Whitened CLIP as a Likelihood Surrogate of Images and Captions

Roy Betser [1] **Meir Yossef Levi** [1] **Guy Gilboa** [1]

Abstract

Likelihood approximations for images are not trivial to compute and can be useful in many applications. We examine the use of Contrastive Language-Image Pre-training (CLIP) to assess the likelihood of images and captions. We introduce *Whitened CLIP*, a novel transformation of the CLIP latent space via an invertible linear operation. This transformation ensures that each feature in the embedding space has zero mean, unit standard deviation, and no correlation with all other features, resulting in an identity covariance matrix. We show that the whitened embeddings statistics can be well approximated as a standard normal distribution, thus, the log-likelihood is estimated simply by the square Euclidean norm in the whitened embedding space. The whitening procedure is completely training-free and performed using a pre-computed whitening matrix, hence, is very fast. We present several preliminary experiments demonstrating the properties and applicability of these likelihood scores to images and captions. Our code is available here.

1. Introduction

Computing likelihoods for images is a challenging yet valuable task with numerous applications in computer vision, such as image generation (Ramesh et al., 2022) and editing (Kawar et al., 2023). Traditional approaches, including diffusion models, primarily rely on the likelihood gradient or score function, limiting direct likelihood computation (Ho et al., 2020; Song et al., 2020).

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) has become a widely adopted embedding for dual text-image semantics. However, its potential as a likelihood surrogate remains unexplored. This paper introduces *Whitened CLIP* (W-CLIP), a linear whitening transformation of the CLIP latent space, where each feature is standardized to have zero mean and identity covariance. In this whitened space, we validate by statistical tests that the embeddings approximate normal distribution, hence negative log-likelihood estimations are a function of the Euclidean norm in the transformed space.

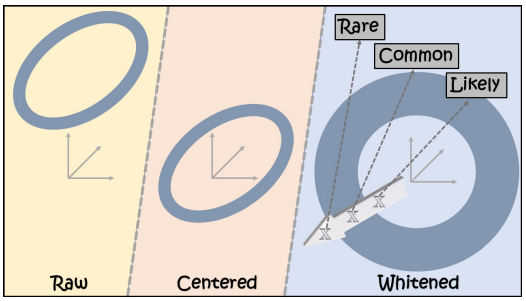


Figure 1. **Raw, centered and whitened CLIP geometry.** The whitened CLIP space is isotropic, transforming the original ellipsoid shaped space into an hypersphere. In this space, the embedding norm reflects likelihood level. Higher norms correspond to lower probabilities.

To the best of our knowledge, this represents the first direct computation of likelihood functions for images and text prompts under the CLIP-learned distribution.

Our main contributions are as follows:

1. We propose *Whitened CLIP* (W-CLIP), based on an invertible linear operation, allowing likelihood assessments while retaining the generative and semantic capabilities of CLIP.
2. We perform quantitative statistical experiments using Anderson-Darling and D'Agostino-Pearson tests, indicating the features in the whitened space can be well approximated by a normal distribution.
3. We are the first to propose a direct computation of likelihood functions for images and text prompts, under the CLIP learned distribution. For images, to the best of our knowledge, this is the first direct likelihood computation with semantic capabilities.

[1]Viterbi Faculty of Electrical and Computer Engineering, Technion - Israel Institute of Technology, Haifa, Israel.. Correspondence to: Roy Betser <roybe@campus.technion.ac.il>, Meir Yossef Levi <me.levi@campus.technion.ac.il>, Guy Gilboa <guy.gilboa@ee.technion.ac.il>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

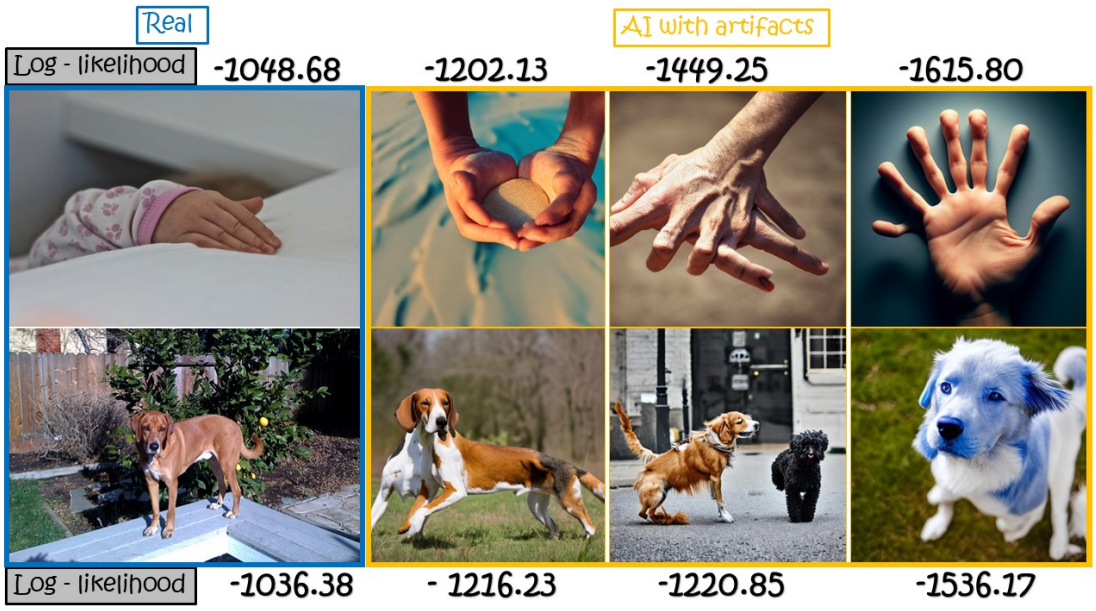


Figure 2. **Log-likelihood of real and generated images with artifacts.** Real images of a hand and a dog (left) and three similar AI generated images with artifacts. Real images have higher log-likelihood than generated images with artifacts.

4. We show W-CLIP can be used to estimate probability drifts in generative models, discover artifacts in image generation and rank statistical deviation of out-of-distribution (OOD) benchmarks, such as ImageNet-C and ImageNet-R, compared to in-distribution (ID) sets.

5. For image manipulation, we use W-CLIP to extend Spherical Linear Interpolation (SLERP) by introducing full-circle SLERP, enabling both interpolation and extrapolation between two given images.

2. Related Work

Estimating the likelihood of images, $P(X)$, is a fundamental task with numerous downstream applications, including super-resolution (Li et al., 2022a; Gao et al., 2023), denoising (Tian et al., 2020; Goyal et al., 2020), and inpainting (Yu et al., 2018; Elharrouss et al., 2020). Early approaches relied on assumptions about natural image smoothness (Geman & Geman, 1984; Ruderman & Bialek, 1993) and patch distribution (Zoran & Weiss, 2011). Generative models such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), Autoencoders (AEs) (Hinton & Salakhutdinov, 2006; Kingma, 2013), Energy-Based Models (EBMs) (Du & Mordatch, 2019; Ou et al., 2024) and Diffusion models (Ho et al., 2020; Song et al., 2020) have further advanced image synthesis by implicitly estimating $P(X)$. However, these methods do not provide explicit access to $P(X)$; for instance, diffusion models approximate the score function, $\nabla_x \log P(X)$, rather than $P(X)$ itself.

In natural language processing (NLP), large language models (LLMs) estimate probabilities directly (Devlin, 2018; Brown et al., 2020), while vision-language models (VLMs), including CLIP (Radford et al., 2021) and other recent models (Desai et al., 2023; Chou & Alam, 2024), embed images and text into a shared space. Despite its success in enabling applications such as captioning (Mokady et al., 2021) and image manipulation (Kawar et al., 2023), CLIP's latent space remains underexplored. Known phenomena include the *Narrow Cone Effect*, where embeddings occupy limited angular space (Schrodi et al., 2024) and the *Modality Gap*, where image and text distributions are disjoint (Liang et al., 2022; Shi et al., 2023; Levi & Gilboa, 2025). Mokady et al. (2021) introduce a mapping network to bridge the modality gap for image captioning.

To the best of our knowledge, this work is the first to analyze CLIP embeddings from a probabilistic perspective and to propose leveraging its latent space as a probability estimator, particularly for the challenging domain of images.

3. Method: CLIP Likelihoods

3.1. Notations

Let $\mathbf{X} = \{x_1, \cdots x_N\}$ be a set of N random vectors of dimension d, $x_i \in \mathbb{R}^d$, where $\mu = \frac{1}{N}\Sigma_1^N x_i$ is the mean vector. We denote by $\hat{x}_i = x_i - \mu$ the centered vector, where $\hat{\mathbf{X}} = \{\hat{x}_1, \cdots \hat{x}_N\}$. Let $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ be the empirical

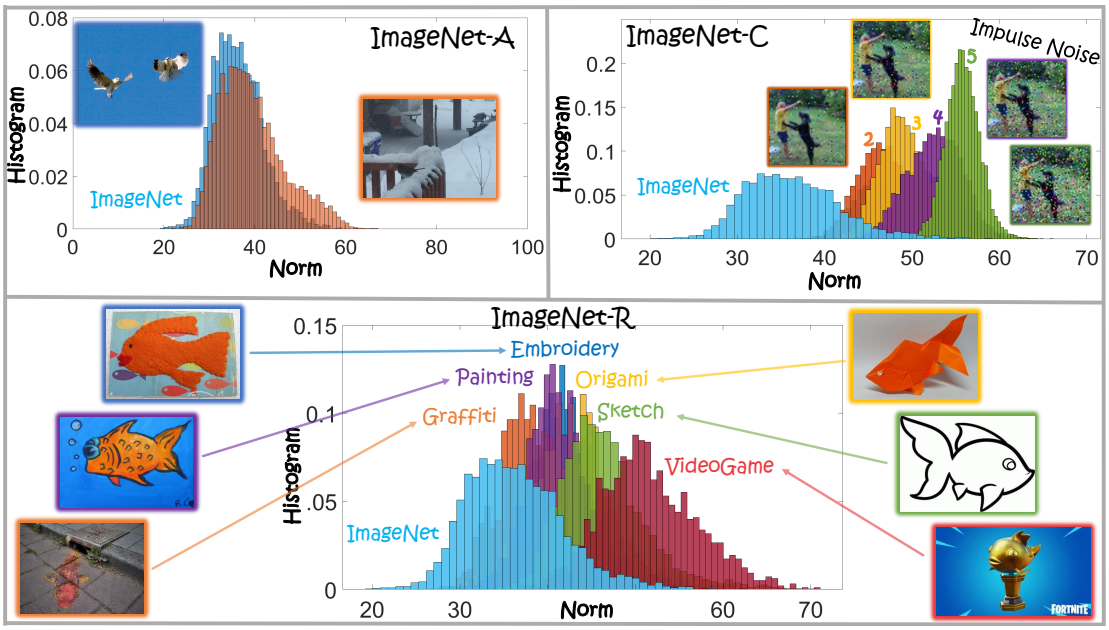


Figure 3. **Norm histograms of ImageNet variations.** Top left: ImageNet-A, comprising of natural adversarial examples, closely aligns with clean ImageNet due to their natural origins. Top right: ImageNet-C histograms under varying impulse noise levels of severity display significantly larger norms than clean ImageNet, indicating distributional deviations. Bottom: ImageNet-R comparison shows that different styles cause varying likelihood shifts, with graffiti closest to real images and video game renditions exhibiting the largest shifts.

covariance matrix of X:

$$\mathbf{\Sigma} = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^\top. \tag{1}$$

We recall that the covariance matrix is symmetric, positive semi definite and that the diagonal contains the variance of each feature in the vector.

3.2. Whitening transform

Given a set of random vectors with a non-singular covariance matrix $\mathbf{\Sigma}$, let $W \in \mathbb{R}^{d\times d}$ be a matrix that satisfies $W^\top W = \mathbf{\Sigma}^{-1}$. We note that W is not unique. A common way to obtain it is by *principal component analysis* (PCA). Let us diagonalize the covariance matrix $\mathbf{\Sigma} = V\Lambda V^\top$, where $\Lambda \in \mathbb{R}^{d\times d}$ is a diagonal matrix of the eigenvalues, λ_i, and $V \in \mathbb{R}^{d\times d}$ consists of the corresponding eigenvectors. Then, the *whitening matrix* W can be defined as:

$$W = \Lambda^{-\frac{1}{2}}V^\top. \tag{2}$$

Note that W is an invertible matrix. A single vector x is *whitened* by $y = W\hat{x}$ and the *whitened matrix* $\mathbf{Y} \in \mathbb{R}^{d\times N}$ corresponding to the raw measurement matrix $\mathbf{X}$ is

$$\mathbf{Y} = W\hat{\mathbf{X}}. \tag{3}$$

$\mathbf{Y}$ consists of *isotropic* random vectors, that is, each vector has zero mean and an identity covariance matrix ($\mu_Y = 0, \mathbf{\Sigma}_\mathbf{Y} = \mathbb{I}$) (see App. D.2). The inverse transform from the whitened space to the original space is performed simply by $x = W^{-1}y + \mu$ and in matrix notation $\mathbf{X} = W^{-1}\mathbf{Y} + \mu \cdot \mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{1\times N}$ is a row vector of 1's.

Given a set of raw CLIP embeddings, The whitening procedure offers three key advantages:

1. W is obtained in a purely data-driven process, without additional meta-parameters.
2. Since the transform is invertible, all existing applications developed in the raw embedding space can be seamlessly integrated with this approach.
3. The computation of W is performed only once and *a-priori*, based on a representative dataset. Memory and computational requirements are very mild, allowing efficient use also in low-resource settings.

It is known that the CLIP latent spaces of images and captions are disjoint (Liang et al., 2022; Levi & Gilboa, 2025). Therefore, we treat the distribution of each modality independently. Additional implementation details and the complete whitening algorithm are in App. D.1, Alg. 1.

3.3. Whitened CLIP embeddings

Our likelihood estimation relies on modeling the whitened CLIP space as following approximately independent and

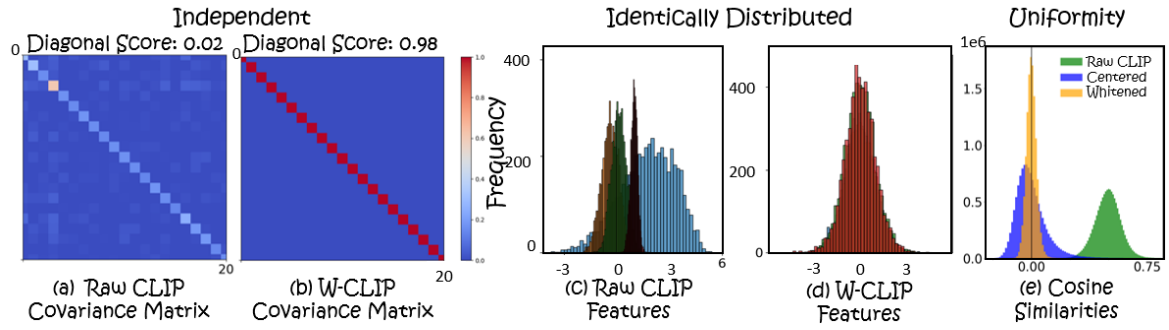


Figure 4. **Raw CLIP and W-CLIP analytic comparison.** The covariance matrices of raw CLIP (a) and W-CLIP (b) demonstrate the effectiveness of the whitening transformation in achieving unit variance and zero correlation among features. Histograms of four CLIP features (c) vary in mean and variance, whereas four W-CLIP features (d) exhibit zero mean and unit variance. Cosine similarity histograms for all image pairs (e) across raw, centered, and W-CLIP embeddings reveal that W-CLIP's cosine similarity is concentrated around zero, indicating significantly improved uniformity compared to the centered and raw CLIP spaces.

identically distributed (i.i.d) standard normal distribution. We thus first examine the validity of this approximation.

Normal distribution tests. To assess how well the whitened embeddings approximate normal distribution, we employ two statistical tests: Anderson-Darling (Anderson & Darling, 1954) and D'Agostino-Pearson (D'agostino & Pearson, 1973). The Anderson-Darling test evaluates how well the empirical cumulative distribution function (CDF) matches the expected CDF of a normal distribution, placing higher weight on the tails to detect deviations. The D'Agostino-Pearson test combines skewness and shape characteristics measures to assess normality, offering sensitivity to both symmetric and asymmetric deviations. See App. D.3 for additional details regarding these tests, specifically Eq. (12), Eq. (13). For stability, the 5000 embeddings of MS-COCO validation set (Lin et al., 2014) are divided into 20 equal groups of 250 samples each. As shown in Tab. 1, the results validate that a normal distribution is a good approximation for both image and text embeddings. Specifically, in both tests, more than 90% of the text features, and more than 98% of the image features conform to a normal distribution, with average scores that satisfy the test criteria by a large margin. Additional details regarding these tests, empirical statistics, and plots are provided in App. D.3.

Independent and identically distributed (i.i.d). Let us break this assumption into independence and identical distributed conditions. For the former, in normal distribution, non-correlation is a sufficient condition for independence. In Fig. 4, the 20 first features of the covariance matrices of raw CLIP embeddings (a) and W-CLIP embeddings (b) are presented. While the CLIP embeddings exhibit correlations between features, the covariance matrix of the whitened embeddings is almost exactly diagonal, indicating that the features are uncorrelated. This is expected, since the whitening transform is designed for exactly this purpose. A metric measuring the proximity of a matrix to being diagonal (in the range $[0, 1]$ with 1 being exactly diagonal) is

$$\text{Diagonal Score} = \frac{\sum_i |\mathbf{\Sigma}_{i,i}|}{\sum_{i,j} |\mathbf{\Sigma}_{i,j}|}, \tag{4}$$

where $\mathbf{\Sigma}_{i,j}$ is an element at row i and column j of the covariance matrix. Scores of the full matrices verify that, provided the normal distribution model is valid, the independence assumption holds as well. Regarding the latter, in Fig. 4 the CLIP features exhibit varied mean and variance values (c), while W-CLIP features have all zero mean and unit variance (d), see further results in Fig. 18, App. D.3. Consequently, the whitened embeddings can be approximated reasonably well as i.i.d. features.

Table 1. **Anderson-Darling and D'Agostino-Pearson scores for image and text embeddings.** The *Avg.* column contains the average score for all features, and *Normal Features* represents the percentage of features passing the normal distribution test based on their average score. The threshold score (in brackets) indicates the required condition for normal distribution, with the sign showing whether higher or lower results imply normal distribution.

	Avg.	Normal Features
	Anderson-Darling (< 0.752)	
Image	0.4890	98.3%
Text	0.5926	90.1%
	D'Agostino-Pearson (> 0.05)	
Image	0.3624	99.3%
Text	0.2568	99.2%

3.4. Log probabilities using W-CLIP

Embedding likelihood. The explicit likelihood of a d-dimensional random vector, x, with i.i.d standard normal variables is:

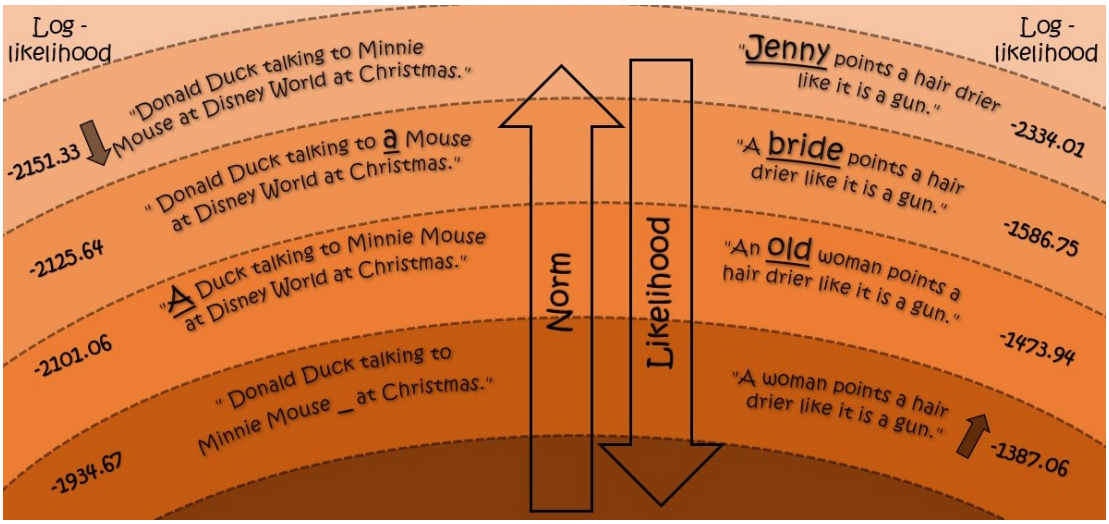


Figure 5. **Likelihood variation for different levels of details.** The original MS-COCO caption is marked with an arrow, with deviations underlined. Left: Removing details, such as character names or locations, increases likelihood. Right: Adding specificity, such as replacing "woman" with "bride" or "Jenny", decreases likelihood.

$$P(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\|x\|^2\right), \tag{5}$$

where $\|x\|^2 = x^\top x$. The log-likelihood is:

$$\ell(x) = \log P(x) = -\frac{1}{2}\left(d\log(2\pi) + \|x\|^2\right). \tag{6}$$

Thus, we propose $\ell(x)$ to be an approximation of the log likelihood of image or caption instances, based on the W-CLIP embedding. To the best of our knowledge, this is the first method to directly obtain a probability score for image or text embeddings using CLIP and the first probability computation for images which is not based on low-level patch statistics but on high-level semantics. In contrast, natural language processing (NLP) language models can directly approximate the negative log-likelihood (NLL) for a text prompt. The relationship between our log-likelihood measure and those of language models is discussed in Sec. 4.4.

Norm distribution in W-CLIP. According to Eq. (6), the norm is directly related to the log-likelihood. We thus highlight some consequences and recall the distribution of norms under standard normal statistics. We first note that the most probable sample resides at the center of the whitened embedding space, nevertheless, the likelihood of sampling this singular point out of the entire space is zero in practice. In general, high-dimensional normal distributions have close to zero mass near the origin. This follows a phenomenon called *Thin Shell* (App. D.4), which reveals that the majority of the distribution is concentrated near the surface of a sphere of radius $\sqrt{d}$. The chi distribution (χ_d), is the appropriate model for the distribution of norms in the whitened space. We denote the norm of x by $S = \sqrt{\sum_{i=1}^{d} x_i^2}$. The log-likelihood of S is:

$$\log(P(S)) = C(d) + \left(\frac{d}{2} - 1\right)\log(S^2) - \frac{1}{2}S^2, \tag{7}$$

where $C(d) = -\log\left(2^{(\frac{d}{2})-1}\Gamma(\frac{d}{2})\right)$ and Γ is the Gamma function. The expected value and standard deviation of S are:

$$\mathbb{E}[S] = \sqrt{2}\,\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}, \quad \mathrm{Std}(S) = \sqrt{d - \mu_S^2}. \tag{8}$$

For large d, $\mathbb{E}[S] \to \sqrt{d - \frac{1}{2}}$. The comparison between theoretical and empirical measurements in Tab. 2, based on MS-COCO, reaffirms the assumed framework of normal distribution. The mean and standard-deviation of the whitened image embeddings closely align with the expected values, while the text embeddings exhibit slightly greater deviation; a trend consistent with the results in Tab. 1.

Table 2. **Empirical and theoretical measurements.** For $d = 768$; relative deviation of the empirical (Emp.) from the theoretical (Theo., Eq. (8)) values are shown in brackets.

	Mean (Emp. / Theo.)	Std (Emp. / Theo.)
Image	27.43/27.7(0.98%)	3.94/3.96(0.55%)
Text	28.49/27.7(2.85%)	5.72/6.60(13.24%)

4. Experiments

All the experiments in this section employ the CLIP ViT-L/14 model and utilize the MS-COCO validation set to

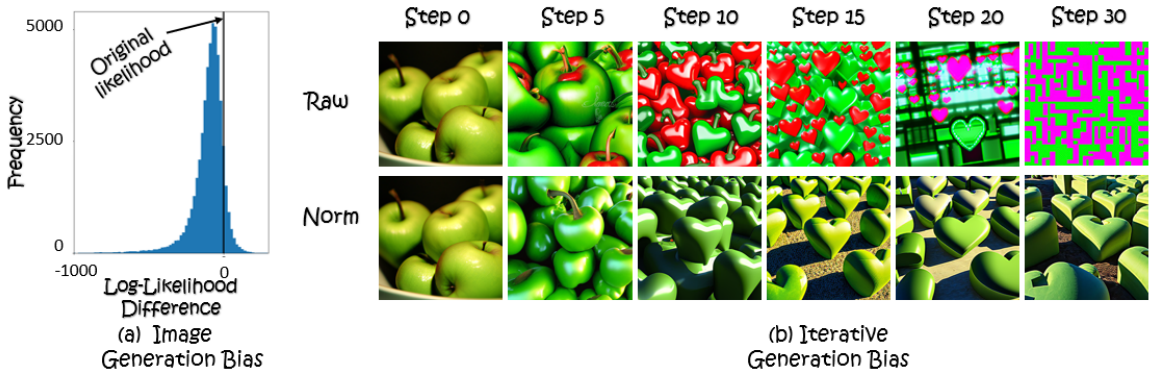


Figure 6. **Bias in image generation.** Left: Using CLIP-encoded images in MS-COCO validation set as a condition for generating new images. The histogram shows a bias towards lower likelihoods in generated images. Right: Iteratively using UnCLIP to generate images encoded by CLIP with a fixed seed. The raw process gradually becomes noisy (Top), whereas with normalization (to $\sqrt{d}$ at each encoding step), the content drifts but remains within the natural and reasonable image space.

compute the whitening matrix W.

4.1. Attributes of W-CLIP

Text complexity. Our observation is that more complex and specific words, such as names, are expected to yield lower likelihood scores. In Fig. 5 we show results of caption editing. Words are replaced with either generic or specific terms, for example, by adding or removing names. The likelihood scores adjust accordingly, decreasing for more specific terms and increasing for more generic ones. Additional examples are available in Figs. 25, 26, App. H.

Uniformity enhancement. An additional desirable property promoted naturally in the whitened space is *uniformity* (Wang & Isola, 2020). Fig. 4.e presents a histogram of cosine similarities between all possible image pairs (predominantly composed of negative examples) in the raw, centered and whitened spaces. In the whitened space, cosine similarities are concentrated near zero with smaller variance. In contrast, the centered space exhibits higher variance, while the raw CLIP space has similarities centered around 0.5 with high variance. These results indicate that the whitened CLIP distribution is more uniform.

4.2. Data Analysis using W-CLIP

Artifact detection. An important attribute of any image likelihood function is its capacity to discriminate between authentic and synthetic images, with particular emphasis on identifying artifacts present in synthetic counterparts. In Fig. 2, we compare the likelihood of real images, and AI-generated ones from the SynArtifact dataset (Cao et al., 2024) containing notable artifacts. All generated images have lower likelihoods than their real counterparts. Additional examples are provided in Fig. 9, App. B.1.

Domain shift. Fig. 3 evaluates a subset of ImageNet (Deng et al., 2009), as presented in Kan et al. (2018), in comparison to ImageNet-A (Hendrycks et al., 2021b), ImageNet-C (Hendrycks & Dietterich, 2019), and ImageNet-R (Hendrycks et al., 2021a). Here we show the distribution of *norms* of each set, instead of the *likelihood* estimation. Following Eq. (6), we have

$$\|x\| = \sqrt{-2\ell(x) - d\log(2\pi)},$$

where $\ell(x)$ is the log-likelihood estimation of x. Thus it is a simple monotonic transformation, which in some cases may serve as an alternative, more convenient, visualization. One should notice that a higher norm indicates lower likelihood. ImageNet-A consists of natural adversarial images. Since the images are natural, their norm distribution is similar to that of ImageNet, apart from a slight shift toward higher values. ImageNet-C introduces common corruptions (e.g., impulse noise), with higher noise levels corresponding to lower likelihoods and distributions consistently below ImageNet. ImageNet-R assesses robustness to domain shifts with renditions like art, graffiti, and video games. Renditions closer to real images, like graffiti, have lower norms than video games, but all renditions exhibit higher norms than ImageNet. For additional Imagenet-C corruptions see Fig. 11, App. B.3.

4.3. Image manipulations

Image generation bias and variance. Generative models may produce outputs which are more likely or less likely than intended. Here we give an example how this can be quantified, allowing to obtain likelihood-bias and likelihood-variance of a generator, as shown in Fig. 6. Image generation was performed using UnCLIP (Ramesh et al., 2022), conditioned by a CLIP embedding. In this experiment, each image from MS-COCO validation set was encoded using CLIP and subsequently ten images were generated by UnCLIP with

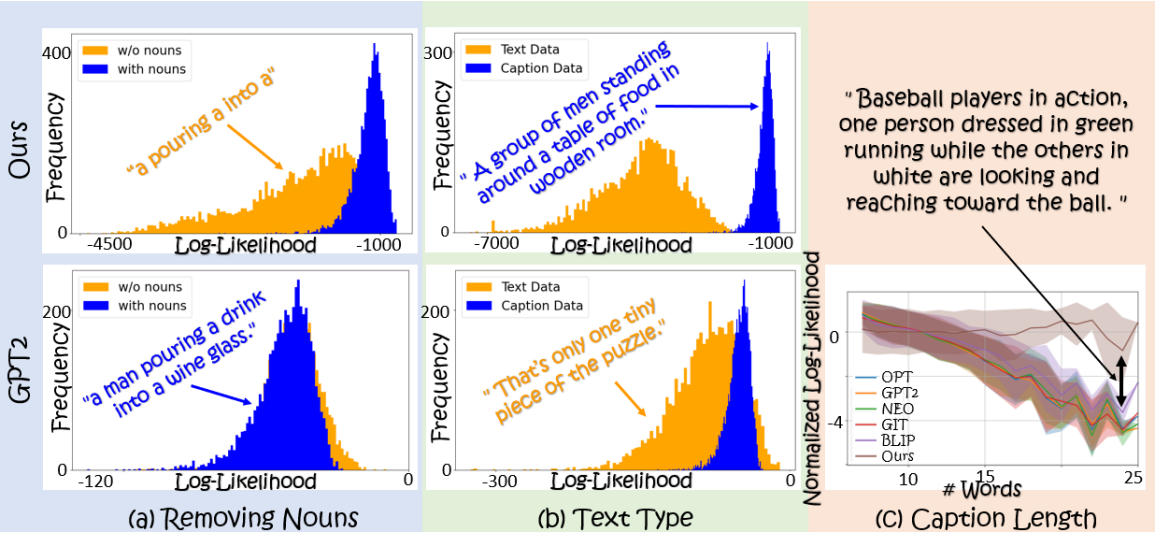


Figure 7. **Differences between likelihood functions.** Our proposed likelihood estimation is highly sensitive to grammatical errors (a), demonstrated by the removal of all nouns from the captions, and text type (b), where *Text Data* refers to a general text dataset (OpenWebText) and *Caption Data* refers to MS-COCO captions. However, it remains less sensitive to caption length (c). In contrast, language models are highly sensitive to caption length and treat captions as being within the distribution of general text. The removal of nouns from the captions causes only negligible changes to the overall distribution of the model's likelihood. Histograms of all language models are in Figs. 12, 13, App. C.

different random seeds. A histogram showing the likelihood differences between the original and the generated embeddings is provided in Fig. 6.a. The results demonstrate a clear bias towards lower likelihoods. This result indicates that our likelihood approximation method can potentially be leveraged as a generated image detector. Thorough investigation of this task will be conducted in future work.

To further understand this phenomena, we implemented an iterative sequence, starting with a real image encoded by CLIP. The resulting embedding is used to generate a new image, which is re-encoded to CLIP. This iterative process caused the generated images to drift in content, quickly degrading into noise. According to the chi distribution, an embedding is most likely to have a norm of approximately $\sqrt{d}$ (Eq. (8)). We use this to normalize each embedding in W-CLIP to a norm of $\sqrt{d}$ and project back to CLIP space, mitigating this issue. While the iterative process still has a semantic drift, reasonable images were consistently produced. Fig. 6.b illustrates this process. Additional experiments are provided in App. G (Figs. 23, 24).

Full circle SLERP. Ramesh et al. (2022) propose spherical interpolation (SLERP) on image CLIP embeddings to interpolate between images. SLERP is defined as:

$$\text{SLERP}(t;\mathbf{E}_1,\mathbf{E}_2)=\frac{\sin((1-t)\theta)}{\sin(\theta)}\mathbf{E}_1+\frac{\sin(t\theta)}{\sin(\theta)}\mathbf{E}_2, \quad (9)$$

where $\mathbf{E}_1$ and $\mathbf{E}_2$ are embeddings, θ is the angle between them (calculated as the normalized dot product), and $t\in[0,1]$ is the interpolation step. SLERP assumes embeddings lie on a hypersphere (Liang et al., 2022; Wang & Isola, 2020) and is mathematically valid for t beyond $[0,1]$. Fig. 20 (App. F) illustrates SLERP on a 2D circle. When one point is off the circle, interpolation forms an ellipse near the perimeter; if shifted from the origin, the ellipse deviates significantly further. Full-circle SLERP uses an interpolation degree ω, with $t=\frac{\omega}{\theta}$ in Eq. (9). In the raw CLIP space, full-circle SLERP often produces noise, with reasonable images only near and between the original embeddings. In the whitened space, it generates consistent images across all angles, with semantic diversity, indicating embeddings remain within the distribution. Images from full circle SLERP examples are in Fig. 8 and App. F (Fig. 21). In Fig. 8, at 300 degrees, which is extrapolation of the source embedding to the reverse direction from the destination embedding, the dog with a bottle of bear (source embedding) becomes a man sitting next to bottles of bear. This is an interesting extrapolation result, not specifically guided. In order to further evaluate this phenomena, and quantify it to quantitative measures we perform an additional experiment. Using MS-COCO validation set, for each image, we performed full-circle SLERP in both the raw CLIP and W-CLIP embedding spaces. In this process, a source image is interpolated toward a destination image along a circular path within the embedding space. Crucially, the image generated at the 180° position from the source—referred to as the "opposite image" (generated from the "opposite

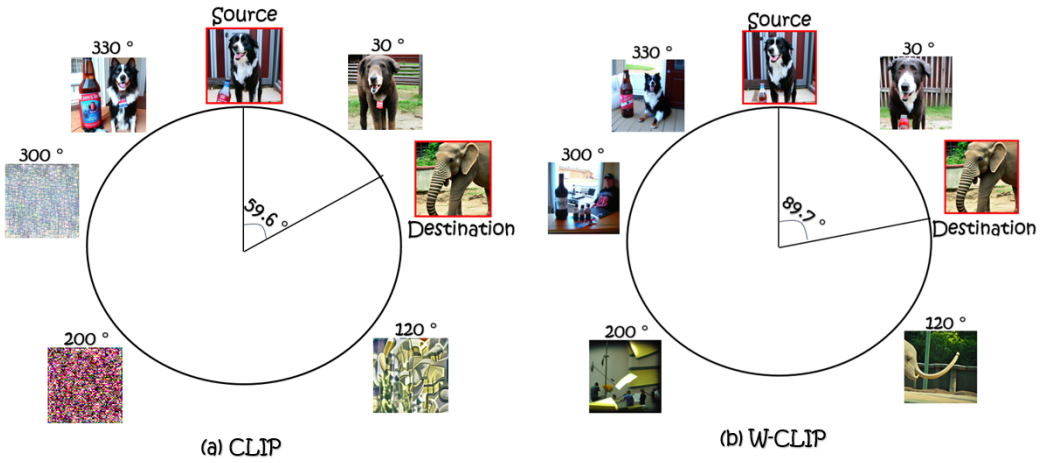


Figure 8. **Full circle SLERP example.** The full circle SLERP is performed in both the raw CLIP space (a) and in the W-CLIP space (b). The different angle between embeddings in both space is presented. In the raw CLIP space the full circle SLERP results with noise for most of the degrees not between the source and destination embeddings. In the W-CLIP space for all degrees real images are generated.

embedding")—is invariant to the chosen destination and determined solely by the source. While other positions along the path are influenced by the destination embedding, the opposite embedding is a fixed, symmetric counterpart.

We generate these opposite images using both CLIP and W-CLIP embeddings and observe a stark contrast: in the CLIP space, opposite images degrade into structured noise, whereas in the W-CLIP space, they remain visually natural and semantically meaningful, as shown in Fig. 8. The structured noise produced by CLIP exhibits 4×4 pixel blocks and a restricted color palette, suggesting synthetic artifacts. We provide a large visual example in App. F (Fig. 22).

To quantify these differences, we compute Total Variation (TV), Entropy, and the percentage of extreme saturation values (top or bottom 1% of the pixel range). All metrics are computed per channel and averaged per image across three sets: original MSCOCO images, CLIP opposites, and W-CLIP opposites. The results are summarized in Tab. 4. These findings confirm that W-CLIP opposites are statistically similar to natural images, whereas CLIP opposites exhibit significantly reduced entropy and variation and much higher percentage of saturation values, indicating a lack of natural structure.

4.4. Relations to language model probabilities

In natural language processing (NLP), large language models (LLMs) minimize the negative log-likelihood (NLL) during training to learn a probability distribution over sequences. At inference, the NLL is computed by summing the negative log probabilities of each token in the prompt, conditioned on previous tokens. The final NLL is averaged over all tokens, with lower NLL scores indicating higher sequence likelihood under the model's learned distribution (Bishop & Nasrabadi, 2006; Murphy, 2012). Our proposed log-likelihood score (Eq. (6)) approximates the likelihood of text prompts based on a single embedding vector for the entire prompt. We evaluated MS-COCO validation set captions using our method and various language models. Both LLMs (GPT-2 (Radford et al., 2019), NEO (Black et al., 2021) , OPT (Zhang et al., 2022)) and VLMs (BLIP (Li et al., 2022b), GIT (Wang et al., 2022)) were tested. Our method computes log-likelihood values in a different range of values, with correlation values between 0.33 and 0.48 with all language models. See Fig. 14, App. C, for full details. In Fig. 7, we highlight three main differences between our likelihood score and those of language models:

1. **Caption length.** All language models approximate lower mean likelihood scores as caption length increases. In contrast, our likelihood score is less sensitive to caption length, Fig. 7.c. Levy et al. (2024) recently demonstrated a degradation in LLM performance on long inputs, particularly in reasoning tasks.

2. **Text type.** While text models are trained on general text, CLIP is trained specifically on captions of images. We sampled 5,000 sentences from OpenWebText (Gokaslan et al., 2019), a general text dataset, ensuring that their lengths are comparable to those of MS-COCO captions, and compared both likelihood histograms. For LLMs captions align with the general text distribution, whereas VLMs and our method result in separable distributions (Fig. 7.b).

3. **Grammatical errors.** Wu et al. (2023) demonstrates that ChatGPT performs poorly on datasets containing grammatical errors, particularly on long sentences. On

Table 3. **Cross dataset comparison.** *COCO*: MS-COCO, *F8k*: Flickr8k. *Data T*: dataset used for tests, *Data W*: dataset used to calculate the whitening transform. *Avg. AD, DP*: the average Anderson-Darling, D'Agostino-Pearson p-value test scores (threshold is under 0.752 and above 0.05, respectively). *Correlation* is calculated between likelihood scores on the test data using different whitening matrices.

		Image			Text		
Data T	**Data W**	**Avg. AD**	**Avg. DP**	**Correlation**	**Avg. AD**	**Avg. DP**	**Correlation**
COCO	COCO	0.489	0.362	0.69	0.592	0.257	0.74
	F8k	0.466	0.380		0.574	0.282	
F8k	COCO	0.641	0.317	0.77	0.735	0.226	0.88
	F8k	0.522	0.329		0.626	0.242	

Table 4. **Opposite image comparison.** Total Variation (TV), Entropy and percentage of saturation pixels (SAT[%]) for natural images (MS-COCO) and the opposite images generated using CLIP and W-CLIP embedding spaces. W-CLIP represents better the statistics of natural images.

Method	**TV**	**Entropy**	**Sat [%]**
MS-COCO	222.3	7.3	4.2
CLIP	156.7	4.8	55.5
W-CLIP	215.9	7.2	6.4

the other hand, we noticed our method is sensitive to grammatical errors and nonsensical inputs. To test this, we remove all the nouns from the MS-COCO captions and compare likelihood before and after. Our likelihood score is significantly affected, while the language models demonstrate less sensitivity, Fig. 7.a.

In Figs. 12, 13, App. C, histograms of all other language models are available. In Table 5, we quantify the separation between the likelihoods of different data types and captions with and without nouns. We employ the AUC metric, as defined in Eq. (10) (App. C), to evaluate the separation between distributions. Additional text examples are provided in Fig. 15, App. C. It is shown that the likelihood approximated using W-CLIP positively correlates with language model likelihoods but contains unique information derived from CLIP's learned distribution.

Table 5. **Likelihood separation with grammatical errors and different text types.** AUC values indicating the separation between likelihood distributions. *Type* compares the separation between captions and general text prompts, while *Nouns* compares the separation between original captions and the same captions with nouns removed. Vision-language models (VLMs) show a high separation for different text types and slightly higher separation when removing nouns compared to language models (LLMs). Our method yields the best separation, especially for *Nouns*.

	LLMs			VLMs		Ours
	GPT2	**OPT**	**NEO**	**BLIP**	**GIT**	
Type	0.8	0.8	0.77	0.92	0.97	**0.999**
Nouns	0.43	0.58	0.58	0.66	0.69	**0.94**

4.5. Data generalization

As W-CLIP is completely data-driven we test its generalization capabilities. Flickr8k (Hodosh et al., 2013), similarly to MS-COCO, is a benchmark for image-captioning tasks that emphasizes real-world imagery and descriptive diversity. In Tab. 3, we compare results using MS-COCO and Flickr8k as both the whitening and testing datasets. We evaluate the normal distribution test scores (Anderson-Darling and D'Agostino-Pearson) as in Sec. 3.3, and the correlation of likelihoods computed for the same data, using different datasets for whitening. The results show that whitening with one dataset and testing on another yields similar normal distribution test scores for features, and moderate to high correlations between likelihoods. Additional ablation studies, including different dataset size and utilizing a different CLIP model, are provided in App. E. These findings confirm that, although W-CLIP is data-driven, it generalizes well across datasets within the same domain. However, as shown in Fig. 3, W-CLIP is sensitive to domain shifts.

5. Conclusion

This paper introduces Whitened CLIP, transforming the raw CLIP latent space into an isotropic space. Whitened CLIP is statistically verified to approximate well normal distribution with independent and identically distributed (IID) components, and exhibits enhanced uniformity. The key contribution of this work is the proposal of a direct computation of likelihood functions for images and text prompts within the CLIP-learned distribution. Embeddings in the whitened space approximately follow the standard normal distribution, enabling the use of the squared Euclidean norm to estimate log-likelihood. These likelihood functions effectively identify artifacts, domain shifts, and demonstrate sensitivity to the complexity of details in text captions. Biases in generative models can be detected by comparing the likelihood of generated images to those of real images. Furthermore, the introduction of full-circle SLERP in the whitened space facilitates both interpolation and extrapolation between images. We believe the results of this research can further benefit numerous applications.

Acknowledgements

We would like to acknowledge support by the Israel Science Foundation (Grant 1472/23) and by the Ministry of Science and Technology (Grant No. 5074/22).

Impact Statement

This work advances Machine Learning by enhancing the understanding of a foundation model (CLIP) and utilizing it to derive direct likelihood functions for images and captions. Potential societal consequences of our work are related to downstream tasks that may rely on our findings. None of these consequences need to be specifically highlighted here.

References

Anderson, T. W. and Darling, D. A. A test of goodness of fit. *Journal of the American statistical association*, 49 (268):765–769, 1954.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata*, 58(2), 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cao, B., Yuan, J., Liu, Y., Li, J., Sun, S., Liu, J., and Zhao, B. Synartifact: Classifying and alleviating artifacts in synthetic images via vision-language model. *arXiv preprint arXiv:2402.18068*, 2024.

Chou, J. C.-C. and Alam, N. Embedding geometries of contrastive language-image pre-training. *arXiv preprint arXiv:2409.13079*, 2024.

D'agostino, R. and Pearson, E. S. Tests for departure from normality. empirical results for the distributions of b^2 and $\sqrt{b}$. *Biometrika*, 60(3):613–622, 1973.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., and Vedantam, S. R. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Du, Y. and Mordatch, I. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

Elharrouss, O., Almaadeed, N., Al-Maadeed, S., and Akbari, Y. Image inpainting: A review. *Neural Processing Letters*, 51:2007–2028, 2020.

Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., and Zhang, B. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10021–10030, 2023.

Geman, S. and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984.

Gokaslan, A., Cohen, V., Pavlick, E., and Tellex, S. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Goyal, B., Dogra, A., Agrawal, S., Sohi, B. S., and Sharma, A. Image denoising review: From classical to state-of-the-art approaches. *Information fusion*, 55:220–244, 2020.

Groeneveld, R. A. and Meeden, G. Measuring skewness and kurtosis. *Journal of the Royal Statistical Society Series D: The Statistician*, 33(4):391–399, 1984.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hodosh, M., Young, P., and Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013.

Kan, W., Howard, A., and Park, E. Imagenet object localization challenge. In *https://kaggle.com/competitions/imagenet-objectlocalization-challenge*, 2018.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007–6017, 2023.

Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Levi, M. Y. and Gilboa, G. The double-ellipsoid geometry of clip. In *International Conference on Machine Learning*. PMLR, 2025.

Levy, M., Jacoby, A., and Goldberg, Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 2024.

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., and Chen, Y. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022a.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022b.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Mardia, K. and Jupp, P. *Directional Statistics*. Wiley, 2000.

Mokady, R., Hertz, A., and Bermano, A. H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Ou, Z. et al. Energy-based models with applications to speech and language processing. *Foundations and Trends® in Signal Processing*, 18(1-2):1–199, 2024.

Paouris, G. Concentration of mass on convex bodies. *Geometric & Functional Analysis GAFA*, 16(5):1021–1049, 2006.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 1 (8), 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ruderman, D. and Bialek, W. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

Schrodi, S., Hoffmann, D. T., Argus, M., Fischer, V., and Brox, T. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. *arXiv preprint arXiv:2404.07983*, 2024.

Shi, P., Welle, M. C., Björkman, M., and Kragic, D. Towards understanding the modality gap in clip. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Wu, H., Wang, W., Wan, Y., Jiao, W., and Lyu, M. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. arxiv, 2023.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zoran, D. and Weiss, Y. From learning models of natural image patches to whole image restoration. In *2011 international conference on computer vision*, pp. 479–486. IEEE, 2011.

A. Reproducibility

Our code along detailed instructions is available HERE. The repository includes: 1) Implementation of our method, reproducing most of our experiments, including simple demo notebooks; 2) Whitening matrices; 3) Additional examples beyond those in the paper and in the appendix, specifically video demonstrations of the full circle SLERP.

B. Artifacts and Domain Shifts Examples

B.1. Image artifacts

In Fig. 9 we offer additional examples of real images compared to similar generated images with artifacts, as presented in Fig. 2 in Sec. 4.2.

B.2. Text artifacts

Trying to generate artifacts in text captions we remove the first or last words from a caption, or one of the middle words. Examples in Fig. 10. In all cases the original caption has the highest log-likelihood score.

B.3. ImageNet datasets

In Fig. 11 we provide histograms using different corruptions from ImageNet-C. All corruptions have a lower log-likelihood compared to ImageNet.

Figure 9. **Log-likelihood of real and generated images with artifacts.** Real images of zebras and a surfer (left) and three similar AI generated images with artifacts. Real images have higher log-likelihoods than AI generated images with artifacts.

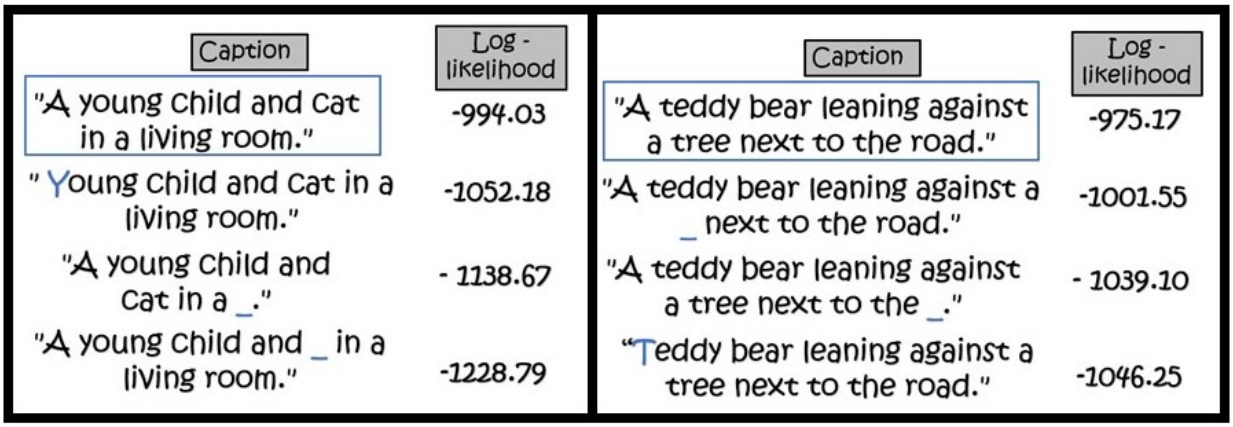


Figure 10. **Log-likelihood of real captions and captions with artifacts.** Real captions, framed with a blue frame and atrifacted captions, where we removed the first or last words from a caption, or one of the middle words. In all cases the original caption has the highest log-likelihood score.

C. Comparison with Language Models Examples

In Fig. 14, we present the log-likelihood values for the MS-COCO validation set using language models, along with the correlation between these log-likelihoods and those computed using our method. Our approach yields log-likelihood scores with larger absolute values and greater variance, aligning with its intended design.

Figs. 12, 13 display histograms replicating the experiments from Figs. 7.a,b for additional models. The LLMs (OPT, NEO) exhibit behavior very similar to GPT-2. The VLMs (BLIP, GIT) also show behavior similar to GPT-2, but with some deviations trending toward our method's likelihood. This observation is reasonable, as these models, like CLIP, are trained (or fine-tuned) on caption data rather than general text data. In Fig. 15, we present additional examples of captions with varying relative likelihood scores. We sort the likelihood scores of 5,000 captions from MS-COCO in ascending order and examine the sorted index for different captions. The comparison includes our method, an LLM (GPT2), and a VLM (BLIP). Each set of examples demonstrates one of the three differences discussed in Sec. 4.4 (e.g. text type, grammatical errors and caption length).

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC, AUC for simplicity) evaluates the ability of a model to distinguish between two classes. It measures the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various threshold levels. The AUC score in Tab. 5 is mathematically defined as:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx \quad , \quad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad , \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{10}$$

where TP are true positives, FP are false positives, TN are true negatives and FN are false negatives. In the context of Tab. 5 the MS-COCO captions are defined as positives and the general text or captions without nouns are defined as negatives.

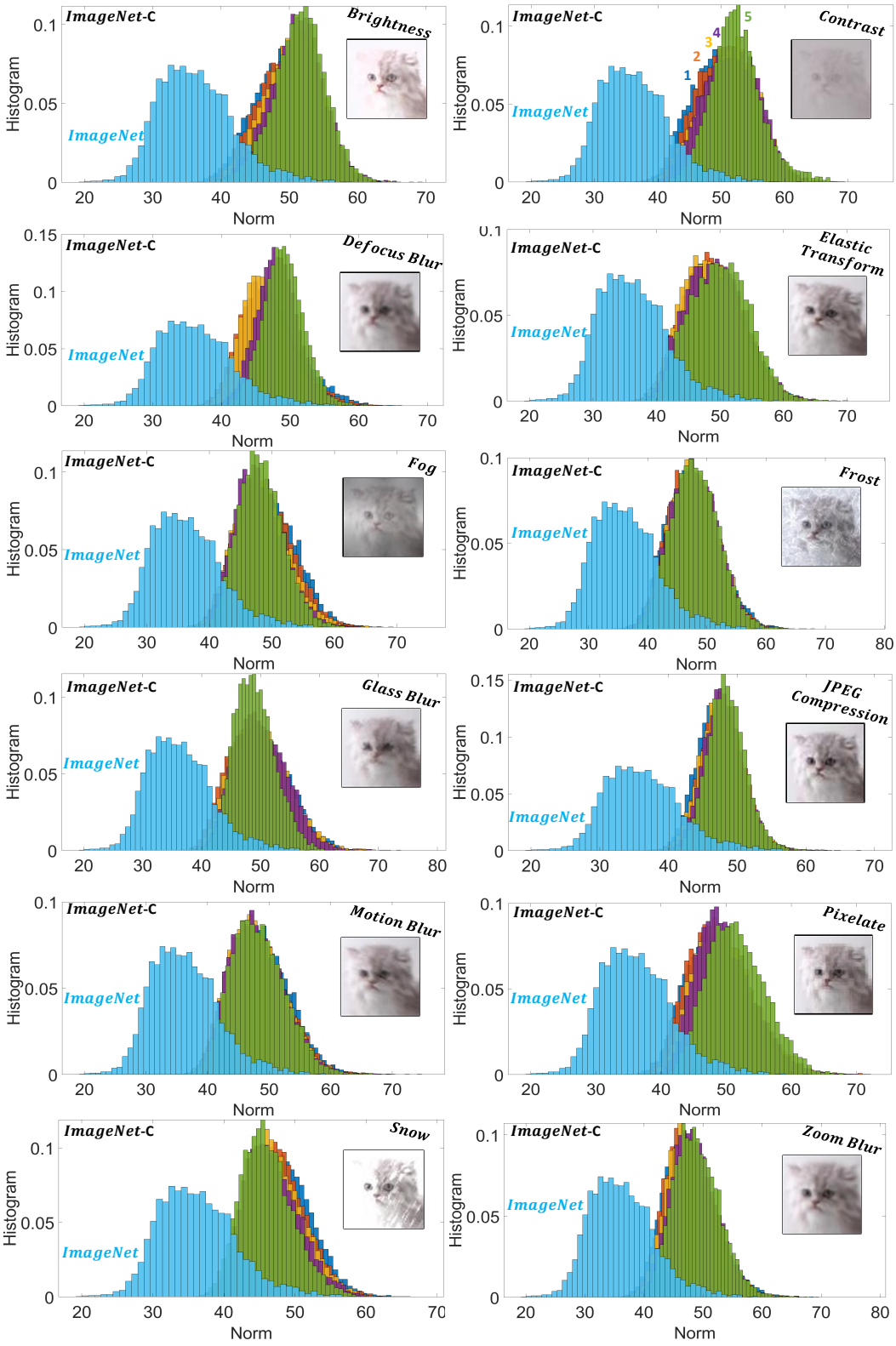


Figure 11. **ImageNet-C histograms on all corruptions.** All corruptions have a significantly higher norm (lower log-likelihood) than ImageNet. For most corruptions, as level of corruption increases the norm increases. Some corruptions do not show this monotonic behavior (motion/glass blur for example) for different levels of corruption.

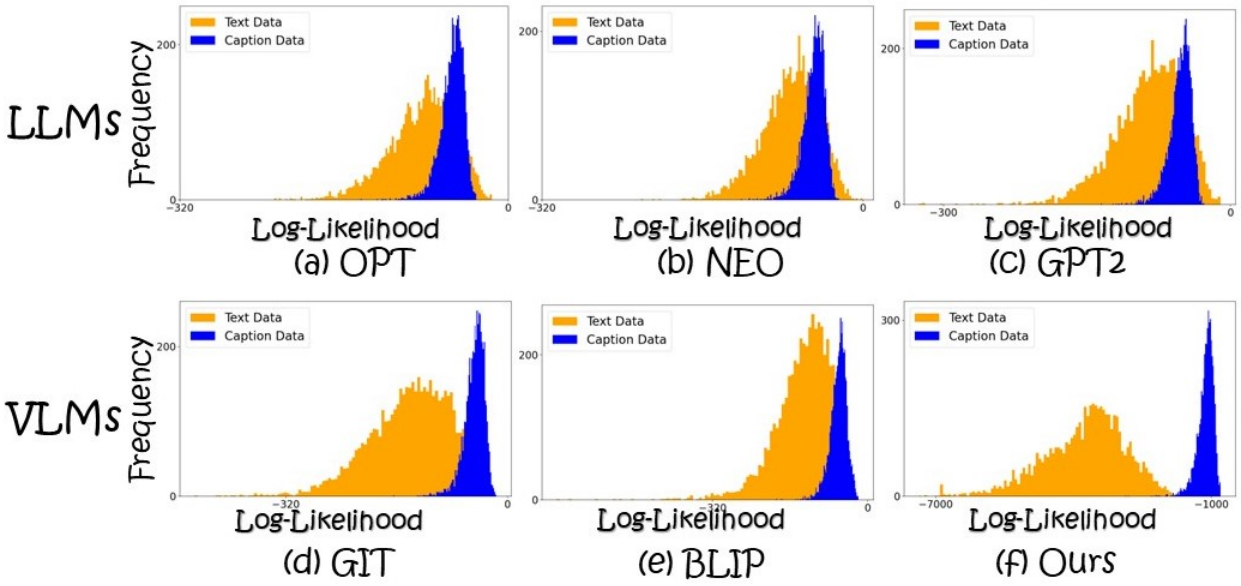


Figure 12. **Likelihood for different text types.** Comparing likelihood values computed for MS-COCO captions and OpenWebText general text sentences. The sentences from OpenWebText are filtered to have similar lengths to MS-COCO captions. LLMs (OPT, NEO, GPT2) treat captions similarly to general text while VLMs (BLIP, GIT) show some separation. No model shows strong separation like our likelihood does (Tab. 5).

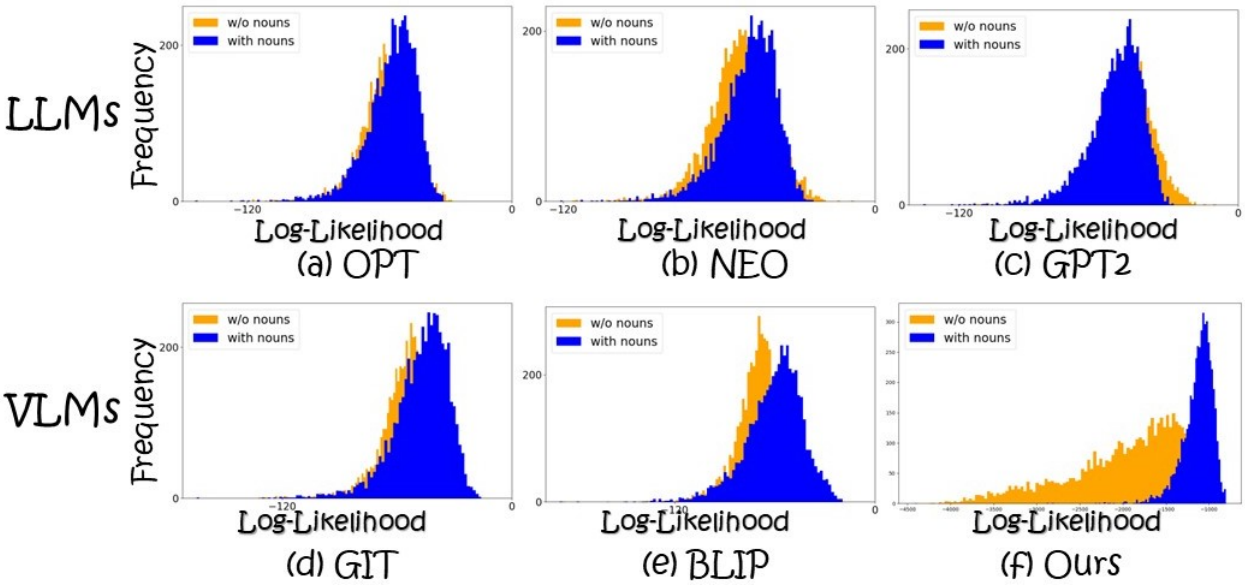
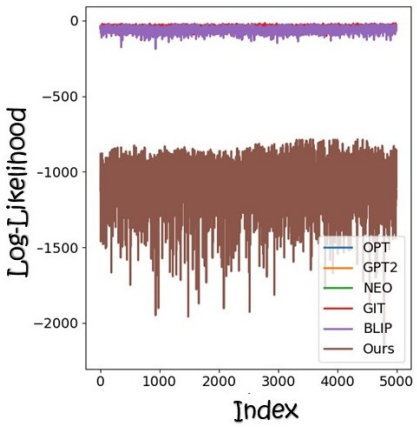


Figure 13. **Likelihood drift when removing nouns** Comparing likelihood values computed for MS-COCO captions with and without nouns. None of the models show a drift like our likelihood (Tab. 5).



Model	Correlation
GPT2	0.33
OPT	0.35
NEO	0.37
GIT	0.42
BLIP	0.48

Figure 14. **Log-likelihood values and correlations.** Log-likelihoods are computed on 5,000 captions from the MS-COCO validation set, with correlations measuring the alignment of each model's log-likelihood and ours.

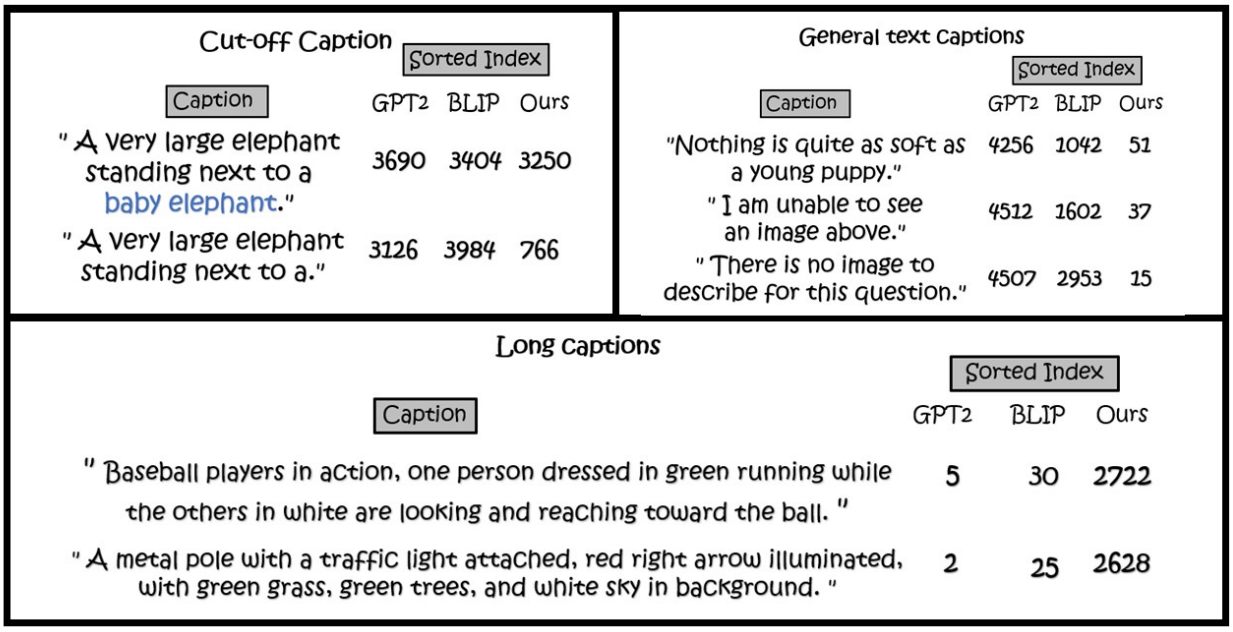


Figure 15. **Examples of differences between language models and our method.** The sorted index represents the position out of 5000 captions from MS-COCO, ranked from low to high likelihood values. The relative likelihood index is compared among GPT2 (LLM), BLIP (VLM), and our method. Top left: Our method shows a significant drop in relative likelihood compared to the language models when the sentence is cut-off. Top right: Captions that do not describe images receive the lowest relative likelihood from our method, the highest from GPT2, and intermediate scores from BLIP. Bottom: Long captions are assigned low relative likelihoods by language models, while our method assigns them average relative likelihood scores.

D. Implementation and Theoretical Details

D.1. Implementation details

As explained in Sec. 3.2 if x is a random vector in R^d with a non-singular covariance matrix Σ (and with zero mean), then W satisfying $W^TW = \Sigma^{-1}$ is called the whitening matrix. One common approach to achieve whitening is through

Principal Component Analysis (PCA), although other methods like Zero-Phase Component Analysis (ZCA) whitening and Singular Value Decomposition (SVD) whitening exist. PCA whitening transforms the data into a new coordinate system defined by the principal components of the covariance matrix. It rescales each component to have unit variance, effectively "whitening" the data. Steps of PCA whitening:

- Compute the covariance matrix of the data.
- Perform eigenvalue decomposition to obtain eigenvalues (Λ) and eigenvectors (V).
- Transform the data:

$$\mathbf{X}_{\text{whitened}} = \Lambda^{-1/2} V^\top \mathbf{X}. \tag{11}$$

The main advantages of PCA whitening are that it ensures that the resulting features are uncorrelated and transforms data along principal axes, which often correspond to meaningful directions in the dataset. It can be efficient for dimensionality reduction, something we do not use in our work. The main limitation is the loss of original geometry (ZCA whitening for instance maintains the original geometry).

When the features in the original data are highly correlated, the matrix W becomes unstable and may not be invertible. To address this issue, we remove one of the highly correlated features and replace it with random noise. In our experiments, this situation occurs only when whitening text embeddings and not with image embeddings. While this introduces some randomness into the process, it has minimal impact on the empirical results. Our full whitening code, together with scripts repeating our experiments is available here.

Algorithm 1 Whitening Process

Input: Dataset $\mathbf{X} \in \mathbb{R}^{N \times d}$, correlation threshold τ
Output: Whitening matrix $\mathbf{W}$

Step 1: Compute Correlation Matrix.
Calculate the correlation matrix:

$$\mathbf{C}_{ij} = \frac{\text{Cov}(\mathbf{X}_i, \mathbf{X}_j)}{\sigma_i \sigma_j}.$$

Step 2: Remove Highly Correlated Features.
Identify feature pairs (i, j) where $|\mathbf{C}_{ij}| > \tau$.
For each pair, remove one feature (e.g., j) and replace it with random noise $\mathbf{r}$, Denote the updated dataset as $\mathbf{X}'$:

$$\mathbf{r} \sim \mathcal{N}(0, 0.1).$$

Step 3: Compute Covariance Matrix.
Calculate the covariance matrix:

$$\Sigma = \frac{1}{N}(\mathbf{X}'^\top \mathbf{X}').$$

Step 4: Perform Eigenvalue Decomposition.
Decompose Σ into eigenvalues Λ and eigenvectors V:

$$\Sigma = V \Lambda V^\top.$$

Step 5: Compute Whitening Matrix and Transform Data.
Calculate the whitening matrix:

$$W = \Lambda^{-1/2} V^\top,$$

where $\Lambda^{-1/2}$ is a diagonal matrix with elements given by the inverse square root of the eigenvalues:

$$\frac{1}{\sqrt{\lambda_i}}.$$

D.2. Isotropic random vectors

An isotropic random vector is one where all components are identically distributed and statistically independent, with zero mean and unit variance, e.g it's covariance matrix $\mathbf{\Sigma}$ is the unit matrix. In other words, an isotropic vector is uniformly distributed across the space, exhibiting no directional bias. Such vectors often arise in high-dimensional statistical models and machine learning applications, where isotropy ensures that the data's statistical properties are invariant to rotation or translation (Mardia & Jupp, 2000). Isotropic distributions are particularly relevant in contexts such as embedding spaces, where uniformity and independence across features simplify analysis and facilitate probabilistic modeling (Vershynin, 2018).

D.3. Normal distribution tests

Normality tests assess whether a dataset follows a Normal distribution, a critical assumption in many statistical methods. As discussed above in Sec. 3.3 the Anderson-Darling test evaluates how well the empirical cumulative distribution function (CDF) matches the expected CDF of a normal distribution, placing higher weight on the tails to detect deviations (Anderson & Darling, 1954). The D'Agostino-Pearson test combines skewness and shape characteristics measures to assess normality, offering sensitivity to both symmetric and asymmetric deviations (D'agostino & Pearson, 1973). The Anderson-Darling test statistic is defined as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\left(\ln F(y_i) + \ln\left(1 - F(y_{n+1-i})\right)\right)\right] \tag{12}$$

Where n is the sample size, $y_1 \leq y_2 \leq \cdots \leq y_n$ are the ordered data samples and $F(y)$ is the Cumulative Distribution Function (CDF) of the hypothesized distribution. The D'Agostino-Pearson test statistic combines skewness and kurtosis:

$$K^2 = z_1^2 + z_2^2 \quad , \quad z_1 = \frac{g_1}{\sqrt{\frac{6}{n}}} \quad , \quad z_2 = \frac{g_2 - 3}{\sqrt{\frac{24}{n}}} \tag{13}$$

where K^2 is the D'Agostino-Pearson test statistic, z_1, z_2 are the standardized skewness and kurtosis. g_1, g_2 are the sample skewness and kurtosis and n is the sample size. For details regarding skewness and kurtosis please refer to Groeneveld & Meeden (1984).

These tests are well-suited for high-dimensional data as they are robust to various types of distributional departures, making them effective for validating Normal approximations in the context of our proposed whitened embedding spaces. Bellow we present statistics of image (Fig. 16) and text (Fig. 17) embeddings, on both tests. Mean values (on all groups of data) with standard deviation and histograms of mean values are presented. In addition, in Fig. 18 we present the mean and variance of all the whitened features, demonstrating minor deviation from the expected values (zero mean and unit variance).

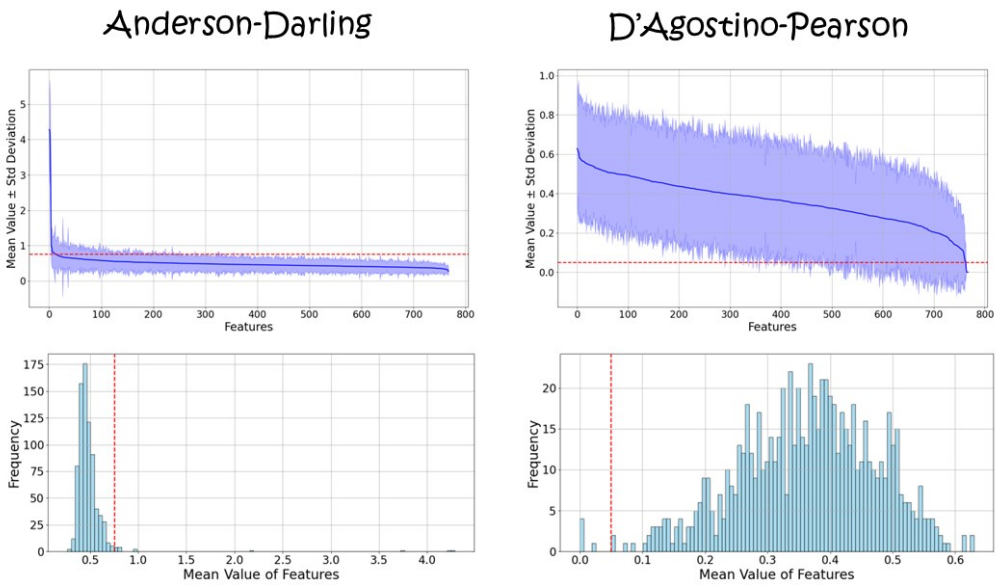


Figure 16. **Normal distribution tests on image embeddings.** Top row - mean value and standard deviation per feature, over all groups of embeddings. Bottom - histogram of mean values of each feature. In all plots the red line represents the test threshold. Left - Anderson-Darling test, threshold is 0.752, lower is better. Right - D'Agostino-Pearson test, threshold is 0.05, higher is better.

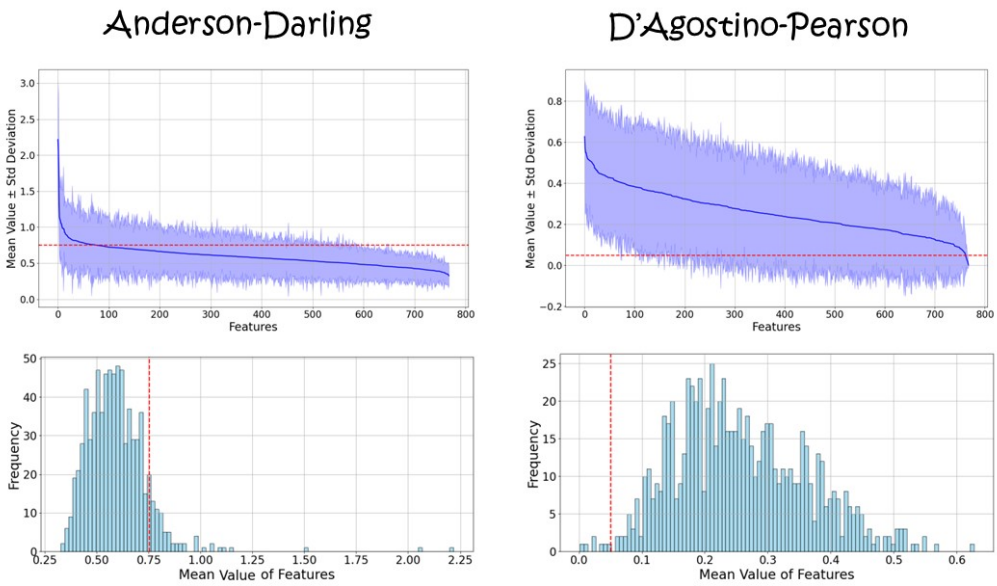


Figure 17. **Normal distribution tests on text embeddings.** Top row - mean value and standard deviation per feature, over all groups of embeddings. Bottom - histogram of mean values of each feature. In all plots the red line represents the test threshold. Left - Anderson-Darling test, threshold is 0.752, lower is better. Right - D'Agostino-Pearson test, threshold is 0.05, higher is better.

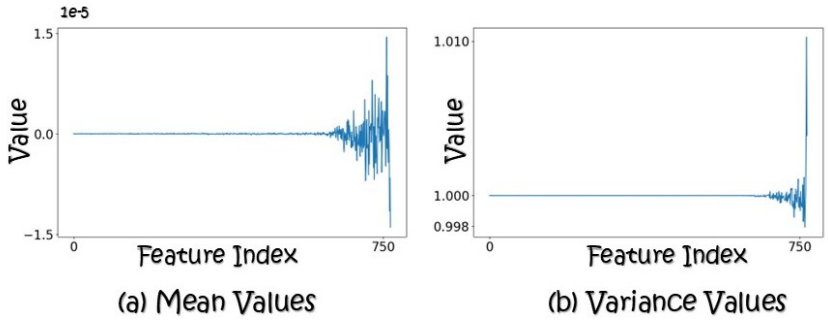


Figure 18. **Mean and variance of all whitened features.** We show the mean and variance of all the 768 features of the whitened embeddings. There are minor deviations from 0 (for mean) and 1 (for variance). For mean values the deviation is up to 0.0015% and in the case of the variance the deviations are up to 1%.

D.4. Thin shell theory

The thin shell theory (Paouris, 2006) is a concept in high-dimensional geometry. According to the thin shell theory most of the volume of a high dimensional convex space is concentered near the surface of the space. Specifically, for a convex space $K \subseteq \mathbb{R}^d$, the majority of points in K lie at an approximated distance r from the origin. This can be formally described in terms of the concentration of measure, where the typical distance of a random point from the origin is concentrated around a specific radius:

$$\mathbb{P}(\|x\| \in [r-\epsilon, r+\epsilon]) \approx \frac{C(d)}{\epsilon^d} \tag{14}$$

where x is a random sample from the space K, r is the typical radius of the space, ϵ is a small deviation ($\epsilon \ll 1$) and C is a constant that depends on the dimensions d. This result indicates that as the dimension d grows, the concentration near the thin shell becomes sharper. The thin shell phenomenon is closely related to the chi distribution described above (Eq. (7)). Specifically relating Eq. (8) to Eq. (14) we get $r = \sqrt{d - \frac{1}{2}}$. Combining both phenomena, as d increases, the space expands and concentration near the surface emerges because the majority of the space's mass resides near its boundary. Consequently, most points are located near the surface, even as the overall space grows.

E. Ablation Study with Different Data and CLIP Model

We apply the whitening transform to embeddings of MS-COCO validation set using a second CLIP model - CLIP ViT-B/32, which encodes embeddings with 512 features (compared to embeddings with 768 features encoded by CLIP ViT-L/14). Results are in Tab. 6. Results are very similar to CLIP ViT-L-14, used in the paper, verifying that our method is general for different CLIP models.

Table 6. **Normal distribution tests using a different CLIP model** *Avg. AD, DP* - the average Anderson-Darling, D'Agostino-Pearson p-value test scores (threshold is under 0.752 and above 0.05, respectively). MS-COCO validation set tested using CLIP ViT-B/32, which has 512 features in each embedding.

	Avg. AD	**Avg. DP**
Image	0.65	0.31
Text	0.61	0.25

We conduct an ablation study, examining the influence of the size of the data used for computing the whitening matrix W. For each size (1k, 2k, 3k, 4k) we randomly sampled 5 subsets of MS-COCO validation set. The average scores with standard deviation are plotted in Fig. 19. The tests are performed on the full MS-COCO validation set (5k images). For the D'Agostino-Pearson test, for all data sizes the tested embeddings comply with the normal distribution. For the Anderson-Darling test using 1k samples for computing W results with tested embeddings that do not comply with the

normal distribution. Using at least 2k results with tested embeddings that comply with the normal distribution. For both tests, using more data to compute W results with improved results. As the whitening transform is completely data-driven it is expected that using additional data improves the results. However we note that the improvement between using 4k and 5k samples is small.

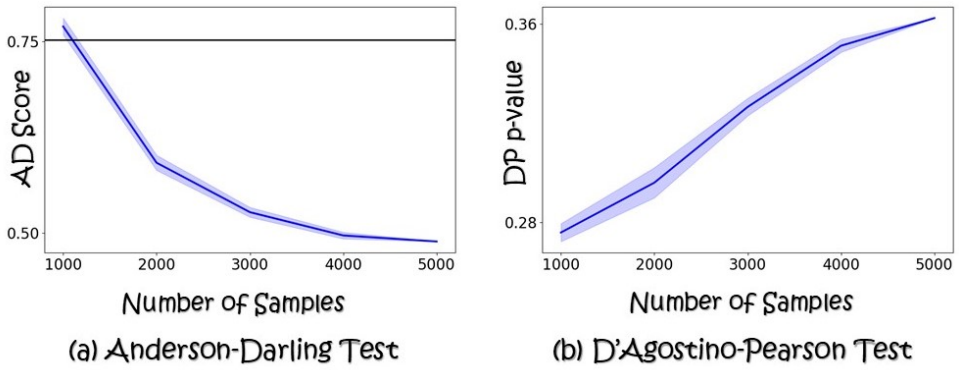


Figure 19. **Normal distribution tests with different data sizes** Anderson-Darling average scores and standard deviation (a) and D'Agostino-Pearson p-value average scores and standard deviation (b). Threshold is under 0.752 (marked with a black line) and above 0.05, respectively. Computing the whitening matrix W with different data sizes. For each size (1k, 2k, 3k, 4k) we randomly sampled 5 subsets of MS-COCO validation set and present the average score with standard deviation. The tests are performed on the full MS-COCO validation set (5k images).

F. Full circle SLERP Examples

In Fig. 20 a simple 2D scenario of full circle SLERP is demonstrated. The main observation is that if the source and destination points are on a circle around the origin (allowing small deviations) the full circle SLERP points (blue) remain on (or near) the original circle (orange). However, if the circle is skewed from the origin the SLERP points deviate far from the original circle.

We present an additional example of sets of images from a full circle SLERP, as discussed in Sec. 4.3, in Fig. 21. As in Fig. 8, also in this case it is clear that a full circle SLERP is not practical in the raw CLIP space, while resulting with real images throughout the full circle in the W-CLIP space.

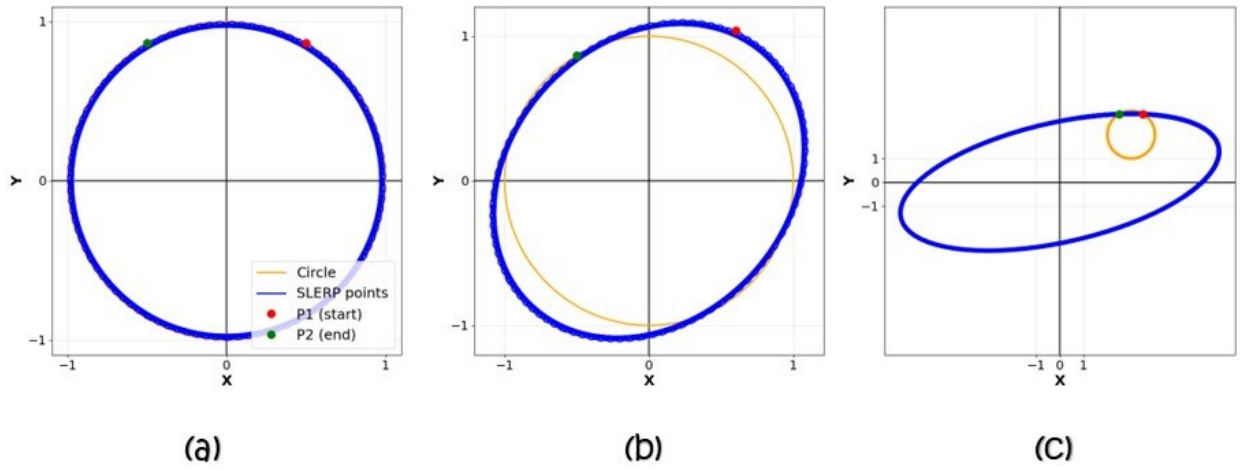


Figure 20. **2D full circle SLERP example.** The SLERP points are in blue and the circle perimeter is in orange. Examples of a simple 2D case of full circle SLERP. When both points are on the circle (a) the SLERP points follow the circle perimeter perfectly. If one of the points deviates from the circle (b) the SLERP points form an ellipse, but remain close to the circle perimeter. If the circle is skewed from the origin (c) the SLERP points form a large ellipse, that distances far from the circle perimeter.

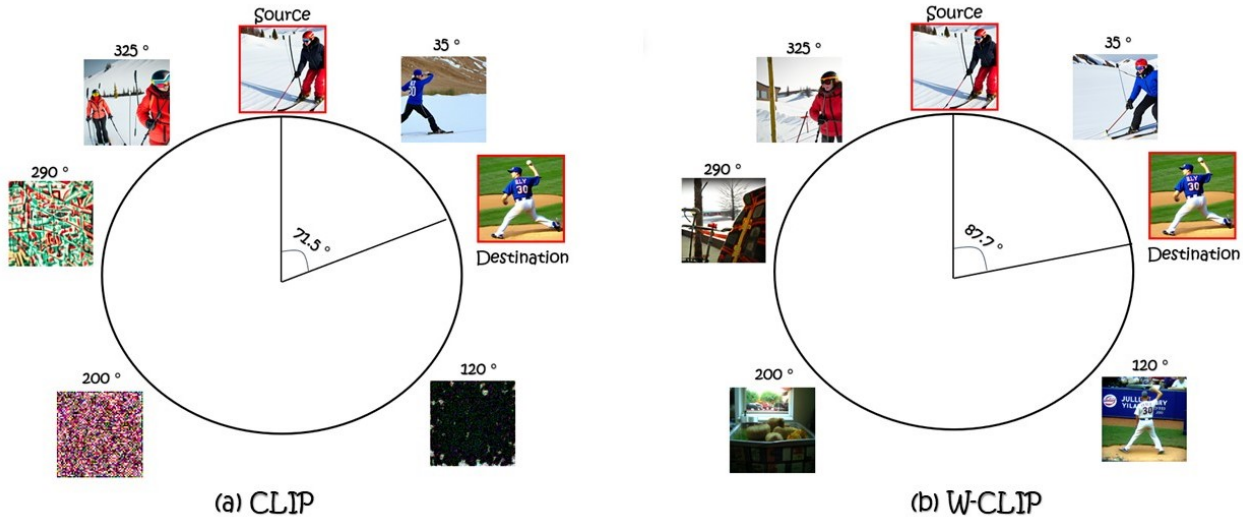


Figure 21. **Full circle SLERP example.** The full circle SLERP is performed in both the raw CLIP space (a) and in the W-CLIP space (b). The different angle between embeddings in both space is presented. In the raw CLIP space the full circle SLERP results with noise for most of the degrees not between the source and destination embeddings. In the W-CLIP space for all degrees real images are generated.

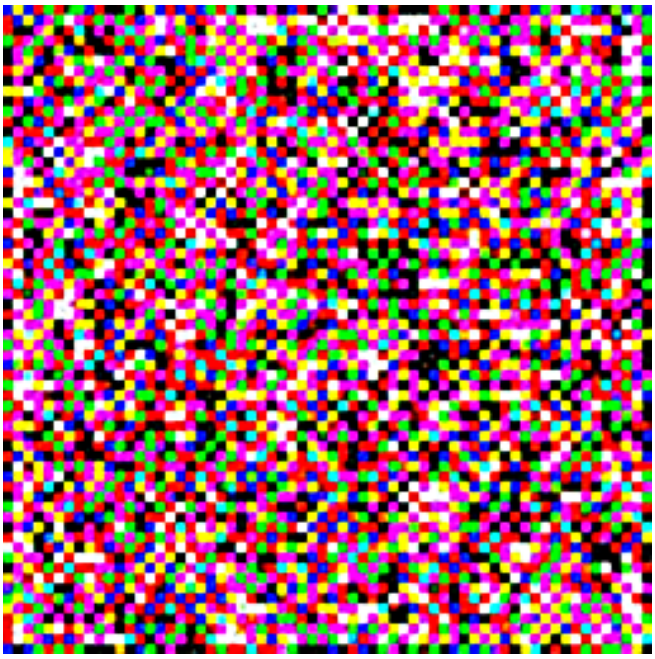


Figure 22. **Opposite image generated in the raw CLIP space.** The structured noise produced by CLIP exhibits 4×4 pixel blocks and a restricted color palette (black ('0' in all color channels), white ('1' in all color channels), red, green, blue, magenta ('1' in red and blue channels), cyan ('1' in green and blue channels), and yellow ('1' in red and green channels)), suggesting synthetic artifacts.

G. Image Generation Bias Examples

In Figs. 23, 24 additional examples of the bias in image generation models are presented. Different random seeds lead to different results due to the image generation model. In all cases the images become noise when no normalization is applied in the whitened space. Normalizing embeddings in the whitened space to have a norm of $\sqrt{d}$ in each iteration results with reasonable images, with varying content.

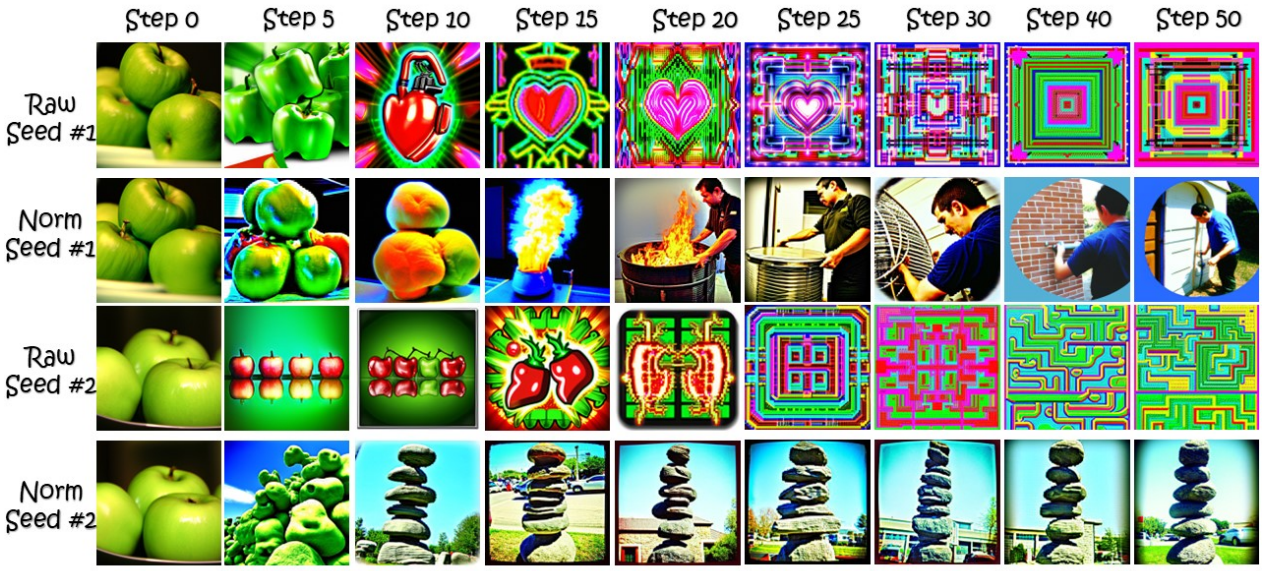


Figure 23. **Generation bias.** Iteratively using UnCLIP to generate images encoded by CLIP with two fixed seeds. The raw process gradually becomes noisy, whereas with normalization (to $\sqrt{d}$ at each encoding step), the content drifts but remains within a natural and reasonable image space.

Figure 24. **Generation bias.** Iteratively using UnCLIP to generate images encoded by CLIP with two fixed seeds. The raw process gradually becomes noisy, whereas with normalization (to $\sqrt{d}$ at each encoding step), the content drifts but remains within a natural and reasonable image space.

H. Text Complexity Examples

In Figs. 25, 26 we repeat the experiment presented in Fig. 5, showing additional examples how adding and removing details from concepts (not the concepts themselves) decreases/increases the likelihood respectively.

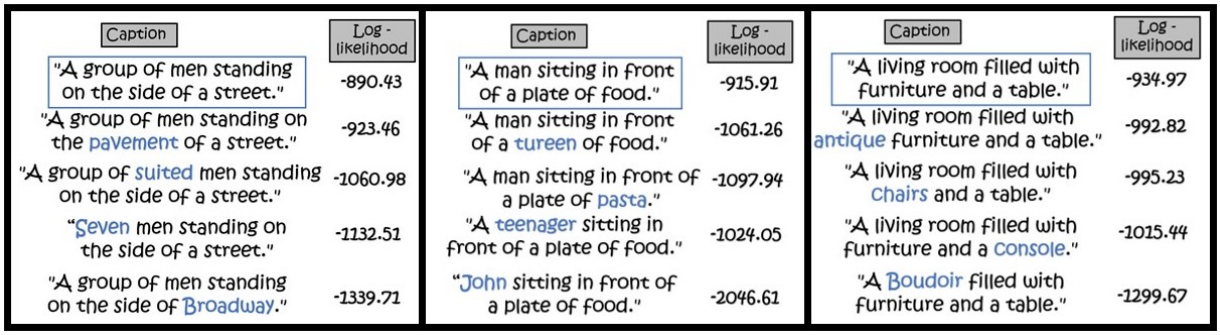


Figure 25. **Adding details to concepts.** The original caption from MS-COCO is framed in blue. Adding details decreases the likelihood.

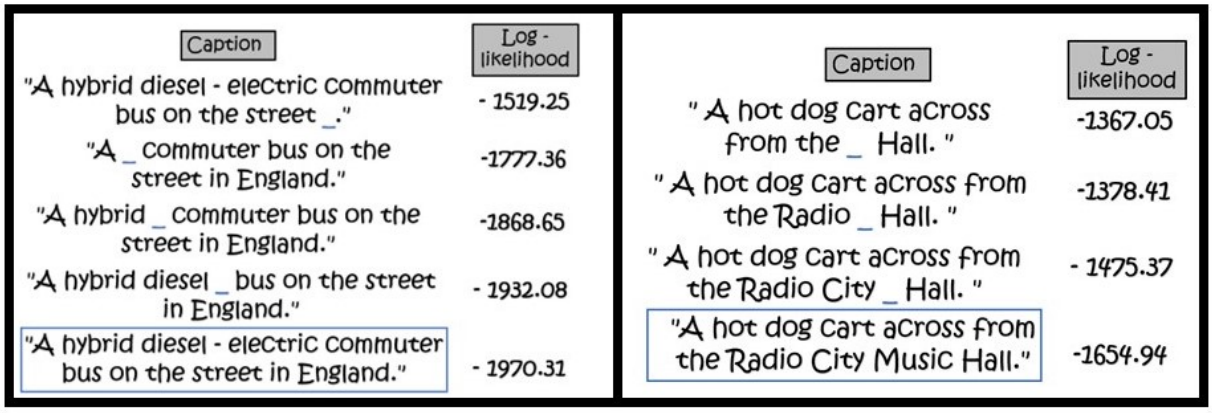


Figure 26. **Removing details from concepts.** The original caption from MS-COCO is framed with a blue frame. Removing different details increases the likelihood.