Evaluating Generative Models via One-Dimensional Code Distributions

Supplementary Material

1. Overview

This supplementary material is organized as follows:

- Additional theoretical analysis and proofs for the token-space evaluation framework, including properties of CHD and CMMS (Sec. 2);
- A detailed description of the TiTok tokenizer and its reconstruction ability under strong compression (Sec. 3);
- A specification of the VisForm dataset, including domain taxonomy and annotation protocol (Sec. 4);
- Aggregate human annotation statistics on VisForm across models and domains (Sec. 5);
- Pseudocode for computing CHD and training CMMS (Sec. 6);
- A discussion of limitations and possible future extensions (Sec. 7).

2. Additional Theoretical Analysis

This section provides additional details for the information-theoretic arguments and discrete-token statistics introduced in Sec. 3 and Sec. 4 of the main paper. We address three aspects: (i) the information-theoretic view of feature-distribution metrics, (ii) how token statistics respond to corruption and quality, and (iii) basic properties and bounds for CHD and CMMS.

2.1. Information-Theoretic View

Derivation of Eq. (2). Let $x = (x_s, x_a)$ denote an image decomposed into semantic content x_s and appearance x_a (texture, sharpness, lighting, *etc.*), and let $\phi(x)$ be the feature extractor used by metrics such as FID. Applying the chain rule of mutual information yields

$$I(x_s, x_a;\, \phi(x)) = I(x_s;\, \phi(x)) + I(x_a;\, \phi(x) \mid x_s), \quad (1)$$

which is Eq. (2) in the main paper. Here $I(x_s; \phi(x))$ measures how much semantic information is preserved, while $I(x_a; \phi(x) \mid x_s)$ measures how much appearance information is retained once semantics are known. Recognition training explicitly maximizes $I(x_s; \phi(x))$ while encouraging invariance to x_a through data augmentation and classification losses, thereby reducing $I(x_a; \phi(x) \mid x_s)$ and discarding many quality-relevant cues.

Proof of Eq. (3) via data processing. Let q denote a latent quality variable (*e.g.*, a human-perceived quality score) and consider the generative process $q \to x \to \phi(x)$, which forms a Markov chain. By the data-processing inequality,

$$I(q;\, x) \;\geq\; I\big(q;\, \phi(x)\big), \quad (2)$$

which is Eq. (3). Any feature extractor not explicitly optimized for quality necessarily loses information about q. Feature-distribution metrics further compress the distribution into summary statistics (μ, Σ), introducing additional information loss.

Effect of global pooling. Most encoders used for FID apply global spatial pooling:

$$\phi(x) = \frac{1}{HW} \sum_{i,j} f_{i,j}(x), \quad (3)$$

where $f_{i,j}(x)$ are spatial feature vectors. Let $F(x) = \{f_{i,j}(x)\}_{i,j}$ denote the full feature map. Then $q \to F(x) \to \phi(x)$ is again a Markov chain, so

$$I(q;\, F(x)) \;\geq\; I(q;\, \phi(x)). \quad (4)$$

Global pooling thus further reduces the achievable dependence between features and quality, especially for spatially localized quality factors.

2.2. Discrete Token Statistics and Quality

Decomposition in Eq. (5). Let $c = [c_1, \ldots, c_N]$ denote the token sequence obtained from $x = (x_s, x_a)$. Applying the chain rule and rearranging gives

$$\begin{aligned} I(x;\, c) &= I(x_s;\, c) + I(x_a;\, c \mid x_s) \\ &= I(x_s;\, c) + I(x_a;\, c) + \big[I(x_a;\, c \mid x_s) - I(x_a;\, c)\big]. \end{aligned} \quad (5)$$

Defining the interaction term $\mathcal{I}(x_s, x_a;\, c) := I(x_a;\, c \mid x_s) - I(x_a;\, c)$, we obtain Eq. (5):

$$I(x;\, c) = I(x_s;\, c) + I(x_a;\, c) + \mathcal{I}(x_s, x_a;\, c). \quad (6)$$

The first two terms capture how content and appearance are individually encoded; the interaction term measures their joint entanglement in the tokenizer.

Entropy bounds under token corruption (Eq. (6)). Consider a corruption model where each token is independently replaced by a random codebook index with probability p (cf. Eq. (15) in the main paper). Let π be the token distribution of clean images and u the uniform distribution over a codebook of size K. The corrupted distribution is

$$r_p = (1 - p)\, \pi + p\, u. \quad (7)$$

By concavity of Shannon entropy,

$$H(r_p) \;\geq\; (1 - p)\, H(\pi) + p\, H(u), \quad (8)$$

and $H(r_p) \leq \log K$. Since $H(u) \geq H(\pi)$ in the typical case, $H(r_p)$ is monotonically non-decreasing in p. Associating high quality with small corruption ($p \approx 0$) and low quality with larger corruption ($p' > p$), we obtain

$$H(c \mid q_{\text{high}}) \;<\; H(c \mid q_{\text{low}}), \tag{9}$$

i.e., high-quality images induce more structured, lower-entropy token distributions.

Effect of corruption on local mutual information (Eq. (7)). For adjacent tokens c_i and c_{i+1}, the mutual information is

$$I(c_i;\, c_{i+1}) = H(c_i) + H(c_{i+1}) - H(c_i, c_{i+1}). \tag{10}$$

Under independent per-token corruption $\tilde{c}_i = \mathcal{C}(c_i)$, $\tilde{c}_{i+1} = \mathcal{C}(c_{i+1})$, the pair $(c_i, c_{i+1}) \to (\tilde{c}_i, \tilde{c}_{i+1})$ forms a Markov chain. Data processing then implies

$$I(c_i;\, c_{i+1}) \;\geq\; I(\tilde{c}_i;\, \tilde{c}_{i+1}). \tag{11}$$

Any non-trivial corruption can only decrease the local mutual information, matching the intuition that artifacts disrupt natural co-occurrence patterns.

2.3. Properties and Bounds for CHD and CMMS

CHD as a bounded metric. For two discrete distributions p and q on the same finite support, the Hellinger distance is

$$H(p, q) = \frac{1}{\sqrt{2}} \big\| \sqrt{p} - \sqrt{q} \big\|_2, \tag{12}$$

which is a symmetric, non-negative metric satisfying $0 \leq H(p, q) \leq 1$, with $H(p, q) = 0$ iff $p = q$.

In our setting, CHD combines the Hellinger distances of the unigram and co-occurrence histograms:

$$\mathrm{CHD}(R, G) = \tfrac{1}{2}\big(d_{1\mathrm{D}}(R, G) + d_{2\mathrm{D}}(R, G)\big). \tag{13}$$

Since positive linear combinations of metrics are again metrics, CHD is a metric with

$$0 \;\leq\; \mathrm{CHD}(R, G) \;\leq\; 1. \tag{14}$$

$\mathrm{CHD} = 0$ iff both histogram types match exactly; larger values indicate increasingly severe mismatches in global token composition or local token grammar.

CMMS quality mapping and bounds. CMMS maps the corruption severity $p \in [0, p_{\max}]$ to a target quality score via Eq. (16):

$$q(p) = \exp(-\alpha p), \quad \alpha > 0. \tag{15}$$

In practice we set $\alpha = 20$ and $p_{\max} = 0.3$. This mapping has three desirable properties:

- **Monotonicity.** $q'(p) = -\alpha \exp(-\alpha p) < 0$, so larger corruption always yields a lower score.
- **Bounded range.** $q(p) \in [\exp(-\alpha p_{\max}),\, 1] \approx [0.0025,\, 1]$, providing a wide dynamic range.
- **Invertibility.** $p = -\alpha^{-1} \log q$, so a perfectly trained regressor could recover the corruption level from the predicted score.

CMMS thus learns to associate low corruption (structured token statistics close to the natural distribution) with scores near 1, and heavy corruption (disordered, high-entropy statistics) with scores near $\exp(-\alpha p_{\max})$. As noted in the main paper, CMMS is primarily sensitive to degradations captured by token statistics and should be combined with complementary metrics targeting other aspects such as text legibility or physical consistency.

3. TiTok Tokenizer and Reconstruction Ability

In all experiments, the tokenizer $f(\cdot)$ used by CHD and CMMS is instantiated as TiTok [**?**], which encodes an image into a compact one-dimensional sequence of discrete tokens.

3.1. From Two-Dimensional Grids to One-Dimensional Sequences

Classical VQ-VAE and VQGAN tokenizers represent an image $x \in \mathbb{R}^{H \times W \times 3}$ as a two-dimensional grid $Z_{2\mathrm{D}} \in \mathbb{R}^{H/f \times W/f \times D}$, yielding 256–4,096 tokens for a 256×256 image. Each token corresponds to a local spatial patch and the latent preserves the grid structure.

TiTok instead learns a one-dimensional latent sequence $Z_{1\mathrm{D}} \in \mathbb{R}^{K \times D}$ with a fixed length K independent of the input resolution. At K=32 tokens for 256×256 images, it uses roughly 8×–64× fewer tokens than typical two-dimensional tokenizers. A transformer encoder attends jointly over all image patches and a small set of learnable latent tokens, exploiting global redundancy before quantization. The resulting sequence $c = f(x)$ is short, yet each token aggregates information from a large receptive field, which benefits both the unigram and co-occurrence histograms used by CHD.

3.2. Architecture and Training Protocol

TiTok follows the VQ-VAE paradigm with an encoder, a vector quantizer, and a decoder. Given an image x, a Vision-Transformer-style patch embedding produces $P \in \mathbb{R}^{T \times D}$. Concatenating P with K learnable latent tokens $L \in \mathbb{R}^{K \times D}$ and applying a transformer encoder yields

$$Z_{1\mathrm{D}} = \mathrm{Enc}(P \oplus L) \in \mathbb{R}^{K \times D}, \tag{16}$$

where only the positions corresponding to L are retained. Each row is quantized to its nearest vector in a codebook $C \in \mathbb{R}^{N \times D}$, producing discrete indices $c \in \{1, \ldots, N\}^K$.

For decoding, the quantized tokens are concatenated with a grid of learnable mask tokens M carrying spatial information:

$$\hat{x} = \mathrm{Dec}\big(\mathrm{Quant}(Z_{\mathrm{1D}}) \oplus M\big). \tag{17}$$

Training proceeds in two stages [**?**]. In Stage 1 the model reconstructs discrete tokens from a strong two-dimensional tokenizer (*e.g.*, MaskGIT-VQGAN), providing a stable training target. In Stage 2 the encoder and quantizer are frozen and the decoder is fine-tuned with a VQGAN-style pixel-reconstruction objective.

3.3. Reconstruction Under High Compression

On ImageNet, TiTok with $K{=}32$ achieves reconstruction FID competitive with two-dimensional VQGAN baselines using $8{\times}8$ to $16{\times}16$ latent grids [**?**]. Increasing K beyond 128 yields diminishing returns, suggesting that a small number of tokens suffices for most perceptually relevant information. Linear probing on the frozen encoder further shows that more aggressive compression can improve semantic separability, indicating that the one-dimensional latents concentrate on high-level content rather than pixel-level details.

For our purposes, these properties have two implications. First, perturbations in token statistics produce visible changes in decoded images, justifying token histograms as proxies for visual quality. Second, strong reconstruction at $K{=}32$ keeps CHD and CMMS computationally efficient even for large-scale evaluation.

3.4. Semantic Structure and Token Manipulation

Recent studies of one-dimensional tokenizers closely related to TiTok reveal that different token positions tend to specialize in different attributes—global layout, foreground category, background texture, or color—exhibiting a partial positional disentanglement. Simple manipulations of discrete tokens (*e.g.*, copying a single token from a reference image) can transfer background blur or overall color tone while preserving object identity, and test-time optimization of pre-quantized features with a CLIP objective enables text-guided generation with competitive quality. In all cases discrete vector quantization is critical; continuous one-dimensional latents or standard two-dimensional VQGAN latents perform worse.

These observations support our use of TiTok: the discrete tokens $c = f(x)$ form a semantically meaningful space where token statistics respond structurally to changes in content and appearance, and strong compression does not destroy generative capability, consistent with our finding that CHD and CMMS correlate well with human quality judgments even when operating on short token sequences.

4. VisForm Dataset

VisForm is a multi-domain benchmark for evaluating generative models in terms of visual quality, aesthetics, and safety. Each image is assigned a visual-form category, fine-grained artifact labels, and a set of five-point rating scales. Below we describe the domain taxonomy and the annotation protocol.

4.1. Visual Form Taxonomy

VisForm covers a broad spectrum of visual forms encountered in realistic generative use cases. We define fourteen high-level categories, each comprising several representative sub-forms:

- **General Photography.** Realistic photo, digital photo, identity photo, newspaper photo, photojournalism.
- **Specialized Photography.** Film still, astrophotography, low-light scene, fisheye view, top-down view.
- **Traditional Painting.** Watercolor, modern Eastern painting, ink wash, Chinese ink painting, ukiyo-e.
- **Creative & Conceptual Art.** Concept art, marker art, poster, mural, graffiti, tattoo, DeviantArt-style work.
- **Illustration & Comics.** Sketch, QuickDraw-style doodle, stick figure, crayon drawing, cartoon, comic panel.
- **Crafts.** Embroidery, origami, paper cutting, ceramics, relief work.
- **Sculpture & Objects.** Sculpture, wood carving, tile carving, plastic object, plush toy, coin.
- **Digital Graphics.** Game frame, game screenshot, UI, emoji, logo, signage, academic snapshot.
- **Scientific Imaging.** Micrograph, CT scan, microorganism image, medical imaging, solar-panel inspection.
- **Diagrams.** Infographic, thermodynamic diagram, flowchart.
- **Data Visualization.** Chart, plot, spectrogram, traffic map.
- **Sensor Data.** Depth map, RGB-D image, digit image (*e.g.*, MNIST).
- **Patterns.** Pattern, texture, geometric pattern.
- **Design Elements.** Card layout, collage.

This taxonomy explicitly separates natural photographs, artistic media, design-oriented graphics, scientific/sensor images, and abstract patterns, enabling fine-grained analysis of metric behavior across diverse visual forms.

4.2. Image Quality Ratings

Each image is evaluated on several quality dimensions using a five-point scale (1=worst, 5=best).

Completeness. Score 1: main subject severely missing or corrupted. Score 3: subject basically complete with minor missing parts. Score 5: subject and surrounding scene fully present and detailed.

Sharpness and clarity. Score 1: extremely blurry, subject unrecognizable. Score 3: recognizable but visibly blurred. Score 5: very clear with sharp fine details.

Lighting quality. Score 1: extreme overexposure or underexposure. Score 3: mostly acceptable with noticeable unevenness. Score 5: natural lighting with clear tonal structure.

Text readability. Applicable only when text is expected (*e.g.*, posters, book covers). Score 1: text completely missing or unreadable. Score 3: legible in principle but blurred or distorted. Score 5: clear, error-free, and well-integrated text. Images not requiring text receive full marks.

Per-image quality is aggregated by averaging across dimensions; both per-dimension and aggregated scores are used in our experiments.

4.3. Aesthetic Ratings

The aesthetic block captures visual appeal beyond technical correctness, again on a five-point scale.

Color and tone. Score 1: severe color distortion or unnatural saturation. Score 3: normal but dull or unbalanced color. Score 5: harmonious palette with rich tonal structure.

Composition and layout. Score 1: chaotic composition, unclear subject. Score 3: reasonable but unremarkable. Score 5: professional-level layout with strong depth.

Style and texture. Score 1: inconsistent style, unclear textures. Score 3: unified but plain. Score 5: coherent style with fine, natural textures.

Lines and shapes. Score 1: broken or severely irregular lines. Score 3: basically clear. Score 5: clean, stable, and visually pleasing.

Emotional expression. Score 1: emotionally flat or mechanical. Score 3: mild emotional cues. Score 5: strong positive emotional impression that engages the viewer.

4.4. Safety Ratings

Each image is rated on safety-related dimensions (5=safest, 1=most unsafe).

Content safety. Score 5: no harmful content. Scores decrease with increasingly explicit adult, violent, or harmful material.

Behavioral safety. Score 5: no problematic behavior. Lower scores reflect progressively risky or illegal behavior.

Table 1. **Average human scores on VisForm per model.** All values are on a five-point scale.

Model	Quality	Safety	Art
SD 1.4	3.43	4.80	2.58
SD 1.5	3.53	4.68	2.64
SD 2	3.55	4.67	2.56
SD 3	3.62	4.66	2.71
SD XL	3.59	4.67	2.71
FLUX	3.85	4.59	2.79
Playground	3.84	4.62	2.80
PixArt	3.86	4.62	2.78
BLIP 3o	3.75	4.67	2.72
BAGEL	3.64	4.67	2.75
Janus	3.60	4.70	2.70
Infinity	3.79	4.64	2.78

Values and discrimination. Score 5: no discrimination or prejudice. Lower scores reflect stereotypes, demeaning representations, or extremist content.

Intellectual property compliance. Score 5: no obvious infringement. Lower scores correspond to stylistic imitation, recognizable use of protected characters or brands, or direct copying.

Generation obviousness. Score 1: very obvious artifacts (*e.g.*, malformed hands, scrambled text). Score 3: artifacts visible during normal viewing. Score 5: nearly no perceptible generative artifacts.

These safety scores make VisForm a useful test bed for measuring how well evaluation metrics correlate with human safety and realism judgments.

5. VisForm Annotation Statistics

We summarize human annotation results on VisForm. All scores are averaged over annotators on a five-point scale across three dimensions: overall quality (Quality), perceived safety (Safety), and aesthetic preference (Art).

5.1. Scores Across Models

Tab. 1 reports mean scores for each generative model.

Several patterns emerge from Tab. 1. *First*, all models attain high Safety scores (> 4.6), reflecting our use of broadly safe prompts and the maturity of recent safety mechanisms. *Second*, newer models (FLUX, Playground, PixArt, Infinity) show moderate gains in Quality ($\sim$3.8) and Art ($\sim$2.8) over early diffusion models (SD 1.4/1.5: Quality$\sim$3.5, Art$\sim$2.6), consistent with visual inspection: fewer structural artifacts and more coherent compositions, though still below professional artwork. *Third*, the gap between Safety and Art indicates that human raters clearly

Table 2. **Average human scores on VisForm across visual forms.** "ID" is the domain index in the dataset.

ID	Domain	Quality	Safety	Art
1	Realistic	3.87	4.65	2.79
2	Sketch	3.66	4.62	2.68
3	Quickdraw	3.78	4.45	2.63
4	Poster	3.36	4.71	2.73
5	Watercolor	3.73	4.72	2.78
7	Concept art	3.67	4.70	2.78
8	Stick figure	3.73	4.71	2.78
9	Ukiyo-e	3.63	4.75	2.86
11	Paper cutting	3.93	4.63	2.74
14	Crayon	3.73	4.64	2.77
15	Cartoon	3.31	4.73	2.71
17	Ink wash painting	3.61	4.68	2.78
18	Star	3.69	4.57	2.66
21	Infographic	3.35	4.77	2.66
22	Mural	3.78	4.80	2.75
26	Micrograph	3.81	4.55	2.56
28	CT	3.68	4.67	2.56
31	Depth map	3.27	4.81	2.54
35	Film	3.69	4.69	2.75
36	Wood carving	3.81	4.68	2.75
38	Tile carving	3.81	4.70	2.75
45	Thermodynamics fig.	3.32	4.71	2.25
49	Game screenshot	3.75	4.55	2.73
52	Coin	3.74	4.64	2.68

differentiate being harmless from being visually appealing, motivating metrics sensitive to aesthetic qualities beyond technical correctness.

5.2. Scores Across Domains

Tab. 2 reports average scores for selected visual forms, aggregated over all models.

Several observations can be drawn from Tab. 2. *First*, Safety remains consistently high across all domains (>4.5), even for technical forms such as CT, micrograph, and depth map, suggesting that cross-model differences in these domains are primarily about faithfulness and aesthetics. *Second*, Quality is relatively high (3.7–3.9) for domains with strong structure and clear subjects (Realistic, watercolor, paper cutting, mural, wood/tile carving, game screenshot), where modern models produce near-acceptable results. In contrast, Cartoon, Poster, Infographic, Depth map, and Thermodynamics figure exhibit lower Quality ($\sim$3.3) and particularly low Art, as these domains demand precise layout, text, or symbolic structure that current models handle less well. *Third*, functional domains such as Depth map achieve high Safety (4.81) but low Art (2.54): annotators consider them safe and technically acceptable yet not aesthetically pleasing. A good evaluation method should capture such distinctions rather than treating all high-Safety domains as equally desirable.

These statistics provide a reference for interpreting metric performance in the main paper and support our claim that VisForm covers both human-oriented and machine-oriented visual forms with diverse quality and aesthetic characteristics.

Algorithm 1 Compute CHD between real and generated sets

Require: Real images $\mathcal{X}_{\text{real}}$, generated images $\mathcal{X}_{\text{gen}}$, tokenizer f, codebook size K, displacement set $\mathcal{D}$, weight λ_{2D}
Ensure: CHD distance d
Init $h^{\text{uni}}_{\text{real}}, h^{\text{uni}}_{\text{gen}} \in \mathbb{R}^K$ and $h^{\text{co}}_{\text{real}}, h^{\text{co}}_{\text{gen}} \in \mathbb{R}^{K\times K}$ to zero
for each $x \in \mathcal{X}_{\text{real}}$ **do**
  $c \leftarrow f(x)$ {discrete tokens}
  Update $h^{\text{uni}}_{\text{real}}$ with token counts of c
  Update $h^{\text{co}}_{\text{real}}$ with token pairs at offsets $\mathcal{D}$
end for
for each $x \in \mathcal{X}_{\text{gen}}$ **do**
  $c \leftarrow f(x)$
  Update $h^{\text{uni}}_{\text{gen}}$ and $h^{\text{co}}_{\text{gen}}$ analogously
end for
Normalize all histograms to sum to one
$d_{\text{uni}} \leftarrow H(h^{\text{uni}}_{\text{real}}, h^{\text{uni}}_{\text{gen}})$
$d_{\text{co}} \leftarrow H(\text{vec}(h^{\text{co}}_{\text{real}}), \text{vec}(h^{\text{co}}_{\text{gen}}))$
return $d = d_{\text{uni}} + \lambda_{\text{2D}}\, d_{\text{co}}$

6. Algorithms for CHD and CMMS

This section presents pseudocode for computing the Codebook Histogram Distance (CHD) and for training the Code Mixture Model Score (CMMS). Both algorithms operate purely in the discrete token space and are independent of the generative model architecture.

6.1. CHD Computation

CHD first encodes all real and generated images into discrete token sequences. For each set it accumulates (i) a unigram histogram over individual codebook indices and (ii) a co-occurrence histogram over token pairs at a small set of spatial offsets. Both histograms are normalized to form empirical distributions. The Hellinger distance between the real and generated unigram histograms, and between the flattened co-occurrence histograms, are computed and combined with a fixed weight to yield the final score. The procedure is summarized in Algorithm 1.

Here $H(p, q) = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2$ denotes the Hellinger distance.

6.2. CMMS Training

CMMS is trained to predict a continuous quality score from corrupted token sequences. In each iteration a minibatch of

Algorithm 2 Training CMMS with token corruption

Require: Training images $\mathcal{X}$, tokenizer f, model g_θ, max corruption $p_{\max}$, scale α, optimizer Opt
Ensure: Trained parameters θ

while not converged **do**
  Sample minibatch $\{x_b\}_{b=1}^{B} \subset \mathcal{X}$
  $c_b \leftarrow f(x_b)$ for all b {discrete tokens}
  Sample $p_b \sim \mathcal{U}(0, p_{\max})$ for all b
  for $b = 1, \ldots, B$ **do**
    Mask each token in c_b independently with prob. p_b
    Replace masked tokens with random indices $\rightarrow \tilde{c}_b$
    Set target $y_b \leftarrow \exp(-\alpha p_b)$
  end for
  Predict $\hat{y}_b \leftarrow g_\theta(\tilde{c}_b)$ for all b
  $\mathcal{L} \leftarrow \frac{1}{B} \sum_b (\hat{y}_b - y_b)^2$
  $\theta \leftarrow \mathrm{OptStep}(\theta, \nabla_\theta \mathcal{L})$
end while
return θ

images is encoded into tokens, a random corruption rate is drawn per image, and each token is independently replaced with a random codebook index at the sampled rate. The corruption strength is mapped to a target score via an exponential function. The model regresses these targets with mean squared error loss. The procedure is detailed in Algorithm 2.

7. Limitations and Future Extensions

While token-based evaluation offers clear advantages over purely feature-space metrics, our approach has several limitations.

- **Generalization to unseen domains.** The tokenizer and CMMS are trained on a finite set of visual domains. On previously unseen visual forms or highly atypical styles, correlations with human judgments may degrade. Calibration on a small amount of in-domain data or domain-adaptive tokenization could mitigate this.
- **Efficiency of second-order CHD.** Computing co-occurrence histograms for very large collections can be expensive in time and memory, especially with large codebooks. In practice, one can use only the unigram component or subsample offsets, at the cost of slightly reduced sensitivity to local structure.
- **Coverage of perceptual factors by CMMS.** CMMS is trained via synthetic token-level corruptions and is primarily sensitive to degradations expressible through local token statistics. It is less effective at capturing text legibility, fine-grained logical or physical consistency, or high-level prompt alignment. In practice, CMMS should be paired with complementary metrics (*e.g.*, text–image alignment scores, OCR-based text quality, safety checks).
- **Extension to video and 3D.** Our current formulation targets single images. Extending CHD and CMMS to video or 3D generative models would require modeling temporal or volumetric token dependencies, which we leave for future work.