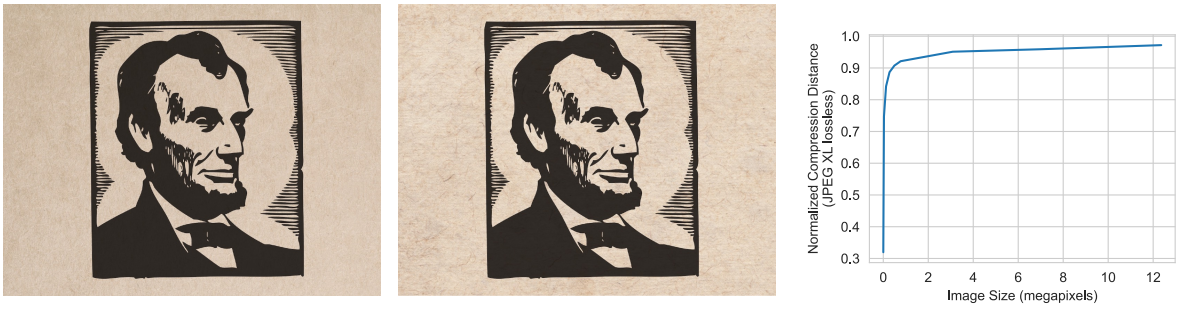

Figure 4. **Conceptual Similarity is not the same as shared information.** Are these two pictures similar? Not according to Normalized Compression Distance, which measures their difference at 97.1% (estimated using JPEG XL lossless compression). However, they share all the structural information — they are the exact same ink print on a piece of paper. The only difference is the randomness of the paper texture. Most people would not consider it a significant conceptual difference, but since NCD cannot differentiate structure from randomness, this slight change accounts for 97.1% of the difference. This problem is inherent in high-dimensional data where information in random variation overshadows structural information. In fact, on the right we plot the NCD distance as we change the resolution of the image, showing that the distance increases drastically as we increase the dimension of the data.

A. Limitations of Information Theoretic Distances

In Fig. 4 we show that two images that are conceptually identical are considered almost completely different by the Normalized Compression Distance (NCD), an information theoretic distance. As we now discuss, this is an intrinsic problem of information theoretic distance, which cannot distinguish differences due to structural properties from differences due to randomness, the latter becoming dominant as the number of dimensions grows (Fig. 4, right).

NCD [25] defines the distance between samples in terms of their common (algorithmic) information. Let x and y be two samples. We denote with the Kolmogorov complexity $K(x)$ the length of the shortest program that can output x (equivalently, up to a constant, its best possible compression cost using commutable function), and with $K(x|y) = K(xy) - K(y)$ the length of the shortest program that can reconstruct x given y as input. If x does not contain any information that is not already contained in y, then $K(x|y) \approx 0$ and we can consider x to be similar to y. Making the role of x and y symmetric, this motivates the following definition:

$$\begin{aligned}\mathrm{NCD}(x,y) &= \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \\ &= \frac{K(xy) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}.\end{aligned}$$

While this is a theoretically viable definition, in practice the shortest coding length $K(x)$ cannot be computed explicitly. However, the distance can be approximated using a strong compression algorithm $Z(x)$ as follows:

$$\mathrm{NCD}_Z(x,y) = \frac{Z(xy) - \min\{Z(x), Z(y)\}}{\max\{Z(x), Z(y)\}}.$$

On the surface, this distance is well positioned to capture shared algorithmic structure, and hence recover meaningful similarities between the samples. However, this is not the case when data can be noisy. The following proposition shows that two pictures that differ only by some slight noise — for example two consecutive photos differing only by sensor noise — always have close-to-maximal NCD.

Proposition A.1. Let s be an image of low complexity $K(s)$. Suppose that two measurements $x := s \oplus n_x$ and $y := s \oplus n_y$ are generated by adding Bernoulli noise $n_x, n_y \sim \mathrm{Bern}(p)$ to s, where $\oplus$ denotes the bit-wise XOR. Denote with $H(p)$ the entropy of the Bernoulli distribution. Then, in the limit of large dimension $D = |s|$ we have:

$$\mathrm{NCD}(x,y) \approx 1 - \frac{K(s)}{D\,H(p)} \xrightarrow{D\to\infty} 1$$

Proof. If the noise p is small, the optimal way to compress x — and similarly y — is to simply encode both s and n_x independently:

$$K(x) \approx K(s) + K(n_x) = K(s) + |s|H(p).$$

Similarly, the cost of encoding x and y together is the cost to encode (once) the shared s and the two noise masks:

$$K(xy) = K(s) + K(n_x) + K(n_y) = K(s) + 2|s|H(p).$$

Source	Model	REPLACE			SWAP		ADD	
		Object	Attribute	Relation	Object	Attribute	Object	Attribute
OpenAI [34]	RN50x64	94.5	83.5	70.6	61.8	66.7	83.3	74.0
LAION [36]	ViT-bigG-14	96.7	88.1	74.8	62.2	74.9	92.2	84.5
	xlm-roberta-large-ViT-H-14	96.9	86.0	72.1	63.8	72.1	93.1	86.1
DataComp [14]	`xlarge`: ViT-L-14	95.5	84.5	67.0	65.0	66.8	91.0	85.0
LLaVA[26]	Cond. Likelihood	78.8	77.7	73.3	77.6	86.0	36.2	76.2
	Meanings as Trajectories [28]	90.4	80.6	78.8	69.9	76.6	75.7	82.8
	CC:DAE (ours)	91.0	82.1	82.2	73.6	78.8	77.0	86.0

Table 3. **Performance on SugarCrepe multi-modal image-caption alignment benchmark.** We show that CC:DAE can be extended to compute similarity between data of different modalities. CC:DAE outperforms or matches all the baseline contrastive-based models (numbers from [19]) on 4 out of 7 tasks. Compared to methods using our same backbone, we significantly outperform the conditional likelihood baseline. We also uniformly outperform [28] which uses the same backbones and trajectories as our method.

Putting all together we get:

$$\mathrm{NCD}(x,y) = \frac{|s|H(p)}{K(s)+|s|H(p)} = \frac{1}{1+\frac{K(s)}{|s|H(p)}}$$
$$= 1 - \frac{K(s)}{|s|H(p)} + o\Big(\frac{K(s)}{|s|H(p)}\Big)$$

as we wanted. □

To demonstrate this effect empirically, in Fig. 4 we generate two images x and y using the same basic picture s, but adding two different noise pattern n_x and n_y (the different paper textures). We then compute NCD_Z using as compressor Z the recent JPEG XL lossless codec (which gave the best compression across the tried codecs). Since Z does not support compression of two images simultaneously, instead of $Z(xy)$ we use the lower-bound:

$$Z(xy) \geq Z(x) + Z(y) - Z(s).$$

The empirical behavior — even if using a suboptimal compression scheme and using correlated noise — indeed follows the theoretical prediction.

B. Additional results

Multi-Modal Similarity on SugarCrepe. Our method can also be used to compute similarity between data in different modalities, as long as they share the description space H. To test this, we evaluate our method on SugarCrepe [19], a vision-language compositionality benchmark framed as a binary classification task: given an image and a pair of candidate captions, the task is to select the right caption for the image. Negative captions in each pair are generated from the ground-truth caption as "compositional distractors" via replacing, swapping, or adding atomic concepts. Since caption pairs differ only by an atomic concept, effective methods require capturing compositional structures in both image and text modalities. We use using CC:DAE for classification by computing the conceptual distance between the image and each caption, and selecting the caption that yields the lowest distance.

For our experiments on SugarCrepe, we generate via multinomial sampling 10 trajectories of maximum 10 tokens from each input image or candidate caption as descriptions for our method. For fair comparison, we apply the same settings for the Meaning as Trajectories baseline.

In Tab. 3, we show that conceptual similarity matches or outperforms all multi-modal contrastive-based models (which are trained specifically for this task) on 4 out of 7 tasks. This holds especially for the hardest benchmarks measured by the poor performance of the paragon contrastive models: *Replace Relation*, *Swap Object*, *Swap Attribute*, and *Add Attribute*. Our model also significantly outperforms the conditional likelihood baseline in most tasks — which notably performs even worse than random guessing on the *Add Object* benchmark — and also uniformly outperforms [28] on all 7 tasks when using the same sampled descriptions for both methods.

Importance of varying C. In Fig. 5 we plot the correlation between human ground-truth and the distance computed by our method for different fixed values of the capacity C. We see that using a large capacity C, and hence more descriptive captions, is actually detrimental: when the capacity C of the description is too high, the correlation between our distance and human ground-truth significantly worsens. This is indeed one key motivation, inspired by Kolmogorov's framework, for varying C. If the description is too complex, we can always find a large number of dif-

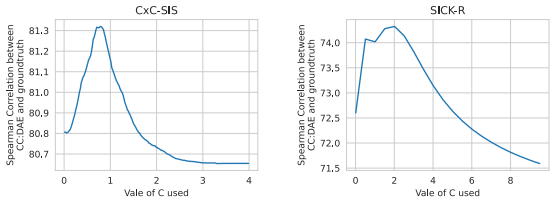

Figure 5. Performance of CC:DAE distance as C increases.

ferences, even if these distinctions are meaningless. This is also formalized by the "Ugly duckling theorem" [42]: if all possible properties are listed, a duckling is as similar to a swan as two swans are to each other, and is supported by our analysis in Appendix A and Figure 4, where we prove analytically that a large C leads to degenerate distances even for two visually indistinguishable images. We also empirically observe that using AUC instead of a single value of the curve improves performance on 11/16 tasks and generally makes the results less sensitive to the choice of C.

Using diffusion models to compute CC:DAE. In Section 5 we introduced a method to compute the CC:DAE distance without requiring access to a generative model, which simplifies the implementation of the method. Similar qualitative results can however be obtained computing $p(x|h)$ directly using a generative model. In Figure 6 we replot Figure 1 using Stable Diffusion as the generative model, and computing the log-likelihood $p(x|h)$ as described in [24].

C. Non-Existence of a Canonical Structure

In this section we want to prove Theorem 4.1:

Theorem (No canonical definitions of structure, informal)**.** 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.

We provide a general sketch of proof in the general setting, and we refer the reader to [13] for a more technical proof in the specific case of $H = \{\text{computable distributions}\}$. First, we need to introduce some additional definitions — adapted from [39] — to formalize the notion of structure of x as all the non-random information contained in x. Given an hypothesis class H, consider the function:

$$\lambda_x(C) = \min_{h \in H} \; -\log p_h(x) - \log p_{\text{code}}(h) \qquad (9)$$
$$\text{s.t.} \; -\log p_{\text{code}}(h) \leq C.$$

This function closely relates to our optimization problem Eq. (1), and can be seen as the compression code for x using a two part code that first specifies a description $h \in H$ — with cost $\ell(h) = -\log p_{\text{code}}(h)$ — and then uses it to encode x using $-\log p_h(x)$ bits. As C grows, and we can use better fitting descriptions, $\lambda_x(C)$ decreases until it reaches a minimum non-zero value, which is the best compression cost achievable using this class. We call a hypothesis h *sufficient statistic* if it witnesses this minimum. A sufficient statistic of x describes all the properties that are compressible under H, however it may also encode random bits of incompressible (non-structural) information. To prevent this, we define a *minimal sufficient statistic* as any sufficient statistic h which has minimal coding length $\ell(h)$ (equivalently, h witnesses the first point where $\lambda_x(C)$ reaches its minimum). Intuitively, a minimal sufficient statistic captures all structural properties and no random properties. This definition was proposed by Kolmogorov, and further formalized by [39] to separate structural and random properties of the data.

We now want to show that, in general, a minimal sufficient statistic will always be trivial if the class H and the decoder function are expressive enough. Hence, picking a canonical class of hypotheses/functions (e.g., all possible functions, all computable functions) always lead to trivial structure and restriction to a smaller non-canonical set (all linear functions, descriptions that are English sentences, etc.) is necessary to talk about structural properties.

Proof. Let H be an hypothesis class. Consider the function

$$\lambda_x(C) = \min_{h \in H} \; -\log p_h(x) - \log p_{\text{code}}(h) \qquad (10)$$
$$\text{s.t.} \; -\log p_{\text{code}}(h) \leq C,$$

The best compression we can achieve using H is given by the minimum over:

$$\min_{h \in H} -\log p_h(x) - \log p_{\text{code}}(h).$$

Suppose however that, because the model class is rich enough, there is an hypothesis h_{search} which can compute:

$$\log p_{h_{\text{search}}}(x) = \min_{h \in H} -\log p_h(x) - \log p_{\text{code}}(h).$$

This can be implemented, for example, by enumerating all elements of H by their coding length $-\log p_{\text{code}}(h)$ and testing all of them until a minimum is found (only finitely many have to be tested, since h whose coding length is too long cannot be minima). Then h_{search} would be a minimal sufficient statistic *at the same* time for all possible samples, hence all samples really have the same structure. Moreover, since all samples would consider h_{search} an optimal description, their conceptual distance would be zero. □

This results motivate our non-canonical choice of restricting to $H = \{\text{natural language sentences}\}$ in defining a conceptual distance. We also note that when H is the class of all possible programs, h_{search} may not be computable due to the halting problem. The sketch of the proof however remains valid for most samples, and in particular all the ones likely to occur as real-world measurements (see [13] for an extended discussion).

D. Connection with Liu et al. [28]

Let x_1 and x_2 be two sentences, and let $p(h|x_i)$ be the distribution over the trajectories h that can extend x_i, where

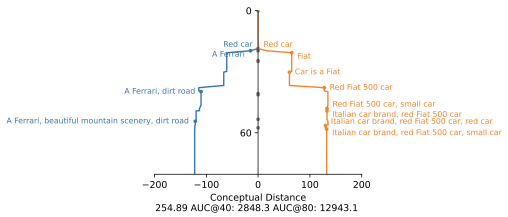

Figure 6. Same plot as Figure 1 but using Stable Diffusion to compute $p(x|h)$.

$p(h|x_i)$ is computed using a large language model. [28] proposes to measures the similarity between x_1 and x_2 based on the similarity between the distribution of trajectories they can generate. Specifically,

$$d_{\text{traj}}(x_1, x_2) = \mathbb{E}_{h\sim\frac{1}{2}(p(h|x_1)+p(h|x_2))}\big|\log p(h|x_1) - \log p(h|x_2)\big|$$

Interestingly, while this perspective is very different from our definition of conceptual distance, we note that it can be derived as a particular case of our method. In particular, consider using the set H of trajectories as the set of sampled descriptions used to compute the conceptual distance. When using the encoder-only method to compute our distance we have:

$$q^*_{x_i}(h|C) = \frac{1}{Z_\lambda} p_{\text{code}}(h)\Big(\frac{p(h|x_i)}{p(h)}\Big)^\lambda, \tag{11}$$

In the case of language, we can assume that the marginal probability $p(h) = \int p(h|x)p(x)dx$ of the trajectory h over all possible prefixes x should be similar to unconditional likelihood $p_{\text{code}}(h)$ of the text. With this assumption, and in the particular case of $\lambda(C) = 1$, the optimal distribution of descriptions reduces to:

$$q^*_{x_i}(h|\lambda = 1) = p(h|x_i).$$

Using this distribution to compute the conceptual distance we have:

$$\begin{aligned} d_{x_1,x_2}(\lambda = 1) = &\mathbb{E}_{p(h|x_1)}[\log p(h|x_2) - \log p(h|x_1)] \\ &+ \mathbb{E}_{p(h|x_2)}[\log p(h|x_1) - \log p(h|x_2)] \end{aligned}$$

assuming that $\log p(h|x_2) - \log p(h|x_1)$ is mostly positive when h is sampled from $p(h|x_1)$ — and vice versa for x_2 — we can rewrite this with absolute values as:

$$\begin{aligned} d_{x_1,x_2}(\lambda = 1) = &\ \mathbb{E}_{p(h|x_1)}|\log p(h|x_2) - \log p(h|x_1)| \\ &+ \mathbb{E}_{p(h|x_2)}|\log p(h|x_1) - \log p(h|x_2)| \\ = &\ 2\cdot\mathbb{E}_{\frac{1}{2}p(h|x_1)+\frac{1}{2}p(h|x_2)}\big|\log p(h|x_1) - \log p(h|x_2)\big|. \end{aligned}$$

Hence we can see $d_{\text{traj}}(x_1, x_2)$ as a particular case of our conceptual distance when using a particular capacity C such that $\lambda(C) = 1$, and making a particular choice of using trajectories as descriptions of the sentences x_1 and x_2. Our results in Tabs. 1 and 3 show that using our distance without these restrictions outperforms [28] when evaluated in the same setting.

E. Using Only the Encoder Model

As described in Sec. 5, for image experiments we use the LLaVA image-to-text encoder $p(h|x)$ to evaluate the likelihood $p(x|h)$, through Bayes' rule:

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

We now want to show that the term $p(x)$ does not affect the distance computation, and can be ignored. First note using the above identity, we have that the reconstruction loss is

$$\ell(x|h) = -\log\frac{p(h|x)}{p(h)} - \log p(x) = \bar{\ell}(x|h) - \log p(x),$$

where we defined $\bar{\ell}(x|h) = -\log\frac{p(h|x)}{p(h)}$. Considering the optimization problem in Eq. (2) to find the optimal description under a capacity constrain:

$$\begin{aligned} q^*_x(h|C) = \underset{q(h)\in\mathcal{P}(H)}{\arg\min}\ &\mathbb{E}_{h\sim q(h)}[\ell(x|h)] \\ \text{s.t. }\ &\text{KL}\left(q(h)\,\|\,p(h)\right) \le C, \end{aligned}$$

we see that $-\log p(x)$ (which does not depend on h) only accounts for an additive constant ($\mathbb{E}_{h\sim q(h)}[\ell(x|h)] = \mathbb{E}_{h\sim q(h)}[-\log\frac{p(h|x)}{p(h)}] + \log p(x)$). Hence $q^*_x(h|C)$ does not depend on $p(x)$. Using Eq. (5), the distance is:

$$\begin{aligned} d_{x_1,x_2}(C) &= \mathbb{E}_{q_1}[\ell(x_1|h)] + \mathbb{E}_{q_2}[\ell(x_2|h)] - \mathbb{E}_{q_\cap}[\ell(x_1|h) + \ell(x_2|h)] \\ &= \mathbb{E}_{q_1}[\bar{\ell}(x_1|h)] + \log p(x_1) + \mathbb{E}_{q_2}[\bar{\ell}(x_2|h)] + \log p(x_2) \\ &\quad - \mathbb{E}_{q_\cap}[\bar{\ell}(x_1|h) + \bar{\ell}(x_2|h)] - (\log p(x_1) + \log p(x_2)) \\ &= \mathbb{E}_{q_1}[\bar{\ell}(x_1|h)] + \mathbb{E}_{q_2}[\bar{\ell}(x_2|h)] - \mathbb{E}_{q_\cap}[\bar{\ell}(x_1|h) + \bar{\ell}(x_2|h)], \end{aligned}$$

and all quantities in the last expression are independent of the value of $\log p(x_i)$.

F. Closed-Form Expressions

We now derive the close form expression for the solution of Eq. (2):

$$\begin{aligned} q^*_x(h|C) = \underset{q(h)\in\mathcal{P}(H)}{\arg\min}\ &\mathbb{E}_{h\sim q(h)}[\ell(x|h)] \\ \text{s.t. }\ &\text{KL}\left(q(h)\,\|\,p_{\text{code}}(h)\right) \le C. \end{aligned}$$

Writing the Lagrangian corresponding to the constrained optimization problem we have:

$$\begin{aligned} \mathcal{L} = &\mathbb{E}_{h\sim q(h)}[\ell(x|h)] + \gamma\big(\text{KL}\left(q(h)\,\|\,p_{\text{code}}(h)\right) - C\big) \\ &+ \alpha\big(\int q(h)dh - 1\big) \end{aligned}$$

Setting to zero the derivatives of $\mathcal{L}$ with respect to $q(h)$ we have:

$$\partial_{q(h)}\mathcal{L} = l(x|h) + \gamma\Big(\log\frac{q(h)}{p_{\text{code}}(h)} - 1\Big) + \alpha = 0,$$

from which we get:

$$\begin{aligned} q(h) &= p_{\text{code}}(h)\exp\Big(-\frac{1}{\gamma}l(x|h) + 1 - \alpha\Big) \\ &= \frac{1}{Z}p_{\text{code}}(h)\exp\big(-\lambda l(x|h)\big) \\ &= \frac{1}{Z}p_{\text{code}}(h)p(x|h)^{\lambda}, \end{aligned}$$

where we defined $\lambda := \frac{1}{\gamma}$ and $Z := \exp(\alpha - 1)$, and we used $l(x|h) = -\log p(x|h)$. Enforcing the constraint that $q(h)$ is a probability distribution and integrates to one (note that $q(h) > 0$ is automatically satisfied), we get:

$$Z = Z_\lambda = \int p_{\text{code}}(h)p(x|h)^{\lambda}dh.$$

Enforcing the remaining constraint on $\lambda = \lambda(C)$ allows finding the optimal distribution for the particular capacity C. While the solution cannot be written analytically, it can be found easily by binary search over λ. Doing it is however not necessary for our method: since our method only cares about the function

$$\beta_x(C) = \mathbb{E}_{q^*(h|C)}[\ell(x|h)]$$

as C varies, rather than solving the constraint we can simply trace the curve

$$\begin{aligned} \beta_x(\lambda) &= \mathbb{E}_{q^*(h|\lambda)}[\ell(x|h)] \\ C(\lambda) &= \mathrm{KL}\left(q^*(h|\lambda)\,\|\,p_{\text{code}}(h)\right) \end{aligned}$$

as λ varies in order to reconstruct the function $\beta_x(C)$. In our experiments, we sample $\lambda \in \mathrm{linspace}(0, 100, 200)$ and linearly interpolate the results to approximate $\beta_x(C)$.

G. Experimental Details for Qualitative Plots

To improve interpretability of the plot, and aid better understanding of the key components of the method, we use the following setup for our qualitative plots. We use as coding length $\ell(h) = -\log p_{\text{code}}(h)$ the log-likelihood assigned to h by a LLaMA language-model. This scales with the complexity and length of the sentence h. In Eq. (2) the complexity C of the hypothesis distribution $q(h)$ is given by $\mathrm{KL}\left(q(h)\,\|\,p(h)\right)$ which measures closeness of the distribution to the prior $p_{\text{code}}(h)$. This means that, when C is low, the method will select all sentences, giving higher probability to short ones. Since distribution of sentences are difficult to visualize, we force $q(h)$ to be a Dirac delta, thus making it more similar to Eq. (1) and effectively selecting the single best description with coding length $\ell(h) \leq C$.

To observe the effect of varying C, it is helpful to sample descriptions that are increasingly longer (larger $\ell(h)$) and more descriptive (smaller $\ell(x|h)$), as we expect longer descriptions to be preferentially picked as C increases. Unfortunately, directly sampling longer descriptions with LLaVA does not result in good descriptions, as the model starts hallucinating information that is irrelevant for the image and thus does not increase $\ell(x|h)$. To remedy this, we use the following beam search method. First, we prompt the LLaVA model to generate short "atoms" of information about the image (using the prompt: "`Describe in 10 short bullet points what you see in the image. Do not provide explanations.`"). We compute a total of 40 atoms for each image of the pair, and combine them in a common dictionary. We also sample 40 atoms that are good descriptions of both images at the same time, by sampling tokens from the ensemble $\frac{1}{2}\log p(h_t|h_{<t}, x_1) + \frac{1}{2}\log p(h_t|h_{<t}, x_2)$ of the log-likelihoods conditioned on each image. We then use a beam search to combine the atoms in the dictionary in increasingly longer sentences that are optimal description for each image — i.e., minimize $\ell(x|h)$. Since evaluating $\ell(x|h)$ with LLaVA at each step of the beam search is expensive, we use the CLIP similarity between h and x as a proxy score. For Fig. 3, we add to the beam search a penalty to preferentially avoid sentences that do not relate to the prompt ("Describe style" or "Describing content") by subtracting from the CLIP score the CLIP similarity between the sentence and the negative prompt. This procedure is used to sample a qualitatively interesting set H of descriptions which helps providing a better intuitive understanding of the conceptual distance. The distance computation is otherwise unaltered. In the quantitative results we instead simply sample $H \sim \frac{1}{2}p(h|x_1) + \frac{1}{2}p(h|x_2)$.