as the reference model.

Table 5 reports MMD^2 values for the discovered modes under different values of η, while the corresponding figures 14, 15 visualize the qualitative behavior of the resulting eigenspectrum and selected modes. Across moderate choices of $\eta \in \{1, 2, 3, 4, 5\}$, the leading disagreement modes remain broadly stable, and several modes continue to exhibit substantially larger discrepancy than the baseline. This indicates that the main semantic disagreement structure identified by PromptSplit is not an artifact of a single hyperparameter choice.

At the same time, the results show that very large values of η change the spectrum qualitatively. In particular, when $\eta = 1000$, only a small number of positive modes remain, indicating that the reference covariance dominates the comparison and suppresses weaker disagreement directions. This behavior is expected from the definition of the operator and supports using $\eta = n/m$ as a practical default: it preserves the dominant disagreement modes while avoiding the collapse observed at excessively large values of η.

B.10. Ablation Study on different image embeddings.

PromptSplit is constructed using DINOv2 features in our main MS-COCO text-to-image experiments. To evaluate whether the discovered disagreement modes depend strongly on this particular representation, we re-evaluate the selected modes using Gaussian kernels built from CLIP and SigLIP embeddings, without changing the original PromptSplit decomposition.

Table 6 reports the corresponding validation scores for the k-means baseline and the leading PromptSplit modes, and the accompanying figure 16 visualizes violin plots of per sample kernel similarities. The overall pattern remains consistent across DINOv2, CLIP, and SigLIP: the PromptSplit modes identified in DINOv2 space remain distinct and structured when

Table 6. Embedding ablation on the MS-COCO text-to-image experiment (SDXL as test, PixArt-Σ as reference). PromptSplit uses DINOv2 embeddings in the main experiments; here we compare mean validation scores across DINOv2, CLIP, and SigLIP 2 for the k-means baseline and the top-5 PromptSplit modes.

Embedding	k-means	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
DINOv2	0.47 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	0.40 ± 0.01
CLIP	0.69 ± 0.01	0.67 ± 0.01	0.66 ± 0.01	0.67 ± 0.01	0.66 ± 0.01	0.67 ± 0.01
SigLIP 2	0.63 ± 0.01	0.61 ± 0.01	0.59 ± 0.01	0.61 ± 0.01	0.57 ± 0.01	0.61 ± 0.01

measured using alternative embeddings. This suggests that the extracted disagreement directions are not merely artifacts of the discovery representation, but reflect more stable semantic differences between the test and reference models.

At the same time, the absolute scores vary across embedding families, which is expected because DINOv2, CLIP, and SigLIP emphasize different aspects of image similarity. We therefore do not expect numerical agreement across embeddings. Rather, the purpose of this ablation is to show that the qualitative separation of the disagreement modes persists under independent feature spaces, providing evidence that PromptSplit captures higher-order differences that transfer beyond the embedding used for discovery.

B.11. Image captioning on ImageNet subclasses.

To further test whether PromptSplit identifies meaningful disagreement modes beyond text-to-image and text-to-text settings, we apply it to an image-captioning scenario on three visually distinct ImageNet subclasses. In this experiment, BLIP-2 is treated as the test model and GPT-4o mini as the reference model. PromptSplit is applied to the paired caption outputs, and the resulting modes are visualized in Figure 17.

The discovered modes remain semantically coherent at the class level. As shown in the figure, PromptSplit groups together images of vintage gas pumps, cassette players and radio equipment, and church interiors, indicating that the method is able to recover structured captioning disagreements tied to consistent visual themes. This supports the view that the operator is not simply separating random examples, but identifying subsets where model behavior differs systematically.

The model outputs also reveal characteristic differences in caption style and content. BLIP-2 tends to produce shorter and more template-like captions centered on the most salient object category, whereas GPT-4o mini typically generates longer and more detailed descriptions, often including attributes, material details, scene context, and finer-grained architectural or object-level cues. These examples illustrate that PromptSplit can surface not only object-category differences, but also higher-order variation in descriptive specificity and semantic emphasis between captioning models.

B.12. Disagreement Identification on MS-COCO Captions.

We generated images using three state-of-the-art text-to-image models—Stable Diffusion XL (SDXL), PixArt-Σ, and Kandinsky on the 30,000 captions from the MS-COCO 2014 validation set. Due to the large dataset size, we applied the scalable random-projection variant of PromptSplit (with projection dimension $r = 3000$) to identify prompt clusters that induce systematic behavioral disagreements in the generated images.

We performed pairwise comparisons between each pair of models (SDXL vs. PixArt-Σ, SDXL vs. Kandinsky, PixArt-Σ vs. Kandinsky) as well as comparisons of each model against the real MS-COCO validation images (serving as a reference distribution). For each comparison, the top 5 identified modes of disagreement with the largest positive eigenvalues. Each mode is illustrated a selection of generated (or real reference) images.

These visualizations highlight interpretable prompt families where the models diverge in visual style, composition, object placement, realism, or alignment with the input text, complementing aggregate metrics such as FID or CLIPScore that do not capture prompt-conditioned differences.

B.13. Additional NQ-Open Experiments.

Figures 23, 24 extend the NQ-Open study to additional LLM pairs and show that PromptSplit isolates prompt-conditioned disagreement in a consistent, interpretable way across comparisons. In each figure, we report the leading positive eigenmodes and we qualitatively inspect the highest-attribution question clusters to reveal the prompt families that most strongly separate

the two models' behaviors. Across pairs, the dominant modes typically correspond to coherent topical categories (e.g., particular entity-centric question types), where the models diverge in answer selection, factual framing, or phrasing. Importantly, alongside these high-disagreement directions, we also include a representative low-eigenvalue (or "similar") mode which shows questions that receive weak attribution under PromptSplit and yield largely aligned answers across the two models.

Figure 4. Additional Qualitative comparison of reference set and PS-guided image generation with SDXL.

Figure 5. Representative lower-ranked PromptSplit modes for SDXL versus PixArt on MS-COCO val2014, where disagreements are subtler than in the leading modes but still organize into coherent clusters reflecting style, number of objects, and other minor visual differences.

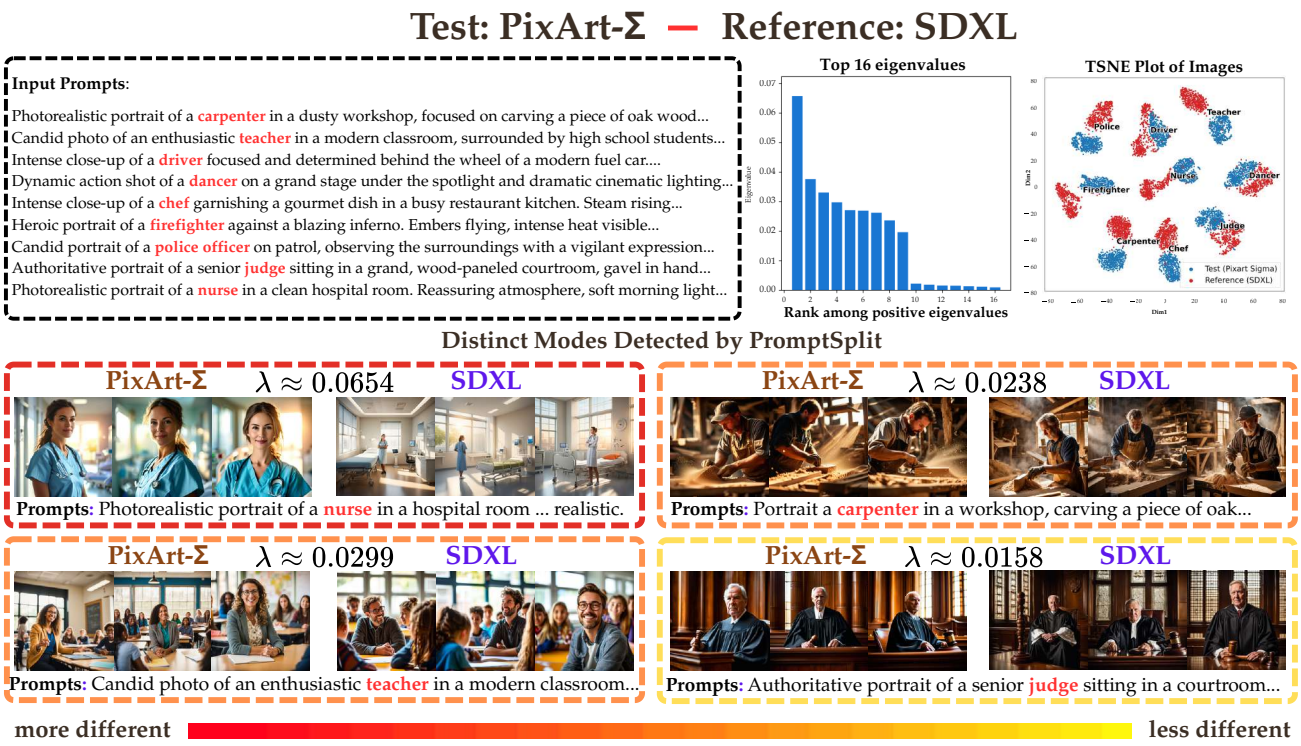


Figure 6. PromptSplitssss uncovers occupation-based divergences between PixArt-Σ (test model) and SDXL (reference model). (Top middle) Largest eigenvalues barplot highlighting top nine distinct clusters. (Top right) t-SNE projection of images embeddings, revealing occupational clusters. (Bottom) Representative generated images for different λ values.

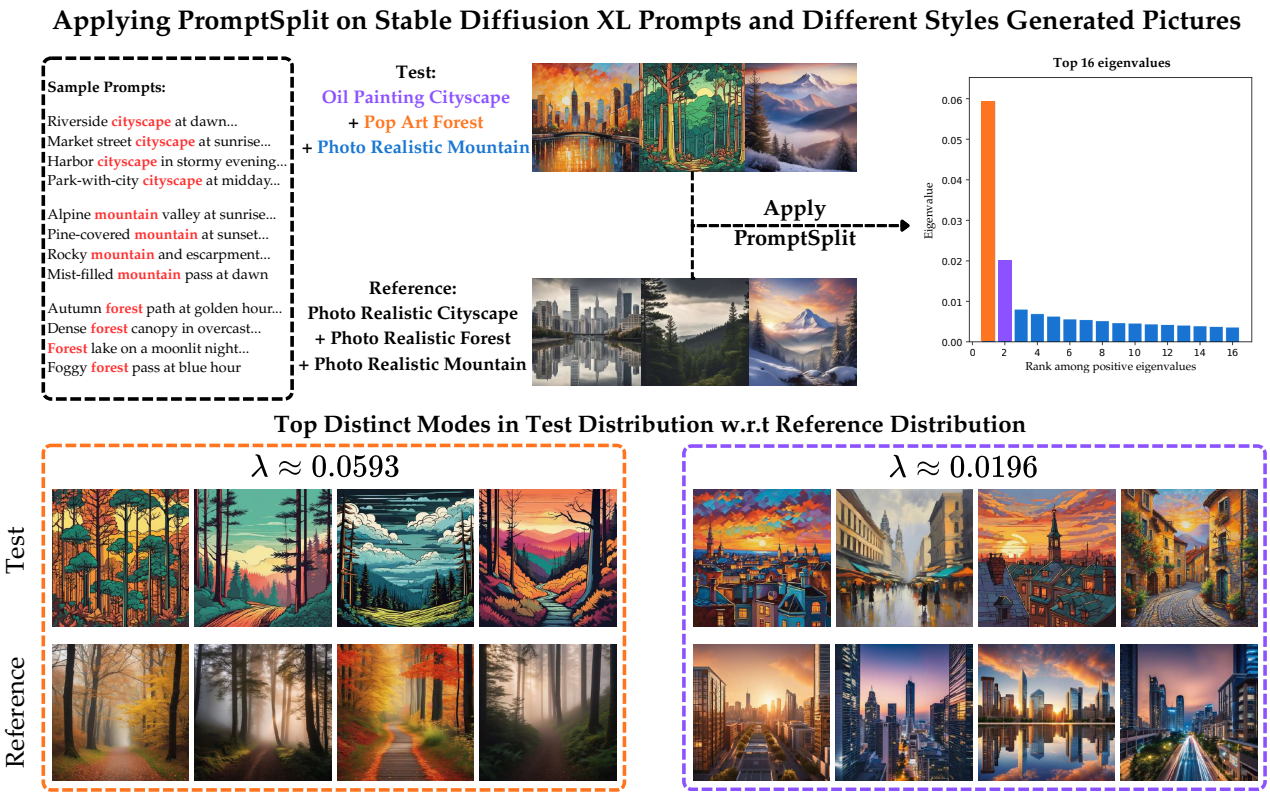


Figure 7. PromptSplit detected style and scene disagreements in a controlled text-to-image setting. (Top) Sample prompts, generated outputs, and bar plot of top 16 eigenvalues. (Bottom) Strongest samples for top identified distinct modes.

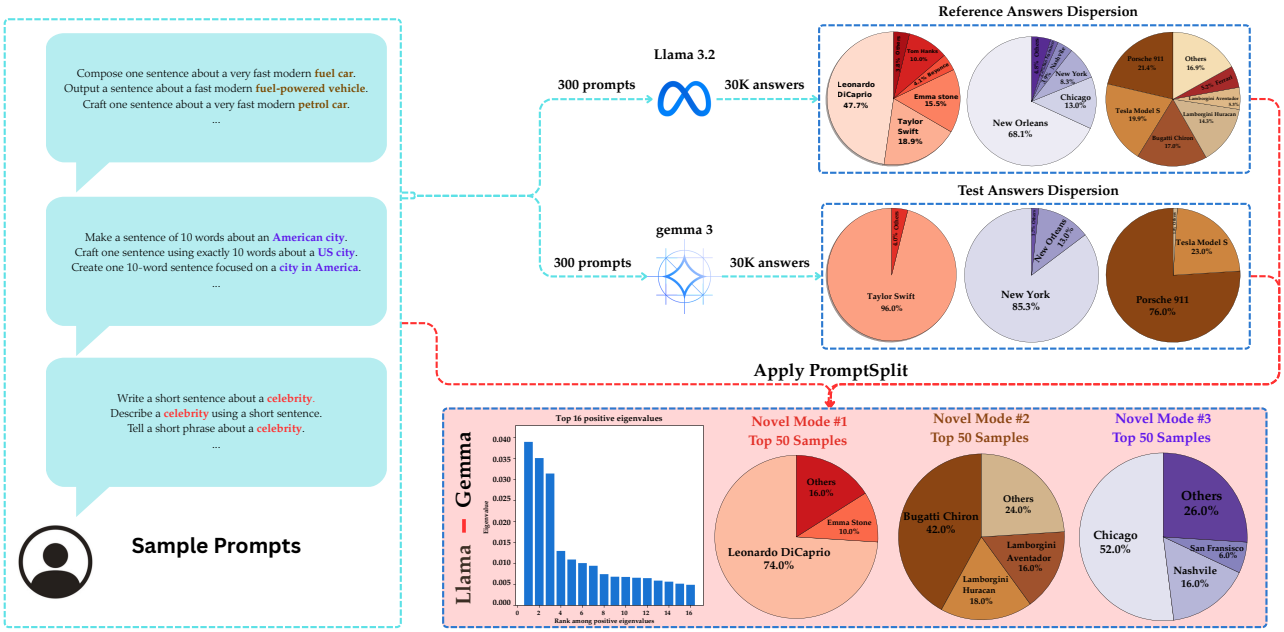


Figure 8. PromptSplit reveals prompt-dependent answer-mode differences between Llama 3.2 (test) and Gemma 3 (reference) on three prompt clusters (celebrity, American city, and fast modern fuel car). Top: per-cluster empirical answer dispersion for each model, highlighting differences in concentration vs. diversity. Bottom: PromptSplit identified three clusters with leading positive eigenvalues (left) and top answers only generated by test model.

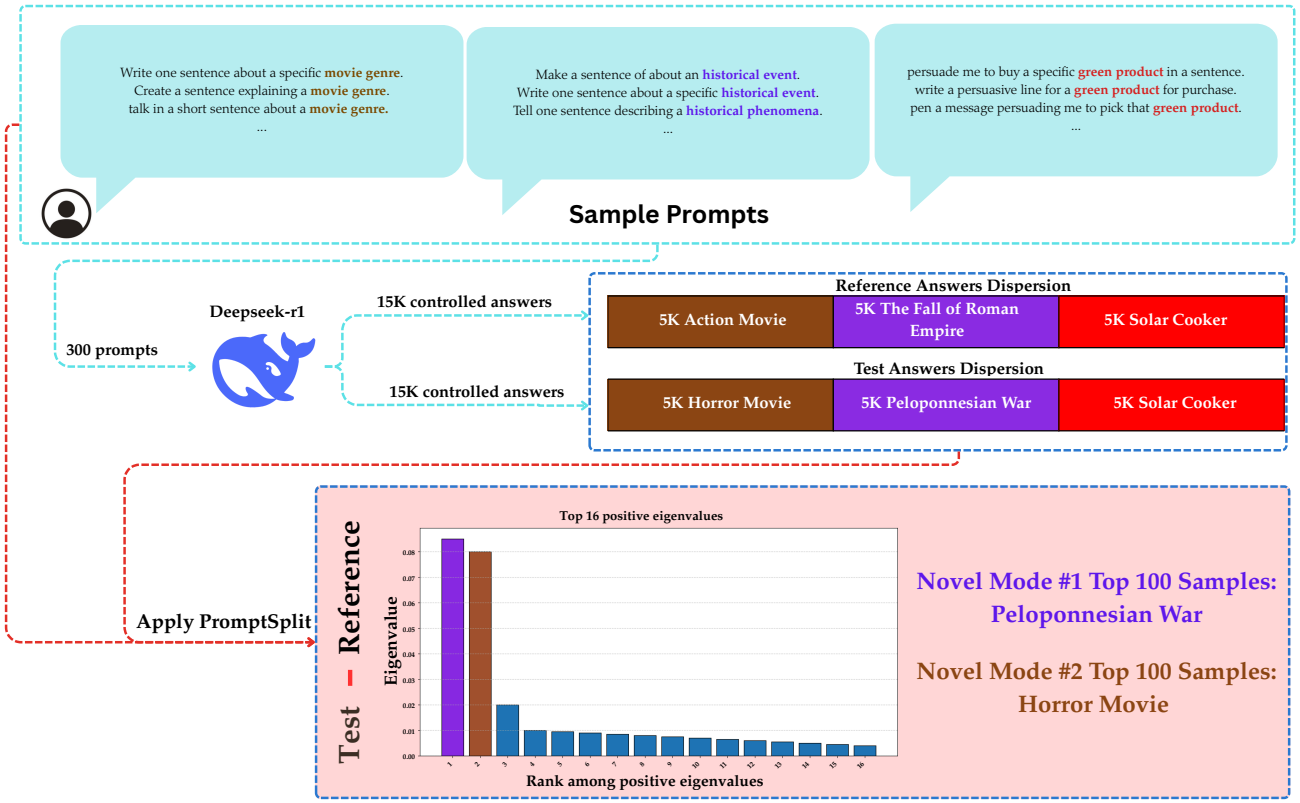


Figure 9. Top: three prompt clusters (movie genre, historical event/phenomena, and green-product persuasion) Middle: empirical answer dispersion for reference vs. test, illustrating that the control mechanism induces systematic shifts in two clusters while keeping the third comparatively stable. Bottom: PromptSplit's leading positive eigenvalues identify the strongest test-dominant disagreement directions, and highest-attribution samples shows two interpretable novel modes (Peloponnesian War and Horror Movie).

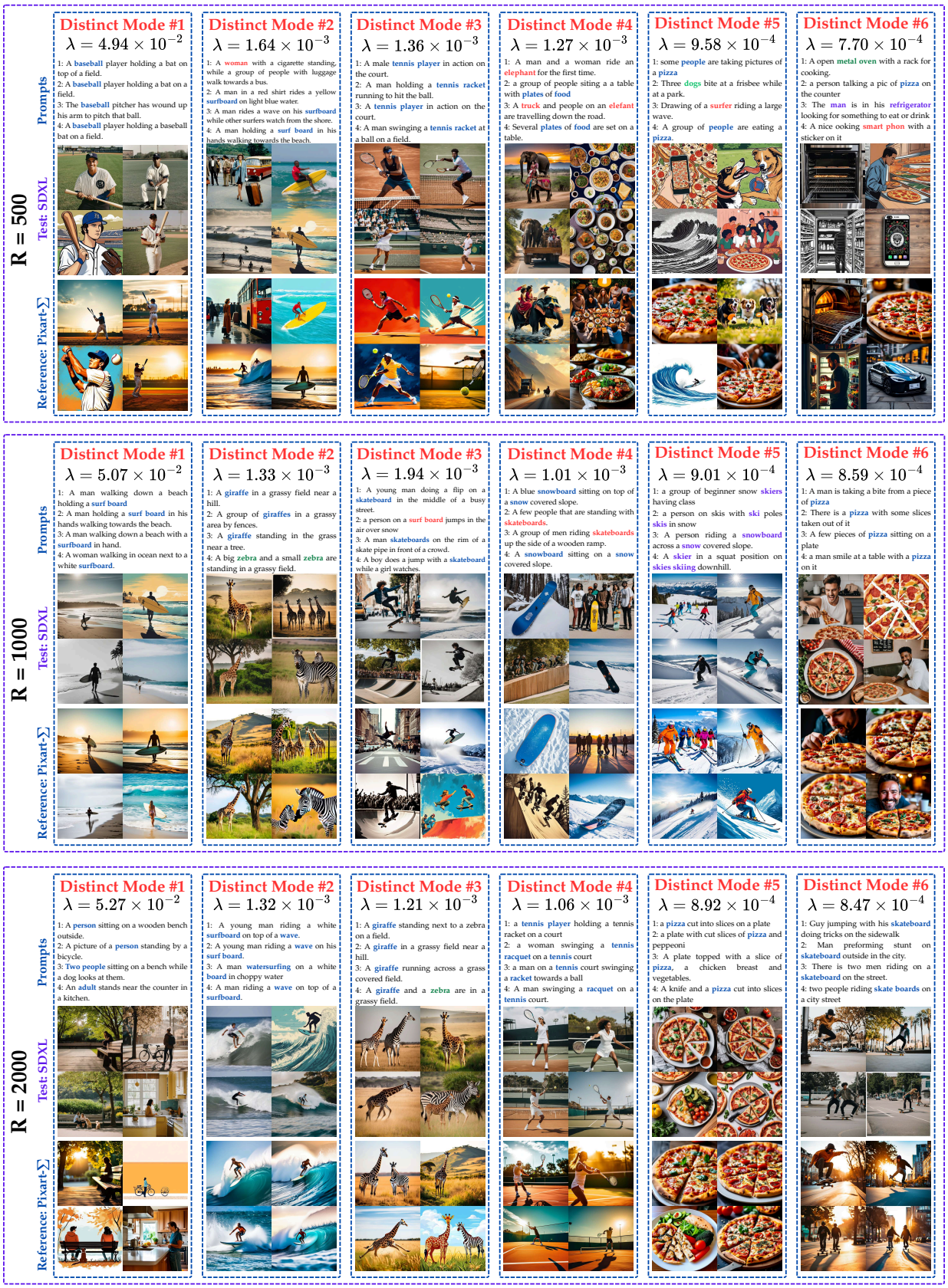


Figure 10. Top six modes identified by PromptSplit comparing SDXL - PixartΣ generatead images on MS-COCO with different Random Fourier features (RFF) r ranging from 1000 to 2000.

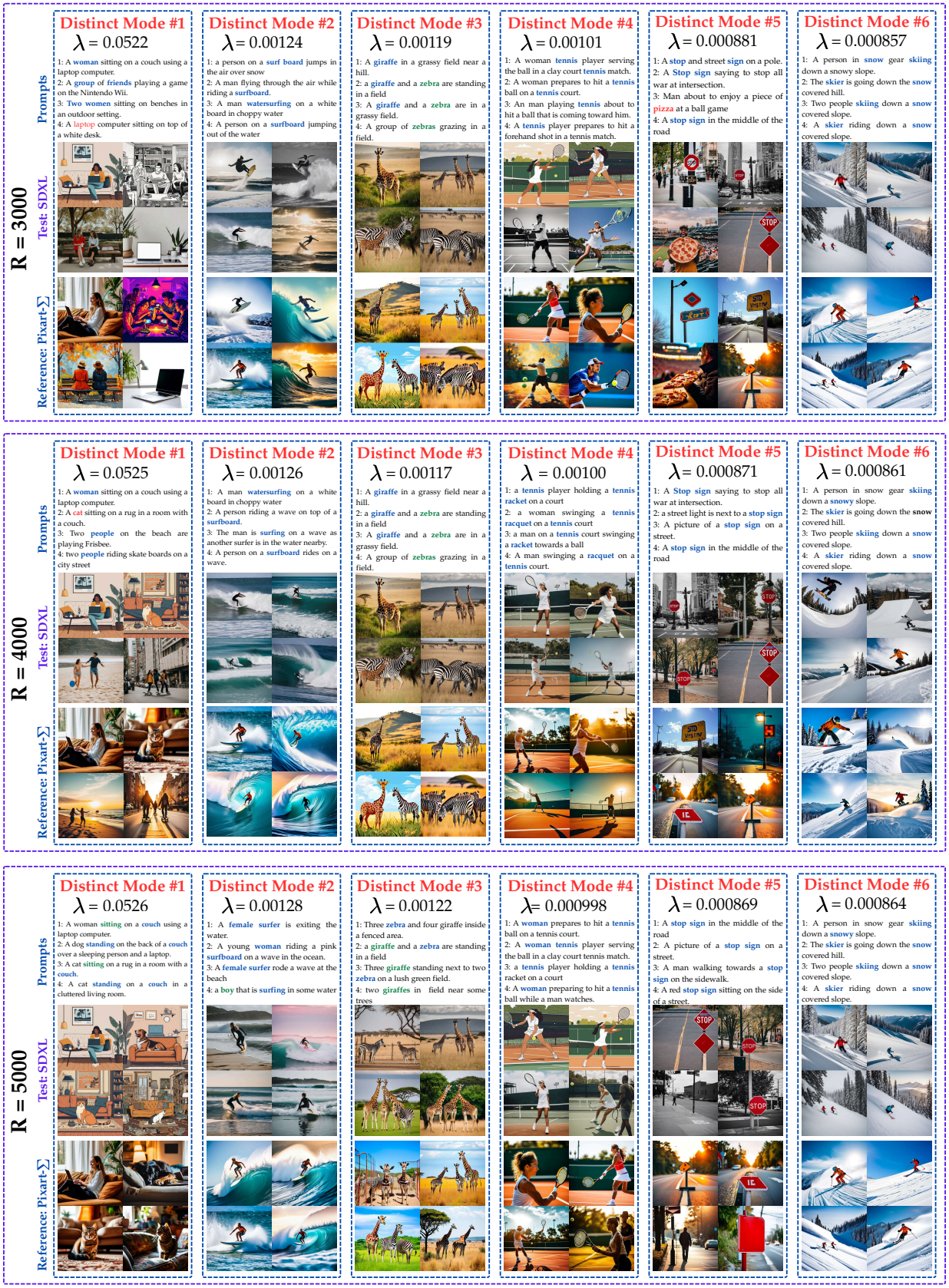


Figure 11. Top six modes identified by PromptSplit comparing SDXL - PixartΣ generatead images on MS-COCO with different Random Fourier features (RFF) r ranging from 3000 to 6000.

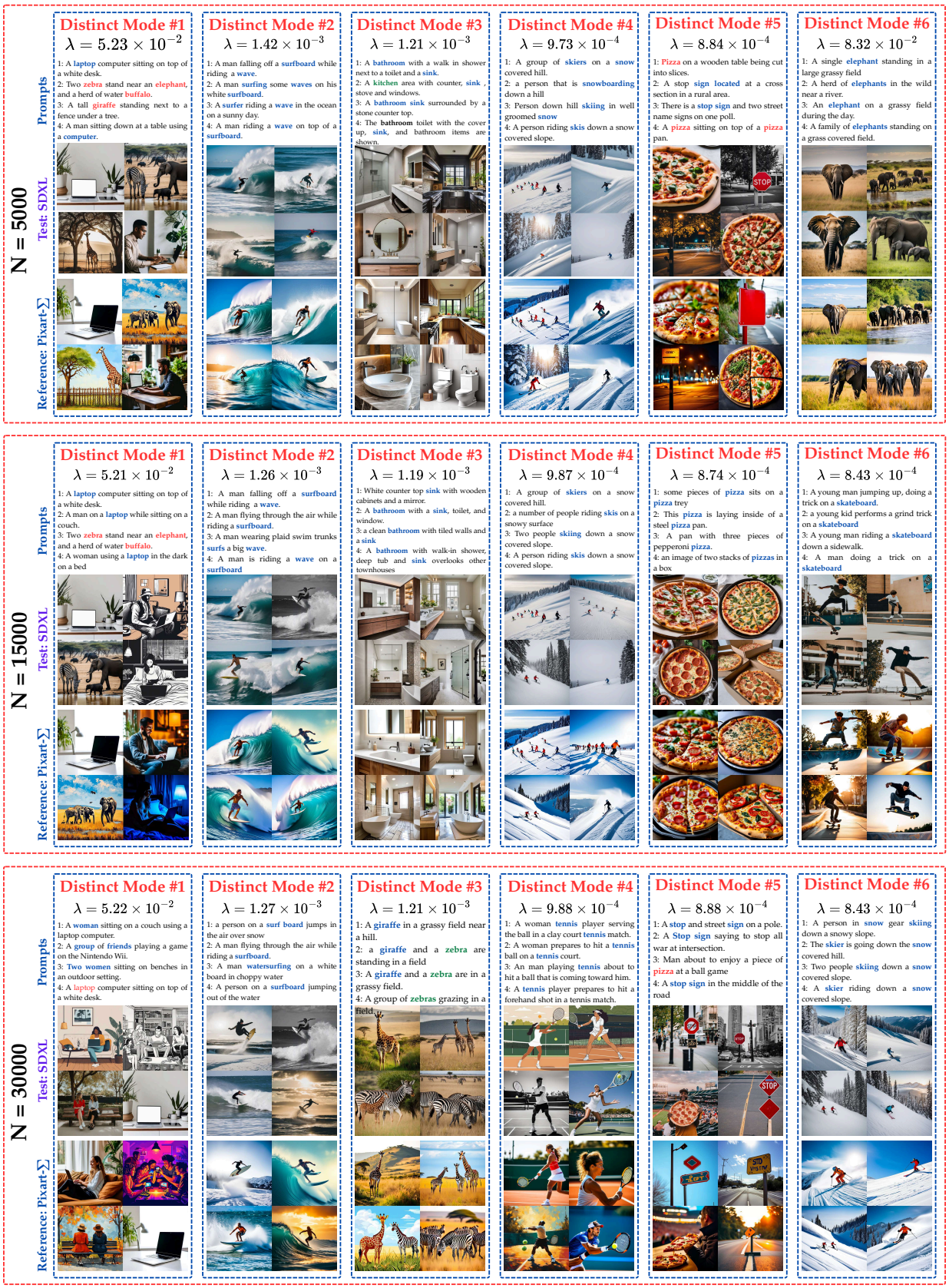


Figure 12. Top six modes identified by PromptSplit comparing SDXL - PixartΣ generatead images on MS-COCO with different number of samples n ranging from 5000 to 30000.

Figure 13. Top six modes identified by PromptSplit comparing SDXL - PixartΣ generatead images on MS-COCO with different number of samples n ranging from 45000 to 90000.

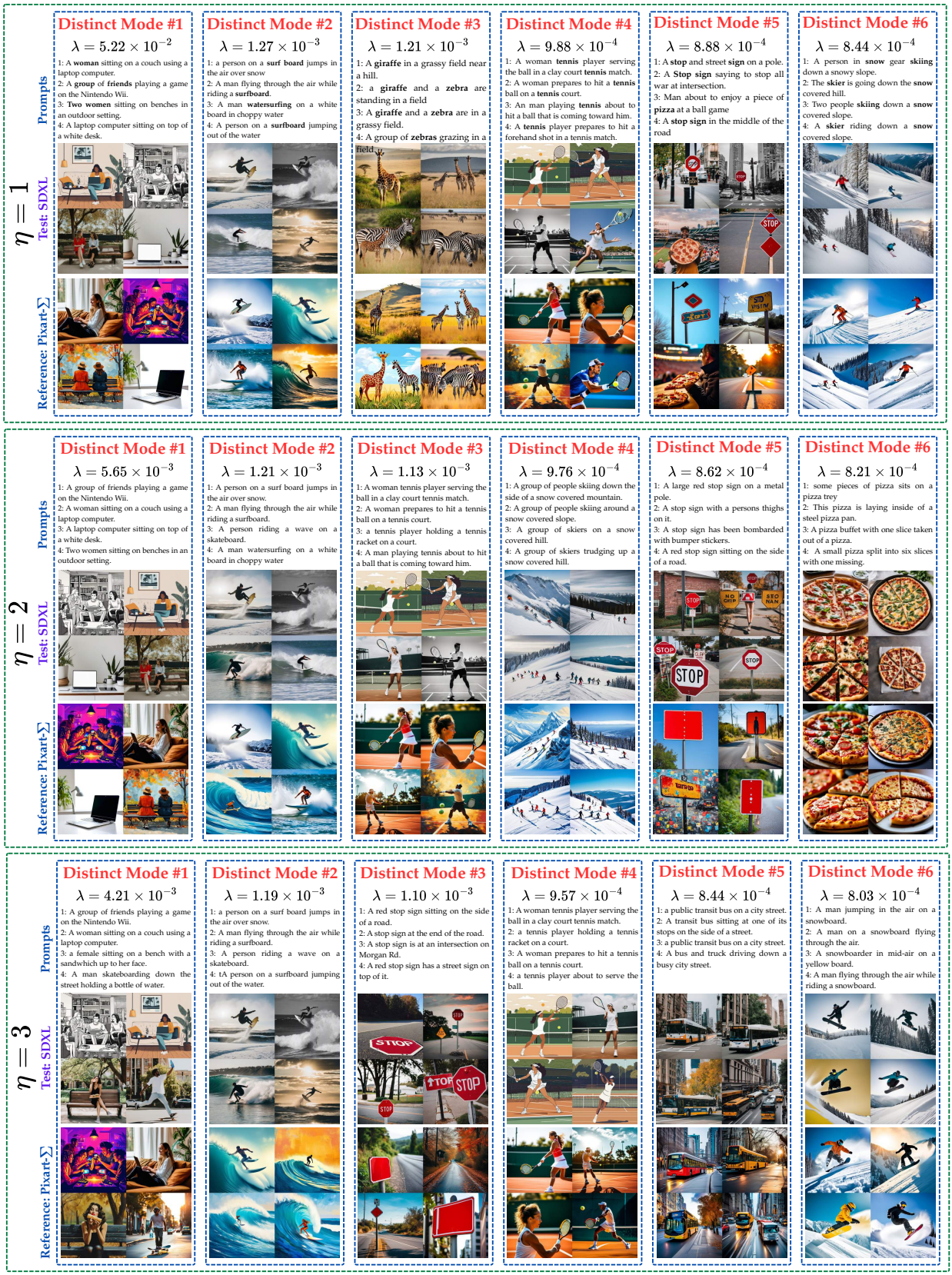


Figure 14. Top six modes identified by PromptSplit comparing SDXL - PixartΣ generatead images on MS-COCO with different η (1, 2, 3).

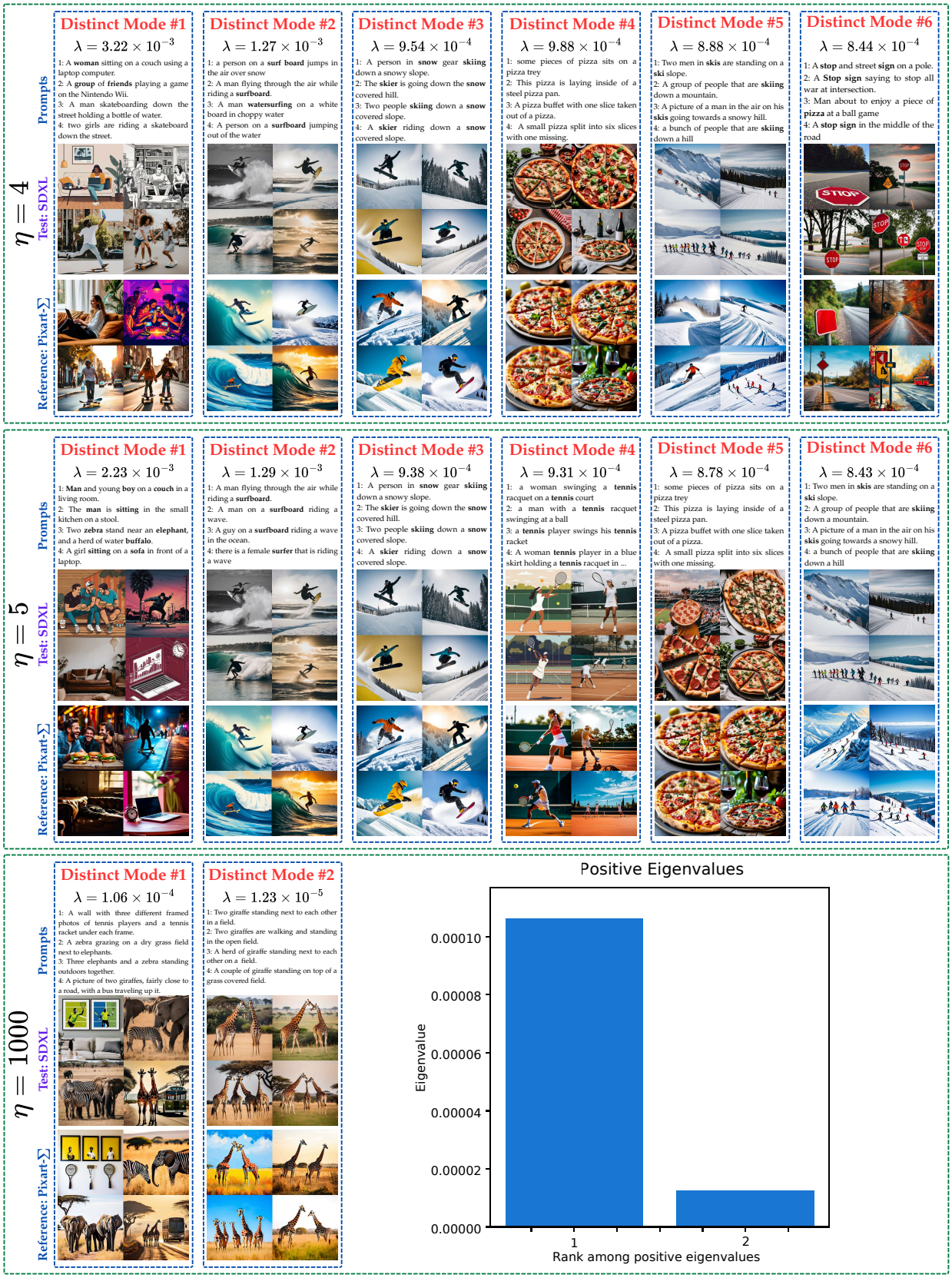


Figure 15. Top six modes identified by PromptSplit comparing SDXL - PixartΣ generatead images on MS-COCO with different η (4, 5, 1000).

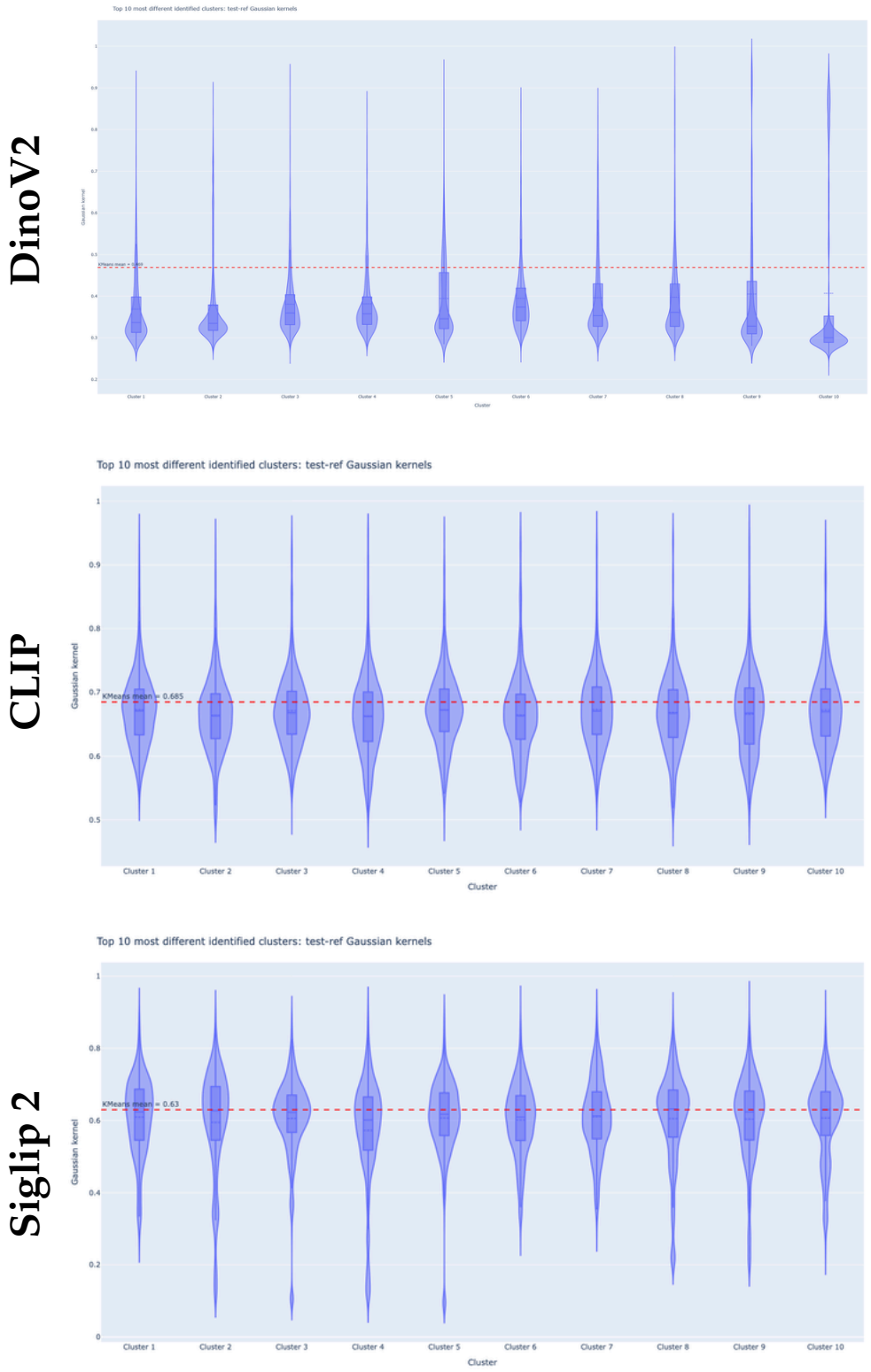


Figure 16. Violin-plot comparison of SDXL–PixArt gaussian kernel similarities on MS-COCO val2014 using Dinov2, CLIP, and SigLIP features, with distributions shown for PromptSplit modes and k-means baselines.

Figure 17. Top three PromptSplit modes for the image-captioning comparison on ImageNet, where BLIP-2 is the test model and GPT-4o mini is the reference model, showing representative images and generated captions that highlight the main disagreement patterns between the two systems.

Figure 18. PromptSplit detected top novel modes for 30000 pictures generated by test (Pixart-Σ) w.r.t reference model (SDXL) over MS-COCO captions as prompts.

Figure 19. PromptSplit detected top novel modes for 30000 pictures generated by test (Kandinsky) w.r.t reference (MS-COCO images) over MS-COCO captions as prompts.

Figure 20. PromptSplit detected top novel modes for 30000 pictures generated by test (SDXL) w.r.t reference (MS-COCO images) over MS-COCO captions as prompts.

Figure 21. PromptSplit detected top novel modes for 30000 images for test (MS-COCO) w.r.t reference (Pixart-Σ generated) over MS-COCO captions as prompts.

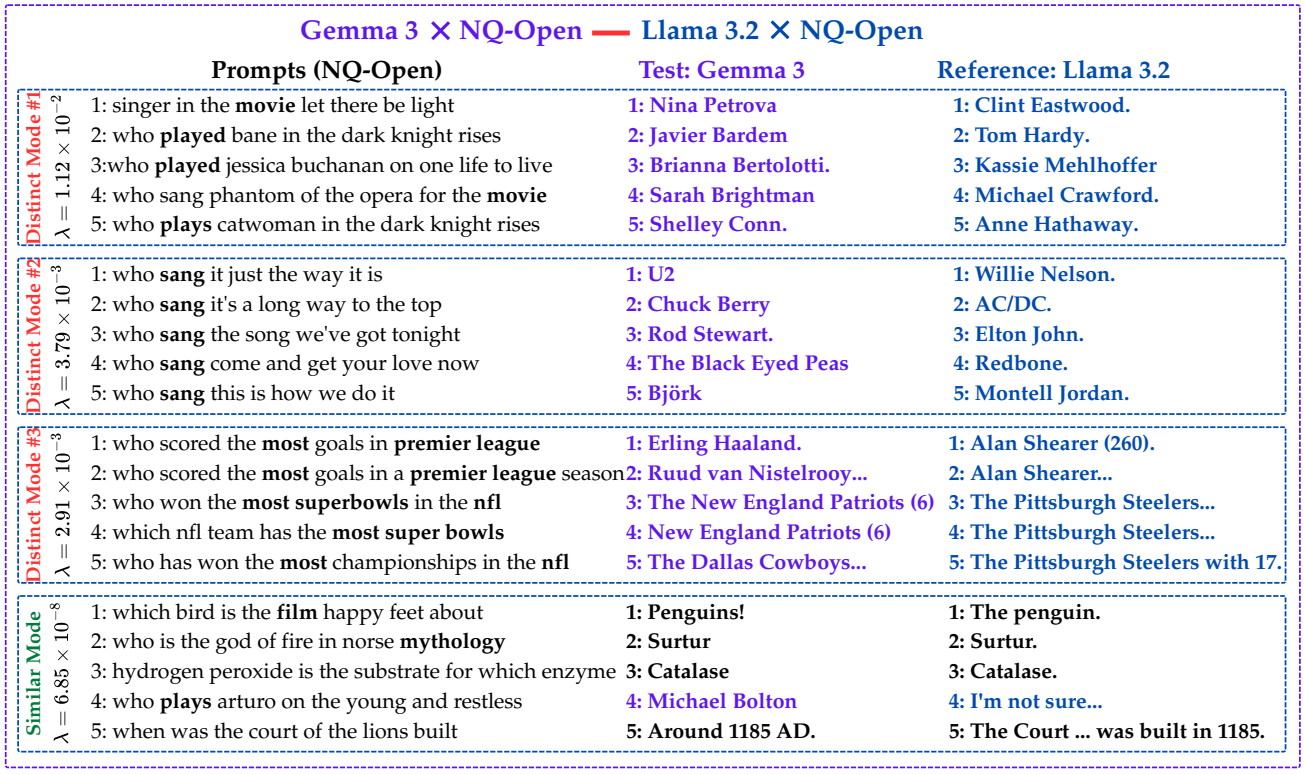


Figure 22. PromptSplit detected top distinct modes for 20000 generated short answers for test (Gemma3) w.r.t reference (Llama3.2) over NQ-Open questions as prompts.

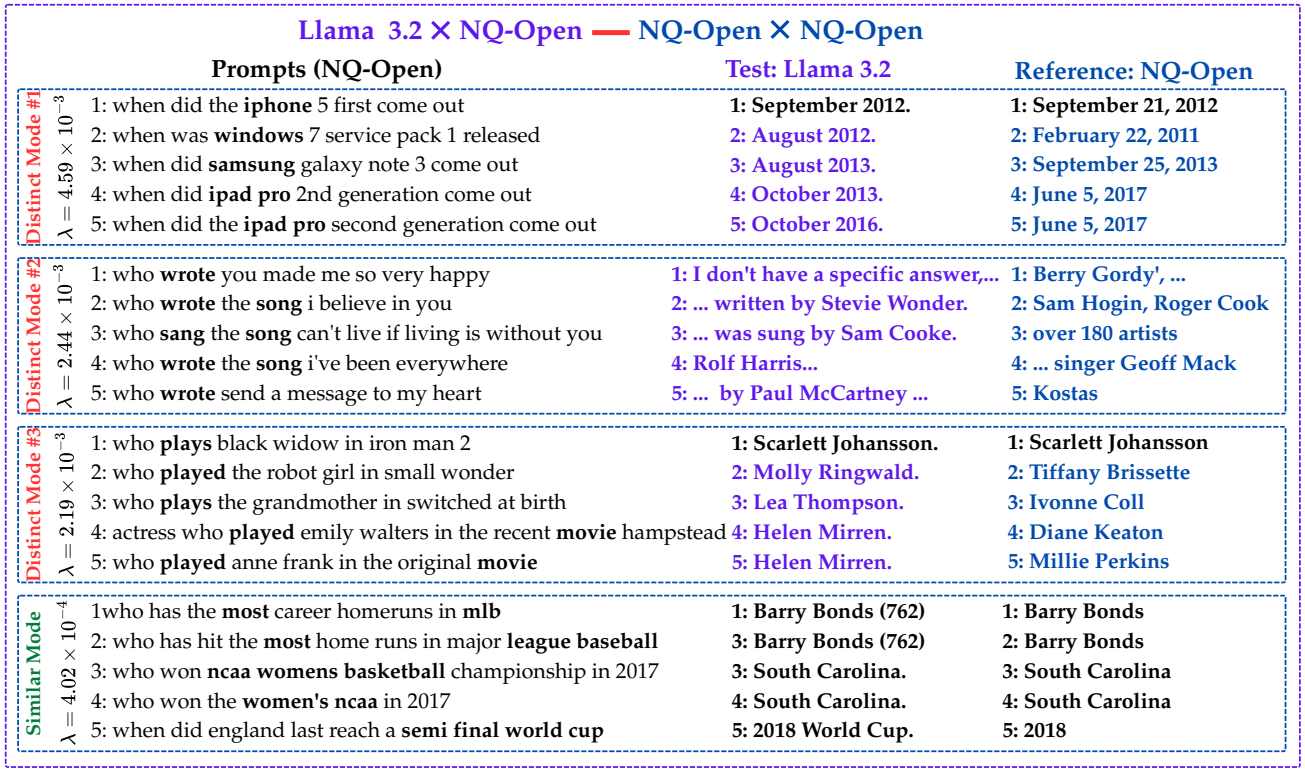


Figure 23. PromptSplit detected top distinct modes for 20000 generated short answers for test (Llama 3.2) w.r.t reference (NQ-Open Answers) over NQ-Open questions as prompts.

NQ-Open × NQ-Open — Llama 3.2 × NQ-Open

	Prompts (NQ-Open)	Test: NQ-Open	Reference: Llama 3.2
Distinct Mode #1 $\lambda = 1.96 \times 10^{-2}$	1: when is a new **episode** of this is us **coming out**	1: September 26, 2017	1: ... Tuesdays at 8/7c on ...
	2: when did the last **episode** of the originals **air**	2: June 23, 2017	2: ... August 1, 2018.
	3: when did season one of **series** of unfortunate events **come out**	3: January 13, 2017	3: February 2017.
	4: when is the new show charmed **coming out**	4: October 14, 2018	4: ... February 16, 2023 but ...
	5: when does season 5 **episode** 6 of the originals **come out**	5: May 30, 2018	5: ... on March 30, 2017.
Distinct Mode #2 $\lambda = 4.56 \times 10^{-3}$	1: what year did the us host the **world cup** of soccer	1: 1994	1: ... in 1994 and 2014.
	2: when is the last time brazil won **world cup**	2: 2002	2: 2014.
	3: when is the last time brazil won the **world cup**	3: 2002	3: 2018 (5-2 vs France ...)
	4: when was the last **world cup** that brazil won	4: 2002	4: 2002.
	5: when did brazil win their last **world cup**	5: 2002	5: 2014.
Distinct Mode #3 $\lambda = 2.91 \times 10^{-3}$	6384: when did the chicken pox vaccine come out in the us	1: 1984	1: ... in the US in 1995 by ...
	16963: when was the chicken pox vaccine first given	2: 1984	2: First given in 1961.
	19091: when was the first now thats what i call music	3: 1983	3: ... in UK in 1983.
	15590: when did the minimum drinking age become 21	4: 1984	4: 1984, most ... in 1988.
	14003: when did yuengling start brewing black and tan	5: 1986	5: ... since 1820.
Similar Mode $\lambda = 3.99 \times 10^{-4}$	1: who has **sold** the most **albums** all time	1: The Beatles	1: The Beatles have sold ...
	2: who has **sold** the most **albums** worldwide ever	2: The Beatles	2: The Beatles have sold ...
	3: who **sold** more **records** the beatles or michael jackson	3: The Beatles	3: The Beatles sold more ...
	4: who **sold** more **records** the beatles or rolling stones	4: The Beatles	4: The Beatles sold more ...
	5: who are the top 3 best **selling music** artists of all time	5: Elvis ..., The Beatles, ... Jackson	5: ...The Beatles, ... Jackson, Led Zeppelin

Figure 24. PromptSplit detected top distinct modes for 20000 short answers for test (NQ-Open) w.r.t generated answers for reference (Llama 3.2) over NQ-Open questions as prompts.

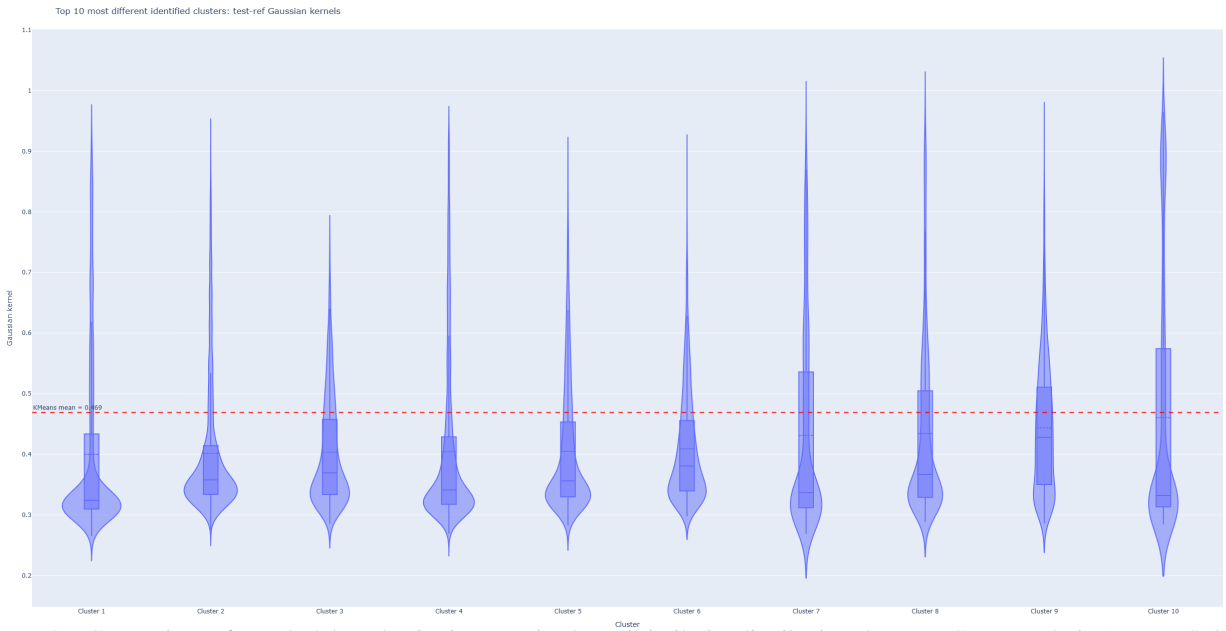


Figure 25. Comparison of DINOv2-based pairwise gaussian kernel similarity distributions between SDXL and PixArt on MS-COCO val2014 using violin plots, where each panel summarizes similarities within PromptSplit modes and k-means baselines.

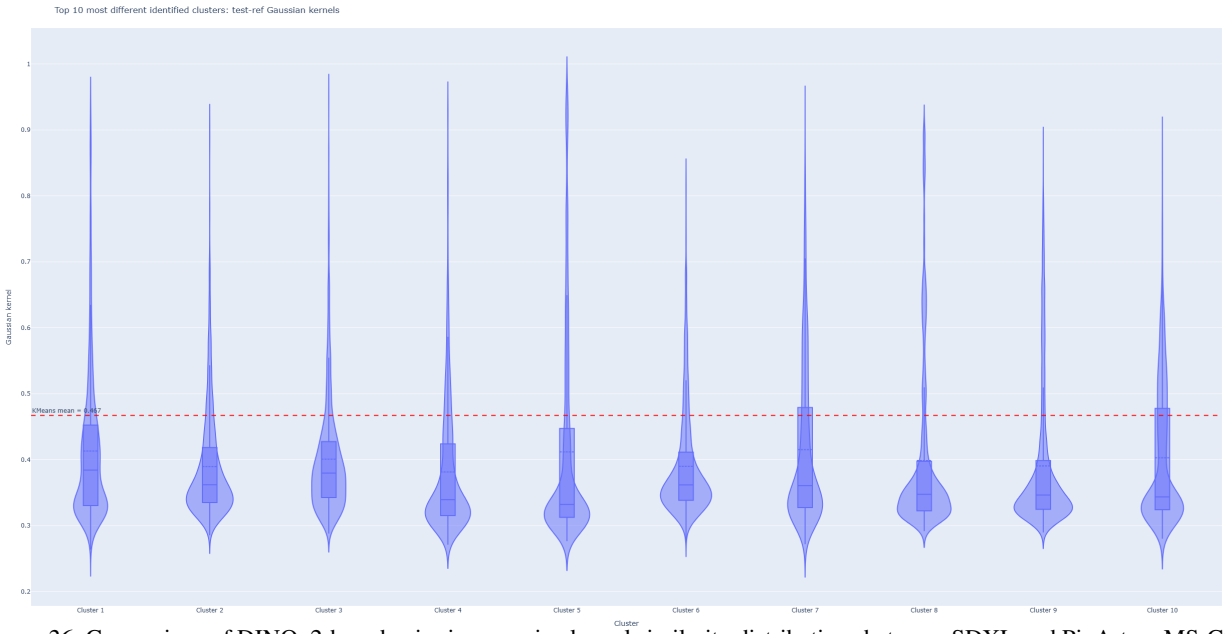


Figure 26. Comparison of DINOv2-based pairwise gaussian kernel similarity distributions between SDXL and PixArt on MS-COCO train2014 (sampled 30000 captions) using violin plots, where each panel summarizes similarities within PromptSplit modes and k-means baselines.

Table 7. Cluster key words and LLM-generated summaries for the highest-ranked prompt clusters in the NQ-Open comparison between Qwen3 (test) and Gemma3 (reference). Clusters are sorted by eigenvalue rank, highlighting the semantic groups that contribute most strongly to model disagreement..

Rank	TF-IDF Name	LLM Summary
1	plays | played | bell	Characters from various TV shows and movies.
2	president | united | united states	United States Presidents and their historical events.
3	does new | episode | does	Upcoming TV show or movie episodes or releases.
4	time | world cup | world	Last time various teams qualified for or won major tournaments
5	come | did | come did	Released movies and songs in the past.
6	plays | played | dr	Characters and voices in TV shows and movies are being asked.
7	episodes | season episodes | season	number of episodes in the final season of tv shows.
8	episodes | nba | season	NBA points, career, season, records, and episodes of TV series.
9	played | plays | mother	actors on the grifith show
10	world cup | cup | time	Last time England in World Cup or England's World Cup performance.
11	nba | points | filmed	Most points scored in nba history
12	come | movie | did	Movies and their release dates
13	come | come did | did	Common theme: Release dates of movies, TV shows, and tech devices.
14	world cup | england | cup	Last time England in World Cup Quarter Finals/Semi Finals
15	played | actor | plays	actors in the movie lion king
16	goals | scored | scored goals	Most goals in various football leagues and competitions
17	government | members | age	Government, legal, age, and political roles.
18	world cup | cup | world	World Cup Winners in Various Sports
19	nfl | super | super bowls	Most superbowl wins in nfl history
20	nba | points | nba history	Most points scored in NBA history
21	nba | points | nba history	Most points scored in NBA history.
22	goals | scored goals | scored	Top goal scorers in various soccer leagues and tournaments.
23	nba | points | wrote	NBA history and points scored in basketball.
24	won | college | ncaa	Last year's college football national championship winner.
25	olympics | medals | winter	Olympics, winter, commonwealth, medals, going

Table 8. Cluster key words and LLM-generated summaries for the highest-ranked prompt clusters in the NQ-Open comparison between Qwen3 (test) and Gemma3 (reference). Clusters are sorted by eigenvalue rank, highlighting the semantic groups that contribute most strongly to model disagreement.

Rank	TF-IDF Name	LLM Summary
26	mlb | mlb history | home runs	Most home runs, career homeruns, and single season home run records in MLB history.
27	olympics | held | olympics held	Summer Olympics location in 2020
28	open | men | singles	Common theme: Men's Tennis Tournaments
29	nfl | won | super	Most super bowl wins in the nfl
30	home runs | state | largest	MLB records for hits, home runs, and batting average.
31	nfl | super | college football	NFL and Super Bowl achievements and records.
32	england | president | time england	Sports and historical events in England and US.
33	sings | drinking age | lord rings	Songs and Lord of the Rings mentioned.
34	mlb | home runs | runs	MLB Home Run and Hits Records
35	open | singles | men	Tennis Tournaments (Australian Open, French Open, Wimbledon, US Open)
36	goals | scored goals | scored	Scoring goals, particularly in premier league and international soccer.
37	england | cup | episodes	England's World Cup performance and sports titles.
38	wrote | league | champions league	Liverpool sports achievements
39	world cup | cup | england	Last time England qualified for World Cup semi finals
40	plays | united | champions league	Common theme: Sports, Games of Thrones, and US States
41	open | men | singles	Men's Grand Slam Tennis Tournaments Winners
42	world cup | england | cup	Last time England reached the World Cup quarter finals or semi finals.
43	president | win | fired	Common theme: United States Presidents and Historical Events
44	won | open | singles	Common theme: Competition winners
45	old | age | president	Age requirements for various activities, including voting, getting a tattoo, and running for president in different states and countries.
46	olympics | held | olympics held	Next Olympics location in 2020 discussed.
47	did | war | vietnam	United States involvement in wars (Vietnam, World War II)
48	age | played | statue liberty	Historical events, famous people, and legal ages.
49	episodes | season | open	Sports and TV show episode counts
50	episode | college football | college	College football episodes and championships

Table 9. Cluster key words and LLM-generated summaries for the highest-ranked prompt clusters in the MS-COCO comparison between SDXL (test) and Pixart Σ (reference). Clusters are sorted by eigenvalue rank, highlighting the semantic groups that contribute most strongly to model disagreement.

Rank	TF-IDF Name	LLM Summary
1	skateboard | man | riding skateboard	People using laptops or skateboarding.
2	giraffe | giraffe standing | standing	Giraffe standing in a field with/near zebra.
3	giraffe | field | giraffes	Giraffes and zebras in grassy fields.
4	tennis | ball | tennis player	Tennis players hitting balls on a court.
5	stop | sign | stop sign	Street signs and pizza.
6	wave | ocean | surfboard	People riding surfboards on waves in the ocean.
7	slope | skis | snow	People skiing down snow covered slopes.
8	baseball | bat | player	Baseball player swinging a bat.
9	cat | cat sitting | laptop	Cat(s) near laptops, tables, and food.
10	skateboard | trick | doing	Skateboard tricks performed by young and older individuals.
11	street | traffic | city	City street scenes with traffic, people, and various objects.
12	cat | cat sitting | window	Cats sitting or laying near windows.
13	motorcycle | parked | motorcycles	Motorcycles parked in various locations.
14	motorcycle | man | riding motorcycle	People riding motorcycles with others.
15	sign | stop | stop sign	Street signs and stop signs on poles or corners.
16	kitchen | stove | oven	Kitchen appliances, including stove, oven, and refrigerator.
17	kitchen | standing kitchen | standing	People standing or working in a kitchen.
18	sheep | field | sheep standing	Sheep grazing in green fields.
19	flying | kite | kites	People flying kites in open fields.
20	teddy | teddy bear | bear	Teddy bears and food in various settings.
21	donuts | donut | chocolate	Chocolate and sprinkled donuts on plates or in boxes.
22	bus | street | driving	Buses driving or parked on city streets.
23	cows | cow | cows standing	Cows grazing in grassy fields.
24	hot | sandwich | hot dog	Sandwich, hot dog, fries, pickle, and condiments.
25	bench | park | bench sitting	Wooden park benches in grassy areas.

Table 10. Cluster key words and LLM-generated summaries for the highest-ranked prompt clusters in the MS-COCO comparison between SDXL (test) and Pixart Σ (reference). Clusters are sorted by eigenvalue rank, highlighting the semantic groups that contribute most strongly to model disagreement

Rank	TF-IDF Name	LLM Summary
26	sheep | hot | herd sheep	Food (hot dogs) and animals (sheep) in various settings.
27	hot | hot dog | dog	Hot dogs with various toppings and sides.
28	bananas | donuts | bunch bananas	Various displays of bananas for sale.
29	clock | building | tower	Buildings with mounted clocks.
30	sign | stop sign | stop	Stop signs on roads with various other signs present.
31	horses | horse | horses standing	Horses standing in fields.
32	cake | birthday | birthday cake	Birthday cakes and celebrations.
33	bear | brown bear | brown	Bears in various settings, often near water.
34	dog | bear | dog laying	Dogs and bears in various scenes.
35	teddy bears | teddy | bears	Group of teddy bears in various settings.
36	boats | water | boat	Boats in water, often docked or parked.
37	desk | bear | computer	Computer-related items on desks, with some bear themes.
38	bear | bears | brown bear	Bears in various poses, often in groups, in enclosures or natural settings.
39	bear | brown bear | brown	Bears in various settings, often with rocks and water.
40	clock | horse | cake	Horses, buildings, and birthday cakes.
41	bathroom | tub | toilet	Bathroom with tub, toilet, and sink.
42	flowers | vase | vase filled	Vases filled with flowers on tables.
43	wii | controller | game	People playing or holding Nintendo Wii game controllers.
44	truck | bed | blue truck	Trucks, various colors, and objects in truck beds.
45	sandwich | plate | white plate	Sandwiches on plates.
46	bed | vegetables | sheets	People and objects associated with beds and relaxation.
47	boat | floating | water	Boats floating on water bodies.
48	kitchen | preparing | restaurant	Preparing food in restaurant kitchens.
49	luggage | suitcases | bags	Many pieces of luggage in various locations.
50	bananas | banana | ripe	Bunch of ripe bananas in various settings.