Interpretable Measures of Conceptual Similarity by Complexity-Constrained Descriptive Auto-Encoding

Alessandro Achille*   Greg Ver Steeg*   Tian Yu Liu   Matthew Trager
Carson Klingenberg   Stefano Soatto
AWS AI Labs

Abstract

Quantifying the degree of similarity between images is a key copyright issue for image-based machine learning. In legal doctrine however, determining the degree of similarity between works requires subjective analysis, and fact-finders (judges and juries) can demonstrate considerable variability in these subjective judgement calls. Images that are structurally similar can be deemed dissimilar, whereas images of completely different scenes can be deemed similar enough to support a claim of copying. We seek to define and compute a notion of 'conceptual similarity' among images that captures high-level relations even among images that do not share repeated elements or visually similar components. The idea is to use a base multi-modal model to generate 'explanations' (captions) of visual data at increasing levels of complexity. Then, similarity can be measured by the length of the caption needed to discriminate between the two images: Two highly dissimilar images can be discriminated early in their description, whereas conceptually dissimilar ones will need more detail to be distinguished. We operationalize this definition and show that it correlates with subjective (averaged human evaluation) assessment, and beats existing baselines on both image-to-image and text-to-text similarity benchmarks. Beyond just providing a number, our method also offers interpretability by pointing to the specific level of granularity of the description where the source data are differentiated.

1. Introduction

Consider the two images in Fig. 1. One could say they are similar: both portray small red Italian cars. Another could say they are different: One is a sports car in an open space, the other a tiny city car in an alley. Is there an objective way of measuring the similarity among images? In some cases, similarity judgments can be influenced by shared concepts: in Fig. 2, two images share compelling stylistic and conceptual similarity, but it is difficult to identify specific visual elements they share. Yet the two were found to be legally "substantially similar" [37]. Can we define an objective notion of 'conceptual similarity' among data?

There have been many attempts at defining an objective notion of similarity based on the number of bits of information that the two samples share [11, 25, 40], but fundamentally they do not capture concepts (Appendix A), which are human constructs. Since humans must play a role in defining similarity among concepts, one way to achieve objectivity is by averaging a large number of human assessments. This is what large-scale neural network models do [34]. However, contrastive-trained models measure similarity based on how easily the data can be distinguished, and any two non-identical samples can be distinguished by random features of no conceptual relevance [22].

Rather than focusing on finding the features that best distinguish the two images, we focus on finding the ones that best *describe* them. Then, we can measure semantic similarity based on how well descriptions of one fit the other. If two samples are similar, a short description should apply equally well to both. If the samples are very different, the description of one will be a poor fit for the other. The more complex the description needed to distinguish the images, the higher their conceptual similarity. The key idea is to focus on optimally describing individual samples, rather than adversarially discriminating them.

Specifically, referring to Fig. 1, we generate multiple descriptions of each sample in increasing level of complexity, sorted by their coding length. Then, we measure the difference of the likelihood of each image conditioned on either descriptions as a function of complexity. That traces a curve that measures distance as a function of complexity. Any two images can be distinguished by their description *at some point* (the more you look, the more *differences* you see). Therefore, similarity should always be relative to a description length. However, when a single number is needed for comparison, we show that the AUC of the distance function is well aligned with human similarity assessments.

Our proposed method, which we call Complexity-Constrained Descriptive Autoencoding, or CC:DAE, is rooted in the idea of the Kolmogorov Structure Function,
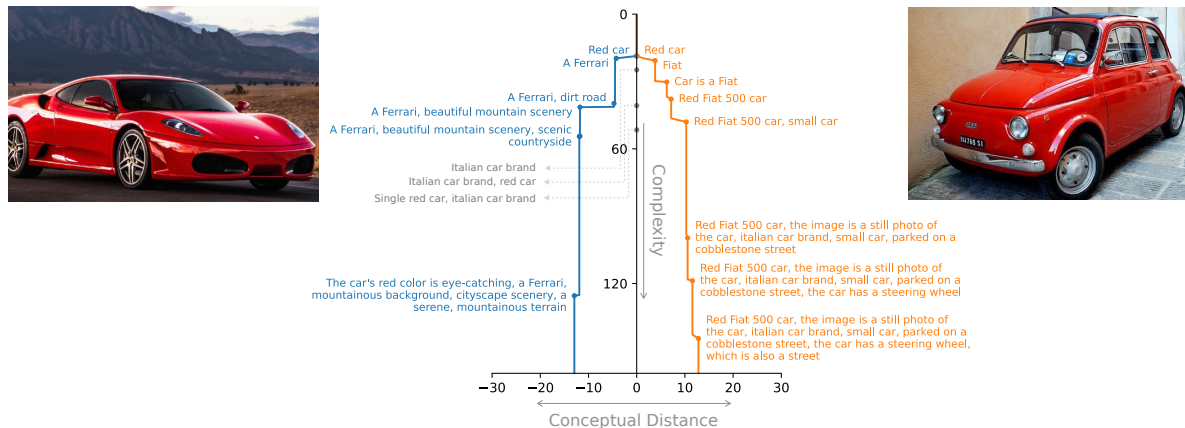
Figure 1. If we describe each image at increasing levels of complexity (blue and orange text), short descriptions apply equally well to both, as measured by their likelihood. However, as the complexity level of the description increases, a gap emerges between the likelihood under the best common description (grey) and the likelihood under the best individual descriptions (blue and orange). For instance, at $C = 36$ the best individual descriptions are "Red Fiat 500 car" and "A Ferrari" whereas the best common description is "Italian car brand" which is not as descriptive. The gap traces two asymmetric curves that measure the conceptual difference between the images at each level of complexity. A single number can be obtained by measuring the area under between curves.

which was introduced to differentiate semantic 'information' from structureless 'noise.' But while Kolmogorov used *programs* to describe the data, and program length as their complexity, we use natural language descriptions and their coding length. One may wonder whether our notion of conceptual distance is canonical in any sense, since it hinges on arbitrary choices, among which the use of language, or even a particular trained language model. Indeed, as we prove formally, no notion of conceptual similarity can be canonical: Using results from [13], we show that Kolmogorov's choice itself does not convey any information about the semantics of data and more generally that no nontrivial notion of common conceptual information can be defined without imposing strong restrictions on the class of representations and encoder/decoder used. The unavoidable need to perform subjective analysis is recognized by copyright legal doctrines, which provide generalized guidelines for analyzing similarity but leave the ultimate determination to be made on a case-by-case basis. As a result, despite efforts to codify similarity into guidelines, one is left with the impression that there are as many notions of similarity as there are judges or juries. One of the advantages of large-scale pre-trained models is that they aggregate content from vast and diverse corpora and, unlike their human sources, their biases can be measured, monitored, and calibrated.

Our method applies to general data, including any combination of text and image and any choice of data-to-text encoder (*e.g.,* captioning models) and a text-to-data generative model (*e.g.,* multi-modal autoregressive models or diffusion models). It can also be used to measure similarity of data across different modalities (*e.g.,* between text and images).

To evaluate the alignment between our notion of concep-

tual similarity and human assessments, we use established human-annotated similarity benchmarks, and obtain state-of-the-art results on the STS sentence similarity benchmarks beating all methods that have not explicitly been trained on human annotated similarity scores. We also surpass CLIP on the CxC-SIS image similarity benchmark.

2. Related Work

Using machine learning to represent conceptual information is a long-running yet elusive goal in machine learning. For instance, the "concept bottleneck" [23] epitomizes a line of work that generate restricted representations based on human specified concepts. Our approach is more general: we do not specify specific concepts, but only demand natural language descriptions with constrained complexity. Lossy compression, for example through capacity-contrained variational auto-encoders [1, 8, 18], can be used to distinguish semantic information in images [16], but the success of these methods is largely due to inductive priors coming from the architecture. In general, there is no reason to expect compression alone to align with human intuition of conceptual information, as we demonstrate in Sec. 4. Another line of work referring to 'visual concept learning' defines concepts in terms of a set of generalizable properties that then can be used in conjunction with 'neuro-symbolic' programs to answer novel queries [30], or in terms of embeddings that can be used to generate images [27].

Recent work on contrastive vision-language models [34] has demonstrated that these models are surprisingly weak at understanding semantic relationships [10, 43], with only modest improvements from conditional diffusion mod-

els [24]. Generating natural language descriptions of images as representations was also considered by [7], who finds the resulting description valuable for a visual question answering task and more useful than the original images for a retrieval task.

Embedding generated by auto-regressive models fine-tuned on human annotations are also frequently used to quantify similarity [12, 15, 20, 32, 35, 44]. Fewer works define similarity using pre-trained auto-regressive models without supervised fine-tuning: [21] uses prompts to condition model outputs; [31] evaluates similarity of two samples via the conditional probability of one given the other. Most similar to our work, [28] defines semantic similarity as the divergence between conditional distributions over future trajectories. Interestingly, [28] can be considered a variant of a special case of our method (Appendix D), which however further benefits from improved interpretability.

3. CC:DAE

In this section we formalize the critical components of our method, including the optimal description of the data using complexity-constrained textual expressions, and the computation of the fit of the description, from which the conceptual distance is measured.

3.1. Optimal Description

Let $x \in X$ be a sample and let $H = \{h_1, h_2, \ldots\}$ be space of hypotheses. We think of each $h \in H$ as a possible description of the sample, for example a natural language sentence. Associated with H and X we consider a *decoder* $p(x|h)$ which computes the conditional likelihood of the input x given the description h. We refer to the negative conditional log-likelihood $\ell(x|h) = -\log p(x|h)$ as the reconstruction loss of x under hypothesis h. A key quantity for us is the *complexity*, or *coding length*, $\ell(h)$ of an hypothesis h. By Kraft's inequality, we can express the number of bits needed to encode an hypothesis h as a negative log-likelihood $\ell(h) = -\log p_{\text{code}}(h)$ for some distribution $p_{\text{code}}(h)$. Conversely, for a given $p(h)$, a lossless compression scheme whose code-lengths almost exactly match $-\log p(h)$ can be constructed with arithmetic coding.

Example. Let X be the space of all images and let H be the set of all English sentences. The cost of encoding $h \in H$ can be computed as the negative log-likelihood $-\log p_{\text{code}}(h)$ assigned to h by a trained large language model (LLM). The log-likelihood $\log p(x|h)$ of an image x given a sentence h can be computed using a conditional image generation model, such as a diffusion model.

Optimal descriptions. Suppose we want to describe a sample x, but we have an upper-bound $\ell(h) \leq C$ on the length of the description h we can use. To find the best description $h_x^*(C)$ to use under this capacity constraint, we have to

Figure 2. These two images have similar art styles, theme, and subject matter. On the other hand, it is difficult to identify specific visual elements that appear in both images. These two images were found to be "substantially similar" [37] based on the arrangement of *similar* features in a *similar* way. How can we measure how similar these images are?

solve the constrained optimization problem:

$$h_x^*(C) = \underset{h \in H}{\arg\min} \ \ell(x|h) \tag{1}$$
$$\text{s.t.} \quad -\log p_{\text{code}}(h) \leq C,$$

which aims at finding the hypothesis h with $\ell(h) \leq C$ which minimizes the reconstruction error for x. This definition however has practical limitations: since the space H is discrete, the optimization in Eq. (1) cannot be performed easily with gradient-based methods, leading to an expensive search over the (potentially infinite) hypothesis space H.

Stochastic relaxation. To simplify the problem, rather than considering the best *single* hypothesis, we can search for a *distribution* of hypotheses $q(h)$ describing the image. This corresponds to a stochastic relaxation of the problem where we find a low-complexity *distribution* of hypotheses $q_x^*(h|C)$ that, on average, have good reconstruction loss:

$$q_x^*(h|C) = \underset{q(h) \in \mathcal{P}(H)}{\arg\min} \ \mathbb{E}_{h \sim q(h)}[\ell(x|h)] \tag{2}$$
$$\text{s.t.} \quad \text{KL}\big(q(h) \,\|\, p_{\text{code}}(h)\big) \leq C,$$

where the optimization is now over the space of distributions $q(h) \in \mathcal{P}(H)$. When $q(h)$ is restricted to being a Dirac delta over a single hypothesis — i.e., $q(h) = \delta_{h_0}(h)$ — we exactly recover Eq. (1). A natural question is in which way the newly introduced term $\text{KL}\big(q(h) \,\|\, p(h)\big)$ generalizes the description length $-\log p_{\text{code}}(h)$ of Eq. (1), aside from reducing to it in the extreme case where $q(h)$ is a Dirac delta. To answer, it can be shown that $\text{KL}\big(q(h) \,\|\, p(h)\big)$ is the expected number of bits required

to specify a fair sample of the distribution $q(h)$ to a receiver that knows $p(h)$ but not $q(h)$ [17, Lemma 1.5]. For this reason, we refer to Eq. (2) as a stochastic relaxation of Eq. (1). The advantage of this formulation is that the solution to the optimization problem can be efficiently approximated by sampling a large language model, as we will discuss later. This formulation also connects to other frameworks (see Sec. 4).

3.2. Conceptual Similarity

As anticipated in Section 1, given two sample x_1 and x_2, we want to define a notion of *conceptual similarity* by measuring how well optimal descriptions of x_1 also apply to x_2 — and vice versa – as the complexity of the description increases. To formalize this notion, let $q_i(h|C)$ denote the family of distributions of optimal descriptions h for the sample x_i as C varies. We can define the function $\Delta_{2 \to 1}(C)$

$$\Delta_{2 \to 1}(C) := \mathbb{E}_{q_1}[\ell(x_1|h)] - \mathbb{E}_{q_2}[\ell(x_1|h)]. \quad (3)$$

which measures how well the description q_2 of x_2 describe x_1, compared to its optimal description q_1. Similarly, we can define $\Delta_{1 \to 2}(C)$ inverting the role of x_1 and x_2.

Conceptual Distance. We define the *conceptual distance* between x_1 and x_2 at complexity level C as

$$d_{x_1, x_2}(C) = \frac{1}{2} \Big(\Delta_{2 \to 1}(C) + \Delta_{1 \to 2}(C) \Big). \quad (4)$$

Figure 1 provides a sample illustration of the typical behavior of this distance. For low values of C (y-axis), the optimal descriptions of one sample — e.g., "a red car" — is an equally good description of both. In this case, both terms $\Delta_{i \to j}$ (blue and orange curves in the plot) will be zero. As the complexity C increases, the optimal description of one sample — e.g., "a Ferrari" — is not an optimal description for the other and the distance becomes non-zero, and will keep increasing as the descriptions become more detailed and specific.

AUC. Rather than being a single number, our definition of conceptual distance is a function of the complexity C of the description. As discussed in Section 4, this is necessary to address some key issues in defining distances and we posit that human perception of similarity relates to how quickly this function grows. To capture this, when a single number is necessary to define a distance, we use the Area Under Curve of the graph, up to some vale $C_{\max}$.

Asymmetric distances. The distance $d_{x_1, x_2}(C)$ in Eq. (4) in written in terms of two *asymmetric distances*, the distance $\Delta_{1 \to 2}$ of x_1 from x_2, and $\Delta_{2 \to 1}$ from x_2 to x_1. While $d_{x_1, x_2}(C)$ provides an overall measure of the distance between samples, each individual asymmetric distances can be useful to measure directed containment relations, for instance to what extent a work is being subsumed by or supersedes another.

Interpretation as intersection. We note that, rearranging the terms, Eq. (4) can also be rewritten as:

$$d_{x_1, x_2}(C) = \mathbb{E}_{q_1}[\ell(x_1|h)] + \mathbb{E}_{q_2}[\ell(x_2|h)] \\ - \mathbb{E}_{q_\cap}[\ell(x_1|h) + \ell(x_2|h)] \quad (5)$$

where $q_\cap(h|C) = \frac{1}{2} q_1(h|C) + \frac{1}{2} q_2(h|C)$. This compares the reconstruction loss obtained by using different optimal descriptions for x_1 and x_2 and the reconstruction loss obtained by encoding both x_1 and x_2 using the *same* descriptions $h \sim q_\cap(h|C)$. Intuitively, if the samples are similar, a shared description equally described each image when compared to picking the optimal descriptions independently. However, as the complexity of C increases, $q_1(h|C)$ and $q_2(h|C)$ will contain increasing details that describe one sample x_1 (respectively x_2) but not equally well the other. In that case $q_\cap$ becomes suboptimal and the distance grows accordingly.

Interpretability. An advantage of our definition of conceptual distance is that the descriptions h that have high probability under $q_\cap$ provide an interpretable explanation for why the two samples should be considered similar (e.g., gray captions in Figure 1). Conversely, looking at h that are likely under q_i but not $q_\cap$ provides an explanation of the unique information in q_i, justifying why q_1 and q_2 should be considered different.

4. Motivation and Discussion

Defining a conceptual distance face three main challenges:
  (i) **Randomness dominates.** Any pair of non-identical real images has a large number of small differences due to randomness which dominates over the few, but important, structural similarities;
 (ii) **Canonical importance of properties.** What properties are important for similarity is non-canonical;
(iii) **Adversarial discrimination.** Any two images may be "adversarially" distinguished by simple and obvious properties (e.g., "the color of the car is different") but ease of *discriminability* should not affect *similarity* (two pictures of the same car model should be similar, even if the color is different).

Corresponding to these challenges there have are three main classes of distance functions that address some of the problems, but fail to capture the others.

Contrastive learning. CLIP-like contrastive learning models by their nature rely on discriminative features that are optimized to discriminate samples, not to describe concepts that humans evince from said samples, thus failing (iii). Since these representations are not trained to align with human conceptual representations, they are a poor fit to measure conceptual similarity as noted in recent works [10, 43]. Indeed, we show in Tab. 2 that our method significantly

outperforms CLIP in outputting human-aligned similarity scores for images. The method also fails (ii) since the model learns semantically relevant properties either through semi-supervised image-text or enforced through hard-coded strong data augmentation.

Information theoretic distances, such as the Normalized Compression Distance (NCD), move away from discriminability and instead define a canonical notion of similarity between samples as the ratio between the amount of shared (algorithmic) information and the total information. However, reflecting (i), unique randomness in the data accounts for most of the information in high dimensional samples. Hence, even when samples are perceptually similar, random information will dominate the distance and similarity of key semantic information is lost (Appendix A).

Structure function. Defining a distance for high-dimensional data requires separating "structural" information of the sample from the information due to randomness and noise, raising the question of how to canonically define what constitutes "structure" and what "noise". Kolmogorov proposed that describing structural information should reduce the reconstruction loss of a sample more quickly than describing random details which are intrinsically incompressible. This is captured by the Structure Function $\beta_x(C)$, defined as

$$\beta_x(C) = \min_{h \in H} \; -\log p_h(x) \qquad (6)$$
$$\text{s.t. } |h| \leq C,$$

which measures the reconstruction error when compressing x using a program $h \in H$ implementing a computable probability distribution $p_h(x)$, as the length $|h|$ of the program varies. Since the optimal descriptions have to reduce the reconstruction loss as quickly as possible, for low values of C they can only contain structural information and not random information. This reasoning can be formalized [39] and allows to define a notion of *algorithmic minimal sufficient statistic* that captures structural but not the random information. Note that this directly relates to our Eq. (1), using $p_{\text{code}}(h) = 2^{-|h|}$, with the only difference that the optimization is done over programs instead of sentences. One of the main contributions of our work is to show that something akin to the structure function can be used to define a notion of conceptual distance, and to introduce a more computationally feasible stochastic relaxation of the framework. However, in its pure form this approach still fails. It can be shown that algorithmic minimal sufficient statistics are trivial for most data [13, Corollary III.13] and, when non trivial, they do not capture useful semantic information but rather capture the mutual information between the sample and the halting problem [13, Theorem III.24]. As we show later, this is a symptom of a more general problem where no expressive enough class of descriptions (such as

general programs) can define non-trivial semantic information. This motivates our choice to restrict descriptions to (non-canonical) subclasses such as language sentences

Conceptual Distance. Our method attempts to tackle all the problems (i)-(iii). First, it does not focus on discriminating samples. Rather we independently find the descriptions of each sample, and then evaluate them on each other. This prevents the method from adversarially picking features that may be good discriminators, but that are not good descriptors – thus addressing (iii). However, we need to ensure that the description will focus on structural properties of the image, and not on random details. Similarly to Kolmogorov's Structure Function approach, we accomplish this by finding optimal descriptions under a capacity constraint which naturally leads the distance to ignore differences due to incompressible randomness for small values of C.

Aside from increasing robustness to unimportant differences, focusing on the initial part of the curve as C grows has other advantages. We posit that human perception of semantic similarity relates to how quickly the distance function grows, rather than its asymptotic value. This motivates our choice of using the AUC of $d_{x_1,x_2}(C)$ up to some small value of C to measure similarity. Indeed, in our experiments we observe that the AUC computed up to some relatively small value of C is better aligned with human similarity annotations than the exact value at any specific C, and that using larger values of C decreases alignment.

Our distance implicitly defines a separation between structural and random components of the images. This depends on their "compressibility" and is in turn dependent on the choice of hypothesis space, encoder and decoder. Our choice of using language as the hypothesis space, and a particular conditional generative model, is not canonical prompting the question of whether a more canonical choice, like Kolmogorov's choice of using generic programs, would be better. However, as the following theorem shows, any model class which is too expressive would not be able to recover a meaningful notion of structural information.

Theorem 4.1 (No canonical definitions of structure, Appendix C)**.** Let H be a class of hypotheses and let $p(x|h)$ be the corresponding decoder. If the decoder $p(x|h)$ is expressive enough to perform perfect test-time optimization, then all samples have the same structure, and the conceptual distance between any pair of samples is zero.

The lack of useful structure emerging from general program classes motivated us to consider an appropriate subclass that could lead to a more useful, though necessarily non-canonical as expressed in property (ii), measure of conceptual similarity. Copyright doctrine acknowledges the absence of a computable notion of similarity, and instead relies on judges and juries to adjudicate each case. We analogize this human decision-making process to a com-

	Zero-Shot Method	STS-B	STS12	STS13	STS14	STS15	STS16	SICK-R	Avg
Paragon	CLIP-ViTL14 [34]	65.5	67.7	68.5	58.0	67.1	73.6	68.6	67.0
	SimCSE-BERT [15]	**68.4**	**82.4**	**74.4**	**80.9**	**78.6**	**76.9**	**72.2**	**76.3**
LLaMA [38]	Cond. Likelihood	44.3	20.8	51.8	38.6	56.0	50.9	56.7	45.6
	Meaning as Trajectories [28]	70.6	52.5	65.9	53.2	67.8	**74.1**	73.0	65.3
	Ours	**72.3**	**56.7**	**67.9**	**56.9**	**68.7**	74.1	**74.3**	**67.3**

Table 1. **Text-to-Text semantic similarity benchmarks.** We evaluate our method using Spearman correlation ($\times 100$) on the STS sentence similarity benchmark. Our method outperforms all zero-shot methods based on vector embeddings. On the SICK-R dataset involving compositional knowledge, conceptual similarity outperforms even the contrastive-trained paragon by 2.1%. This shows that our definition of Conceptual Distance is better aligned with human judgement than standard vector embedding. It also outperforms [28], which can be seen as a particular case of our method for a fixed value of $\lambda = 1$, while greatly improving interpretability. This shows the importance of considering different levels of capacity to recover human judgement.

putable machine-learning process, in which *natural language programs* attempt to describe and distinguish the works in question. To operationalize this notion, we reframed Kolmogorov complexity in Sec. 3 to use natural language as opposed to formal language programs for encoding inputs. Natural language descriptions are generated by models that aggregate subjective assessments not just of a single judge or a small group of jurors, but from textual expressions of the millions of individual contributions used to train large-scale models. While our approach makes a subjective choice of using natural language as a representation of the data, such a representation is naturally fit to the task, since natural language is routinely used, and arguably has evolved, to express and communicate abstract concepts.

5. Practical distance computation

To compute our conceptual distance we need to solve the constrained optimization problem Eq. (2). The solution can easily be written using the Lagrange multiplier method (Appendix F) and leads to the family of optimal distributions:

$$q_x^*(h|C) = \frac{1}{Z_\lambda} p_{\text{code}}(h) p(x|h)^\lambda, \qquad (7)$$

where $Z_\lambda = \mathbb{E}_{p_{\text{code}}(h)}[p(x|h)^\lambda]$ is the normalization factor and $\lambda = \lambda(C) \geq 0$ has to satisfy the condition:

$$\text{KL}\left(q_x^*(h|C) \| p_{\text{code}}(h)\right) = C. \qquad (8)$$

Having $q_x^*(h|C)$ we can compute the various quantities in the definition of the distance. However, given that $q_x^*(h|C)$ is a distribution over an infinite hypothesis space H, we cannot represent it explicitly. Fortunately, all the quantities involved are in the form $\mathbb{E}_q[f(h)]$ for some function $f(h)$ and can be estimated through importance sampling. Let $\pi(h)$ be a proposal distribution and let $h_1, \ldots, h_N \sim \pi(h)$ be samples from it. Then we have (Appendix F):

$$\mathbb{E}_{q^*}[f(h)] = \mathbb{E}_\pi\left[\frac{q^*(h)}{\pi(h)} f(h)\right] \approx \sum_{i=1}^N \alpha_{\lambda,i} f(h_i)$$

$$\alpha_{\lambda,i} := \text{softmax}\left(-\lambda \ell(x|h_j) + \log \frac{p_{\text{code}}(h_j)}{\pi(h_j)}\right)_i$$

Hence, computing the importance weights $\alpha_{\lambda,i}$ for each sample h_i (all needed quantities are available) we can easily compute an unbiased estimator of our distance function.

Using only the encoder model Typically, we have access to an encoder model $p(h|x)$, which outputs description h given the data x — *e.g.,* a captioning model — but may not have a corresponding generative model $p(x|h)$ to evaluate the likelihood of the data given the description. This problem can be circumvented. By Bayes's rule we have

$$p(x|h) = \frac{p(h|x)p(x)}{p(h)}.$$

The encoder model directly provides $p(h|x)$. The likelihood $p(x)$ of the data is however is difficult to evaluate for high-dimensional data. Luckily, the role of $p(x)$ simplifies in our definition of distance (Appendix E) so the term is not needed. Finally, $p(h)$ — which should not be confused with $p_{\text{code}}(h)$ — is the marginal $p(h) = \sum_{x \in X} p(h|x)p(x)$, which can be estimated through sampling. In practice however, we found that the simple approximation: $p(h) \approx \frac{1}{2}p(h|x_1) + \frac{1}{2}p(h|x_2)$ performs well and can be computed for free using only the available two samples.

Choice of $\pi(h)$. To reduce the variance of importance sampling, the proposal distribution $\pi(h)$ should be well aligned with the distributions we want to compute the expectation with, in our case $q_1(h|C)$ and $q_2(h|C)$. Recall that

$$q_i(h|C) = \frac{1}{Z_\lambda} p_{\text{code}}(h) \left(\frac{p(h|x)}{p(h)}\right)^{\lambda(C)}$$

which for $\lambda(C) = 1$ simplifies to the posterior $p(h|x_i)$, and will generally be close to it at most important values of λ, making it a good choice for the proposal distribution. Since we want to be able to use the same samples for both $q_1(h|C)$ and $q_2(h|C)$, we define the proposal distribution as:

$$\pi(h) := \frac{1}{2}p(h|x_1) + \frac{1}{2}p(h|x_2).$$

Sampling from this distribution can be achieved easily by sampling $\frac{N}{2} + \frac{N}{2}$ samples from $p(h|x_1)$ and $p(h|x_2)$.

Choice of $p_{\text{code}}(h)$. The choice of the encoding distribution $p_{\text{code}}(h)$ can be used to align the conceptual distance with human similarity assessments. A natural choice is to take $p_{\text{code}}(h)$ to be the likelihood assigned to the description h by a LLM. Another choice, which is more natural from a computational perspective is to take $p_{\text{code}}(h) = \pi(h)$ in order to reduce the variance of the importance sampling estimation. We found the first choice to give more interpretable results (and we use it in our qualitative plots), whereas the latter choice performs better on large scale benchmarks.

6. Experiments

The goal of this section is to evaluate how well our notion of conceptual distance correlates with ground-truth human annotators. To this end, we empirically compare samples under three different domains — text-to-text, image-to-image, and text-to-image (cross-modal) similarity.

Model choice. To implement our proposed similarity score, we require a model $p_{\text{code}}(h)$ to compute code length of text descriptions h, as well as conditional distributions like $p(h|x)$, where x can represent text or images. For text encoding we use $p_{\text{code}}(h) = \frac{1}{2}\big(p(h|x_1) + p(h|x_2)\big)$ on benchmarks, while on qualitative experiments we use the likelihood of h outputted by LLaMA-13B [38]. We also use LLaMA-13B as conditional text encoder $p(h|x)$, while when x is an image we use LLaVA [26]. In the latter case, images are encoded as a sequence of tokens using a vision transformer, and then fed into a general sequence transformer that is trained on both image and text tokens.

Importance sampling. For comparing a pair of samples (x_1, x_2), we use descriptions $H_{\text{sample}} = H_1 \cup H_2$ sampled from the proposal distribution $\pi(h)$, where $H_1 \sim p(h|x_1)$ and $H_2 \sim p(h|x_2)$ are the set of descriptions sampled from x_1 and x_2 respectively. To ensure fair comparison, we use the same sampling procedure and hyperparameter choices on [28] and generate from each input 20 descriptions of 20 tokens through multinomial sampling.

Sentence similarity. First, we establish whether CC:DAE is able to compute a meaningful notion of conceptual similarity between sentences which aligns with ground-truth human annotations. Towards this end, we leverage the Semantic Textual Similarity (STS) benchmark [2–6, 9, 29], where each sentence pair (x_1, x_2) is labelled with a similarity score attributed by human annotators. To quantify this alignment, we measure the Spearman correlation between our conceptual distance and the ground-truth scores. In particular, for each input sentence x, we sample continuations $h \sim \pi(h) = \frac{1}{2}(p(h|x_1) + p(h|x_2))$ of the two sentences, which we use as descriptions to estimate their distance. Our results in Tab. 1 demonstrates that our method correlates well with human judgements. In particular, conceptual sim-

Method	CxC-SIS	CxC-SITS	Average
CLIP-ViTB/16	72.08	64.60	68.34
Cond. Likelihood	-	29.46	-
Meaning as Trajectories [28]	**81.47**	67.63	74.55
Ours	81.44	**67.71**	**74.58**

Table 2. **Image-Image/Image-Text Semantic similarity benchmarks.** Evaluation on human-alignment of similarity scores on the CxC Semantic Image Similarity (SIS) and Semantic Image-Text Similarity (SITS) benchmarks. CC:DAE outperforms CLIP, which is trained explicitly on the multi-modal similarity task, showing that CC:DAE better captures human intuition compared to contrastive methods. On CxC-SITS the method outperforms baselines using the same backbone and, by leveraging the alignment prompt used by [28], our method achieves comparable performance while benefiting from enhanced interpretability.

ilarity outperforms all methods in the literature that have not been explicitly trained on human annotations, with a relative decrease in average error by -5.8% compared to the next best method, which also uses the same backbone. Moreover, our method outperforms CLIP-ViTL14, which as been trained explicitly on contrastive objectives. Lastly, on the SICK-R (Sentences Involving Compositional Knowledge) dataset, conceptual similarity achieves state-of-the-art performance among zero-shot methods, showing that our method excels at representing compositional structures present in the data.

Image similarity. Next, we want to determine whether our conceptual distances correlates with that of human annotators when used to compare images. To accomplish this, we leverage the Crisscrossed Captions Semantic Image Similarity (CxC-SIS) [33] benchmark containing pairs of images annotated by humans with similarity scores ranging from 0 to 5. To evaluate our method, we measure the Spearman correlation of conceptual distance and human annotations. In Tab. 2, we show that our method outperforms all prior zero-shot methods on CxC-SIS, including reducing the error -33.5% relative to CLIP. This supports the idea that our method, based on comparing samples via descriptions, fares better than methods based on discrimination.

Cross-modal similarity. We further investigate whether conceptual similarity can be applied when inputs do not come from a common modality. In particular, we compare on the CxC-SITS Semantic Image-Text Similarity (CxC-SITS) benchmark, which provide human annotated scores on (image, text) pairs based on their semantic similarity. Our experiments in Tab. 2 demonstrate that our method indeed generalizes well to cross-modal comparisons, reducing error relative to CLIP by -8.8%. To further judge the improvement of our method over the base capability of the backbone, on CxC-SITS we compare against a baseline that selects captions by directly comparing the conditional like-
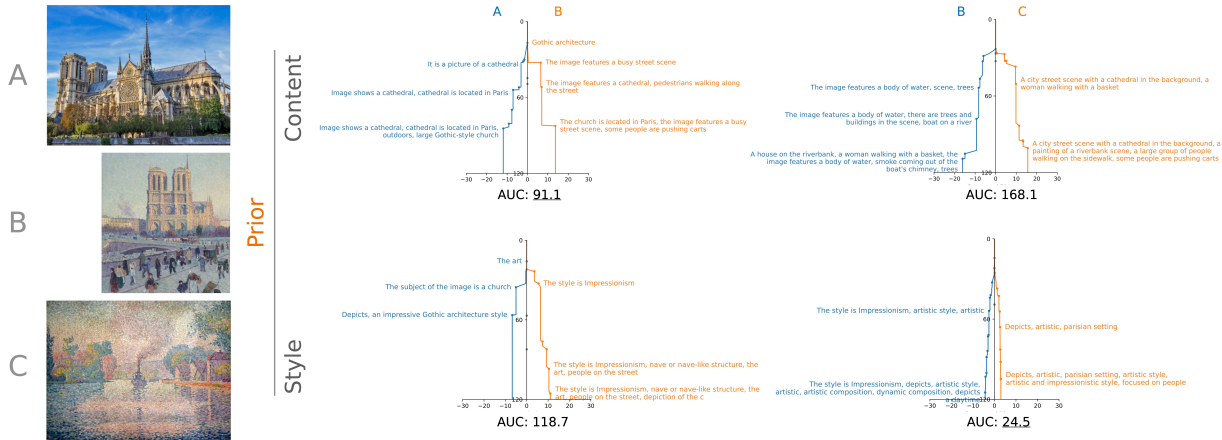
Figure 3. **Role of prompts.** Consider the three images above: which pair is most similar? This depends if we focus on content — the first two depict Notre-Dame, the third a boat on the Seine — or on the style/artistic technique — the first is a photograph, the second and third are paintings in the pointillist style. By changing the prompt ("Describe the style of the image" or "Describe the content of the image"), the user can bias $p(h|x)$ to focus the conceptual distance on one or the other aspect. Note that images A and B are closer under the content prompt, but B and C are closer under the style prompt.

lihood of the text given the image. We note that this baseline fares poorly, which is aligned with results showing that perplexity is be a poor metric for caption alignment [41]. CC:DAE bypasses this limitation by autoencoding both text and images in a common text description space. CC:DAE also matches or outperform — both in Tabs. 1 and 2 — the best zero-shot method [28], which can be seen as a very particular case of our method (Appendix D) while it benefits from enhanced interpretability. Our method also outperforms [28] and comparable baselines on the more challenging SugarCrepe benchmark (Appendix B).

Qualitative results. In Figs. 1 and 3 we illustrate the behavior of our conceptual distance. To generate richer and more interpretable samples from the description space — and to further highlight the trade-off between coding length C and distance — rather than sampling directly $h \sim p(h|x_i)$ we use a beam search to generate effective descriptions of increasing length (Appendix G). For visualization purposes, rather than optimizing a distribution, we restrict the optimization to select only the single best performing description among the samples for each C and display it. In all cases, we see that as expected the distance is zero when restricted to short description (small C). As descriptions become more detailed, the distance starts increasing. Each increase in distance can be interpreted looking at the selected h. While selected descriptions highlight the differences, we can also look at the best description h explaining both images at the same time (see Appendix) to understand what common structure the two images contain which justifies why the distance is not larger.

Prompting. Conceptual distance is influenced by the

choice of the description space H and the decoder $p(x|h)$, or equivalently the encoder $p(h|x)$. In Sec. 4 we proved that making such non-canonical choices is inevitable. In Fig. 3 we show that freedom to modify the encoder probability — for example through prompting — is indeed not a disadvantage but a helpful option. Different prompts such as "Describe the style of the image" or "Describe the content of the image" allow the user to easily bias the conceptual distance to focus on the similarity axis of interest.

7. Conclusions

We introduce CC:DAE, a principled notion of conceptual similarity which aims to explore and resolve past issues in the definition of semantic distance between high dimensional data. Empirically, the method achieves state-of-the-art results for matching human similarity judgments without fine-tuning on human scores. We built on on Kolmogorov's central insight that structured information should be easier to explain than randomness. While using general programs as explanations does not yield acceptable results, selecting natural language sequences as our "programs" is enough to spotlight the properties in images and text that are relevant to humans. Our general strategy can flexibly incorporate other types of explanations, including visual features, to specify aspects of conceptual similarity that are relevant for specific contexts. In our current implementation, the description space is limited to textual description. This prevents the method from efficiently describing similarities deriving from similar visual arrangement, which is however an important space. We leave to future works extension of the method using more generic decoders that can take both textual and visual features in input.

References

[1] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018. 2

[2] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, 2012. 7

[3] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43, 2013.

[4] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91, 2014.

[5] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263, 2015.

[6] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511*. ACL (Association for Computational Linguistics), 2016. 7

[7] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Semantic bottleneck for computer vision tasks. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 695–712. Springer, 2019. 3

[8] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018. 2

[9] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017. 7

[10] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022. 2, 4

[11] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 1

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[13] Péter Gács, John T Tromp, and Paul MB Vitányi. Algorithmic statistics. *IEEE Transactions on Information Theory*, 47 (6):2443–2463, 2001. 2, 5, 13

[14] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 12

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 3, 6

[16] Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. *Advances In Neural Information Processing Systems*, 29, 2016. 2

[17] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. In *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 10–23. IEEE, 2007. 4

[18] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International conference on learning representations*, 2016. 2

[19] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023. 12

[20] Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*, 2022. 3

[21] Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645*, 2023. 3

[22] William B Johnson and J Lindenstrauss. *Extensions of Lipshitz mapping into Hilbert space*. 1984. 1

[23] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 2

[24] Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. *arXiv preprint arXiv:2310.07972*, 2023. 3, 13

[25] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*. Springer, 2008. 1, 11

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 7, 12

[27] Nan Liu, Yilun Du, Shuang Li, Joshua B Tenenbaum, and Antonio Torralba. Unsupervised compositional concepts discovery with text-to-image generative models. *arXiv preprint arXiv:2306.05357*, 2023. 2

[28] Tian Yu Liu, Matthew Trager, Alessandro Achille, Pramuditha Perera, Luca Zancato, and Stefano Soatto. Meaning representations from trajectories in autoregressive models. *arXiv preprint arXiv:2310.18348*, 2023. 3, 6, 7, 8, 12, 13, 14

[29] Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8, 2014. 7

[30] Lingjie Mei, Jiayuan Mao, Ziqi Wang, Chuang Gan, and Joshua B Tenenbaum. Falcon: fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations. *arXiv preprint arXiv:2203.16639*, 2022. 2

[31] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022. 3

[32] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021. 3

[33] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. *arXiv preprint arXiv:2004.15020*, 2020. 7

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6, 12

[35] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 12

[37] Steinberg v. Columbia Pictures Industries, Inc. 663 F. Supp. 706, (S.D.N.Y. 1987). 1, 3

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 6, 7

[39] Nikolai K Vereshchagin and Paul MB Vitányi. Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004. 5, 13

[40] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997. 1

[41] Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. Perplexity from plm is unreliable for evaluating text quality. *arXiv preprint arXiv:2210.05892*, 2022. 8

[42] Satoshi Watanabe. Knowing and guessing: A quantitative study of inference and information. 1969. 13

[43] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 2, 4

[44] Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. An unsupervised sentence embedding method by mutual information maximization. *arXiv preprint arXiv:2009.12061*, 2020. 3