CLIP-like Model as a Foundational Density Ratio Estimator

Supplementary Material

9. Derivations of Eqs. (1) and (2)

In SigLIP, the learning objective is to classify whether a given pair of image and text is a positive pair or a negative pair. Given an image i, the text paired with i could be considered as a sample from the conditional text distribution $p_T(\cdot|i)$, while the negative texts are sampled from the marginal text distribution $p_T(\cdot)$.

Fixing i, let's consider a set that contains one positive sample text paired with i, and ν other texts that are not paired with i. We assign binary labels C where $C = 1$ for the positive sample and $C = 0$ for the negative samples. Therefore,

$$P(t|C = 1; i) = p_T(t|i) \tag{20}$$
$$P(t|C = 0; i) = p_T(t). \tag{21}$$

The prior probabilities of class labels are $P(C = 1) = \frac{1}{1+\nu}$ and $P(C = 0) = \frac{\nu}{1+\nu}$. The posterior probability of positive sample is

$$\begin{aligned} &P(C = 1|t; i) \\ &= \frac{P(t|C = 1; i)P(C = 1)}{P(t|C = 1; i)P(C = 1) + P(t|C = 0; i)P(C = 0)} \\ &= \frac{p_T(t|i)}{p_T(t|i) + \nu p_T(t)}. \end{aligned} \tag{22}$$

Also, the posterior probability of negative sample $P(C = 0|t; i)$ is $\frac{\nu p_T(t)}{p_T(t|i)+\nu p_T(t)}$. Denoting the positive sample by $t^{\mathrm{pos}(i)}$, and the set of the negative samples by $\mathrm{neg}(i)$, the log-likelihood $L_C(i)$ on a set about i is

$$\begin{aligned} &L_C(i) \\ &= \log\left[P(C = 1|t^{\mathrm{pos}(i)}; i) \cdot \prod_{t'\in\mathrm{neg}(i)} P(C = 0|t'; i)\right] \\ &= \log\left[\frac{p_T(t^{\mathrm{pos}(i)}|i)}{p_T(t^{\mathrm{pos}(i)}|i) + \nu p_T(t^{\mathrm{pos}(i)})}\right] \\ &\quad + \sum_{t'\in\mathrm{neg}(i)} \log\left[\frac{\nu p_T(t')}{p_T(t'|i) + \nu p_T(t')}\right] \\ &= \log\left[\frac{p_T(t^{\mathrm{p}(i)}|i)}{p_T(t^{\mathrm{p}(i)}|i) + \nu p_T(t^{\mathrm{p}(i)})}\right] \\ &\quad + \sum_{t'\in\mathrm{neg}(i)} \log\left[1 - \frac{p_T(t'|i)}{p_T(t'|i) + \nu p_T(t')}\right]. \end{aligned} \tag{23}$$

On the other hand, a batch of SigLIP has $\nu + 1$ elements, where the jth element is a text set composed of one text paired with an image i_j and ν other texts, which is the same configuration as the previous one. The SigLIP loss is

$$\begin{aligned} L_{\mathrm{SigLIP}} &= -\frac{1}{N}\sum_{j=1}^{N}\left[\log\left[\sigma(s(t^{\mathrm{pos}(i_j)}, i_j))\right]\right. \qquad (24)\\ &\quad \left. + \sum_{t'\in\mathrm{neg}(i_j)} \log[\sigma(-s(t', i_j))]\right] \\ &= -\frac{1}{N}\sum_{j=1}^{N}\left[\log\left[\sigma(s(t^{\mathrm{pos}(i_j)}, i_j))\right]\right. \qquad (25)\\ &\quad \left. + \sum_{t'\in\mathrm{neg}(i_j)} \log[1 - \sigma(s(t', i_j))]\right] \qquad (26) \end{aligned}$$

where $N := \nu + 1$ and $s(t, i) := a\langle v_t, v_i\rangle + b$ that b is a scalar called logit bias. $\sigma(x) := \frac{1}{1+\exp(-x)}$ is sigmoid function. The last equation holds because $\sigma(-x) = 1 - \sigma(x)$.

Compared with Eq. (23) and Eq. (26), Minimizing L_{NCE} is equal to maximizing summation of $L(i)$ w.r.t. images and the optimal embeddings $\hat{v}_t, \hat{v}_i$ satisfy following equation:

$$\frac{p_T(t|i)}{p_T(t|i) + \nu p_T(t)} = \frac{1}{1 + \exp(-a\langle \hat{v}_t, \hat{v}_i\rangle - b)} \tag{27}$$

when assuming a one-to-one mapping from text to embedding. With some equivalent transformation, we obtain

$$\frac{p_T(t|i)}{p_T(t)} = \nu e^b \exp(a\langle v_t, v_i\rangle). \tag{28}$$

ν and b are constant through samples, and $\int p_T(t|i)dt = \int p_T(t)\exp(a\langle v_t, v_i\rangle)/Z(i)dt = 1$, thus Eq. (1) holds and $Z(i)^{-1} = \nu e^b$ in an ideal modeling.

A batch of SigLIP is also regarded as a batch whose j-th element is an image set composed of one image paired with a text t_j and ν other images, since the batch is made from N image-text pairs and negative samples are prepared from other pairs. So Eq. (2) holds symmetrically.

While SigLIP is learning binary classification, CLIP is based on multi-label classification in a batch. Let's assign an index for each sample instead of a binary label in a text set, and assume that the positive text paired with an image i_j exists at the j-th position in a text set. We denote the j-th sample text by t_j for $1 \le j \le N$. This configuration makes a multi-label classification problem, which predicts

the index of the text that corresponds to the input image. The log-likelihood for image-to-text prediction is therefore:

$$\begin{aligned} L_D(i_j) &= \log\left[P(D=j|t_1,\dots,t_N;i_j)\right] \\ &= \log\left[\frac{p_T(t_j|i_j)\prod_{l\neq j}p_T(t_l)}{\sum_{k=1}^N p_T(t_k|i_j)\prod_{l\neq k}p_T(t_l)}\right] \\ &= \log\left[\frac{\frac{p_T(t_j|i_j)}{p_T(t_j)}}{\sum_{k=1}^N \frac{p_T(t_k|i_j)}{p_T(t_j)}}\right]. \end{aligned} \tag{29}$$

In the same way, the log-likelihood of text-to-image prediction is

$$\begin{aligned} L_D(t_j) &= \log\left[P(D=j|i_1,\dots,i_N;t_j)\right] \\ &= \log\left[\frac{\frac{p_I(i_j|t_j)}{p_I(i_j)}}{\sum_{k=1}^N \frac{p_I(i_k|t_j)}{p_I(i_j)}}\right]. \end{aligned} \tag{30}$$

CLIP loss functions are designed for maximizing the log-likelihood of image-to-text prediction and text-to-image prediction as follows:

$$\begin{aligned} L_{\text{CLIP}} = &-\sum_{j=1}^N \underbrace{\log\left[\frac{\exp(s(t_j,i_j))}{\sum_{k=1}^N \exp(s(t_k,i_j))}\right]}_{\text{image-to-text prediction}} \\ &-\sum_{j=1}^N \underbrace{\log\left[\frac{\exp(s(t_j,i_j))}{\sum_{k=1}^N \exp(s(t_j,i_k))}\right]}_{\text{text-to-image prediction}} \end{aligned} \tag{31}$$

where $s(t,i) := a\langle v_t, v_i\rangle$ is the score of CLIP that does not include logit bias b. The optimal score $\hat{s}$ is proportional to the density ratio of conditional and marginal probabilities for both image and text modality:

$$\hat{s}(t,i) \propto \frac{p_T(t|i)}{p_T(t)} = \frac{p_I(i|t)}{p_I(i)}. \tag{32}$$

10. Derivations of Eqs. (10) and (11)

To treat images and text in the same manner, we abuse p_T like $p_T(t) = p_T(v_t)$ and $p_T(t|i) = p_T(v_t|v_i)$.

By substituting Equation (1) into Equation (8),

$$\begin{aligned} D_{\text{KL}}(i) &= \int p_T(v_{t'}|v_i)\log\frac{p_T(v_{t'}|v_i)}{p_T(v_{t'})}dv_{t'} \\ &= \int p_T(v_{t'})\frac{\exp(a\langle v_{t'},v_i\rangle)}{Z(i)}\log\frac{\exp(a\langle v_{t'},v_i\rangle)}{Z(i)}dv_{t'} \\ &= \frac{1}{Z(i)}\int p_T(v_{t'})\exp(a\langle v_{t'},v_i\rangle)a\langle v_{t'},v_i\rangle dv_{t'} \\ &\quad - \frac{1}{Z(i)}\log(Z(i))\int p_T(v_{t'})\exp(a\langle v_{t'},v_i\rangle)dv_{t'} \\ &= \frac{1}{Z(i)}\mathbb{E}_{t'\sim p_T(\cdot)}\left[a\langle v_{t'},v_i\rangle e^{a\langle v_{t'},v_i\rangle}\right] - \log Z(i). \end{aligned} \tag{33}$$

Note that $Z(i) = \mathbb{E}_{t\sim p_T(\cdot)}[e^{a\langle v_t,v_i\rangle}]$. Eq. (33) can be estimated by text samples $\mathcal{D}_T$ as

$$\begin{aligned} D_{\text{KL}}(i) &= \frac{\mathbb{E}_{t'\sim p_T(\cdot)}\left[a\langle v_{t'},v_i\rangle e^{a\langle v_{t'},v_i\rangle}\right]}{\mathbb{E}_{t\sim p_T(\cdot)}[e^{a\langle v_t,v_i\rangle}]} \\ &\quad - \log\mathbb{E}_{t'\sim p_T(\cdot)}\left[a\langle v_{t'},v_i\rangle e^{a\langle v_{t'},v_i\rangle}\right] \\ &\approx \frac{\sum_{t'\in\mathcal{D}_T} a\langle v_{t'},v_i\rangle e^{a\langle v_{t'},v_i\rangle}}{\sum_{t\in\mathcal{D}_T} e^{a\langle v_t,v_i\rangle}} \\ &\quad - \log\frac{1}{|\mathcal{D}_T|}\sum_{t'\in\mathcal{D}_T} a\langle v_{t'},v_i\rangle e^{a\langle v_{t'},v_i\rangle} \\ &= \sum_{t'\in\mathcal{D}_T} a\langle v_{t'},v_i\rangle \text{softmax}_i(t') \\ &\quad - \log\frac{1}{|\mathcal{D}_T|}\sum_{t'\in\mathcal{D}_T} a\langle v_{t'},v_i\rangle e^{a\langle v_{t'},v_i\rangle} \\ &= \sum_{t\in\mathcal{D}_T} a\langle v_t,v_i\rangle \text{softmax}_i(t) \\ &\quad - \log\sum_{t\in\mathcal{D}_T} a\langle v_t,v_i\rangle e^{a\langle v_t,v_i\rangle} + \log|\mathcal{D}_T| \end{aligned} \tag{34}$$

where $\text{softmax}_i(t) := \frac{e^{a\langle v_t,v_i\rangle}}{\sum_{t'\in\mathcal{D}_T} e^{a\langle v_{t'},v_i\rangle}}$.

Equation (9) is also transformed with Eq. (1) as follows:

$$\begin{aligned} D_{\text{KLR}}(i) &= \int p_T(v_{t'})\log\frac{p_T(v_{t'})}{p_T(v_{t'}|v_i)}dv_{t'} \\ &= \int p_T(v_{t'})\log\frac{Z(i)}{\exp(a\langle v_{t'},v_i\rangle)}dv_{t'} \\ &= \log Z(i)\int p_T(v_{t'})dv_{t'} \\ &\quad - \int p_T(v_{t'})a\langle v_{t'},v_i\rangle dv_{t'} \\ &= \log Z(i) - \mathbb{E}_{t'\sim p_T(\cdot)}\left[a\langle v_{t'},v_i\rangle\right] \end{aligned} \tag{35}$$

With test samples $\mathcal{D}_T$,

$$D_{\mathrm{KLR}}(i) \approx \log \frac{1}{|\mathcal{D}_T|} \sum_{t \in \mathcal{D}_T} e^{a\langle v_t, v_i \rangle} - \frac{1}{|\mathcal{D}_T|} \sum_{t \in \mathcal{D}_T} a\langle v_t, v_i \rangle$$
$$= \log \sum_{t \in \mathcal{D}_T} e^{a\langle v_t, v_i \rangle} - \log |\mathcal{D}_T|$$
$$- \frac{1}{|\mathcal{D}_T|} \sum_{t \in \mathcal{D}_T} a\langle v_t, v_i \rangle \tag{36}$$

11. Hyper-parameters for Sec. 4

We trained ViT-B/32 models on the CC12M [5] dataset using the official OpenCLIP training pipeline. Unless otherwise noted, we follow the default OpenCLIP configuration. We use AdamW with a learning rate of 5×10^{-4}, weight decay of 0.2, and mixed-precision training (AMP). Each training run uses a single node equipped with eight GPUs with a per-GPU batch size of 1000 and eight dataloader workers, resulting in a global batch size of 8000. Models are trained for ten epochs. We enable Importance-Weighted Learning (IWL) using a publicly available ViT-L/14 model pretrained on LAION2B [15] as the reference model. The end-to-end training cost is approximately 70 H200 GPU hours per model.

12. Other results of Sec. 5.3

Figures 5 to 7 show top and bottom 18 images ranked by D_{KLR}, D_{C}, D_{W}. Some bottom samples of D_{C} have shared concepts such as patterned animals, and the top samples of D_{W} resemble the bottom samples of D_{KL}. However, unlike D_{KL}, qualitative differences between top and bottom samples are less distinct. Figures 8 to 10 are the text samples. D_{KLR} and D_{W} capture meaningless captions at their bottom score, while D_{C} seems to represent specificity for image content. Figure 11 is the cumulative N-gram probability of captions grouped by each KL divergence. Interestingly, captions associated with low D_{W} show more diverse N-gram distribution. Further analysis of those metrics is left for future work.

13. SigLIP Results

In Secs. 4 and 5, we empirically examined the applications of CLIP-like models as density ratio estimators using CLIP, which is commonly used in various tasks. In this section, we report the result of Importance-Weighted Learning and KL divergence estimation using ViT-B-16 SigLIP pretrained on the webli dataset [41].

Importance-Weighted Learning Figure 12 shows the downstream task performance for each domain compared with no importance weighting. Although the performance of IWL does not explicitly surpass training without importance weighting on Food101, our method mostly demonstrates higher downstream task performance than the baseline consistent with the results observed with CLIP.

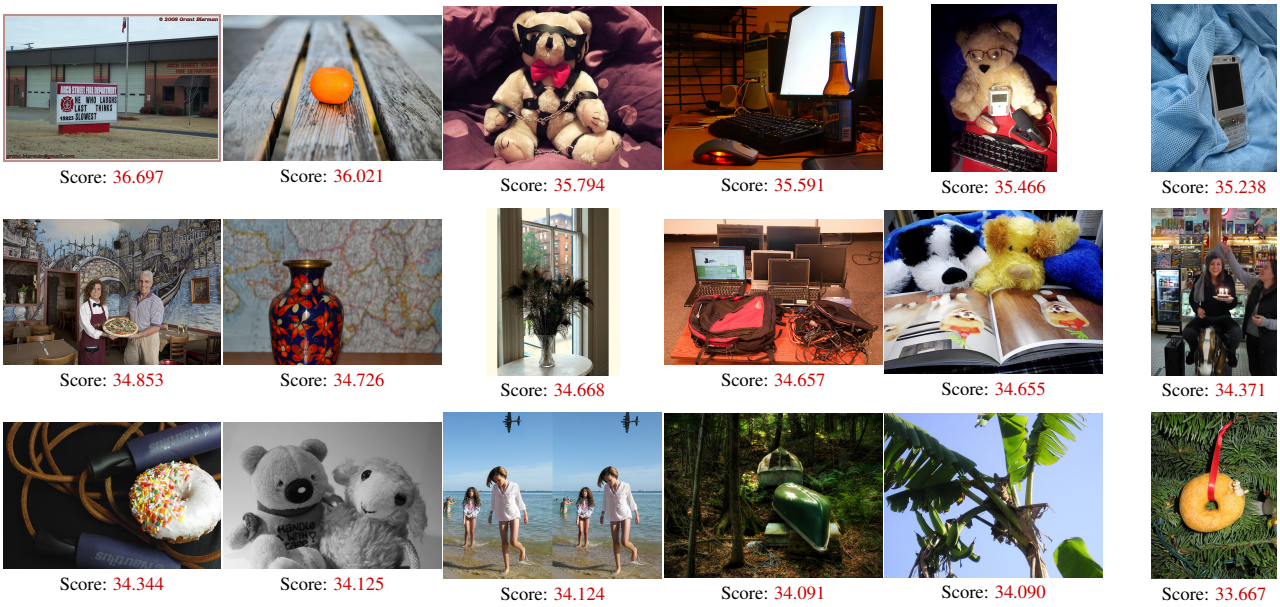
K	d	CLIP			SigLIP		
		$R^2\uparrow$	MSE↓	Pearson↑	$R^2\uparrow$	MSE↓	Pearson↑
8	2	0.9938	4.46e-01	0.9971	0.7722	1.63e+01	0.9457
	8	1.0000	1.05e-02	1.0000	0.9250	1.51e+02	0.9978
	64	1.0000	9.60e-04	1.0000	0.9723	4.84e+01	0.9983
	128	1.0000	1.03e-03	1.0000	0.9924	1.12e+01	0.9998
	512	1.0000	5.35e-03	1.0000	0.9823	4.04e+01	0.9997
16	2	0.9947	4.25e-01	0.9974	0.9250	6.02e+00	0.9847
	8	0.9998	9.53e-01	0.9999	0.1484	4.05e+03	0.3980
	64	0.9998	3.30e+00	1.0000	0.0927	1.33e+04	0.3595
	128	1.0000	2.23e-01	1.0000	0.7778	1.62e+03	0.9589
	512	1.0000	9.38e-02	1.0000	0.9763	1.11e+02	0.9991

Table 3. Density ratio estimation accuracy (R^2, MSE, Pearson) on synthetic data. K is the number of the artificial labels, and d is the dimension of image space. Each metric is calculated using density ratios in test data.

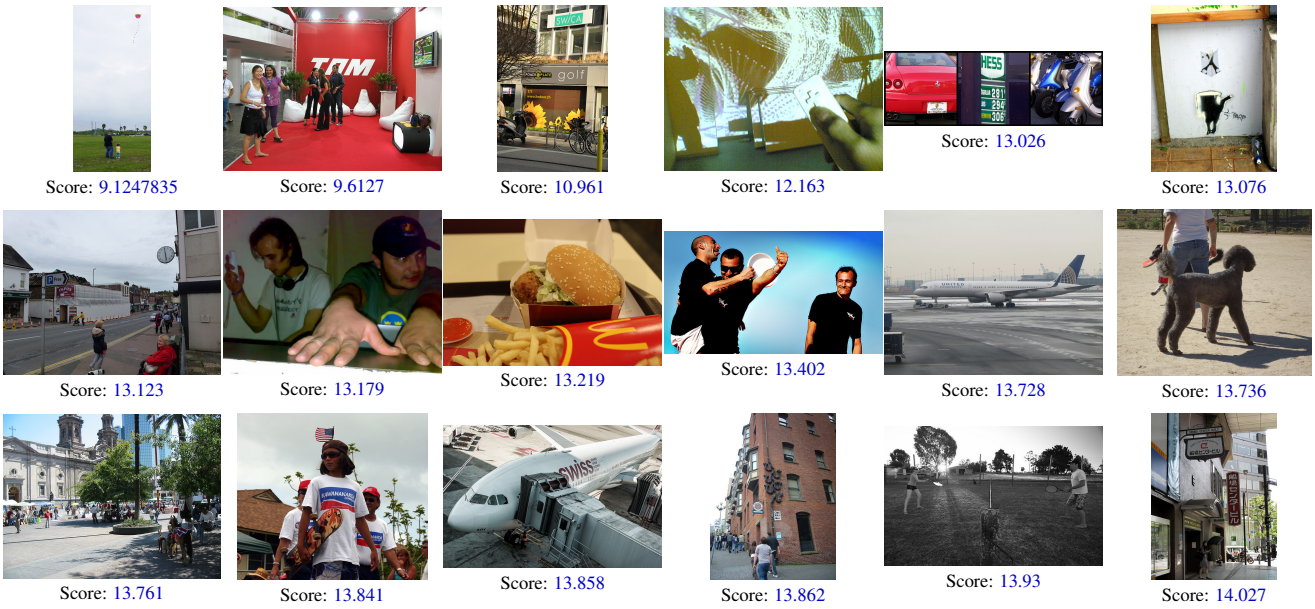
KL Divergence Estimation Figures 13 to 16 show top and bottom 18 images ranked by the four KL-based metrics D_{KL}, D_{KLR}, D_{C}, D_{W} and Figs. 17 to 20 are the text samples. Figure 21 is the cumulative N-gram probability of captions grouped by the four metrics estimated by SigLIP. Consistent with CLIP's results, D_{KL} captures semantic diversity in terms of N-gram frequency.

14. Ground-truth Experiments in a Toy Setting

In the main part of the paper, we have verified the density ratio modeling of CLIP-like models by implementing applications based on density ratio. On the other hand, we also conducted small-scale experiments to directly validate the density ratio modeling of the CLIP-like models. For $K \in \mathbb{N}$, we prepared eight artificial class labels $t \in \{t_j\}_{j=1}^{K}$ sampled from a categorical distribution. This corresponds to captions in CLIP-like models. Then, an image $i \in \mathbb{R}^d (d \in \mathbb{N})$ sampled from a Gaussian mixture model; each peak is a Gaussian distribution $\mathcal{N}(\mu_j, 4I)$ and $\mu_j \in \mathbb{R}^d$ corresponds to t_j. We trained lightweight CLIP-like encoders, implemented as three-layer MLPs, using the InfoNCE and NCE objectives. As shown in Figure 22, the predicted density ratios in 2D image space with eight labels closely match and have high correlation coefficients with the ground-truth ratio. Moreover, Tab. 3 demonstrates that CLIP achieves almost perfect estimation across all settings, while SigLIP shows varying accuracy depending on dimensionality. This validates our theoretical interpretation that the contrastive objectives learn density ratios, complementing the real-data applications in the main paper.



(a) Top 18 images



(b) Bottom 18 images

Figure 5. Top and bottom image examples in MSCOCO captions ranked by D_{KLR} through cosine similarity of CLIP.

15. Estimation Error Analysis

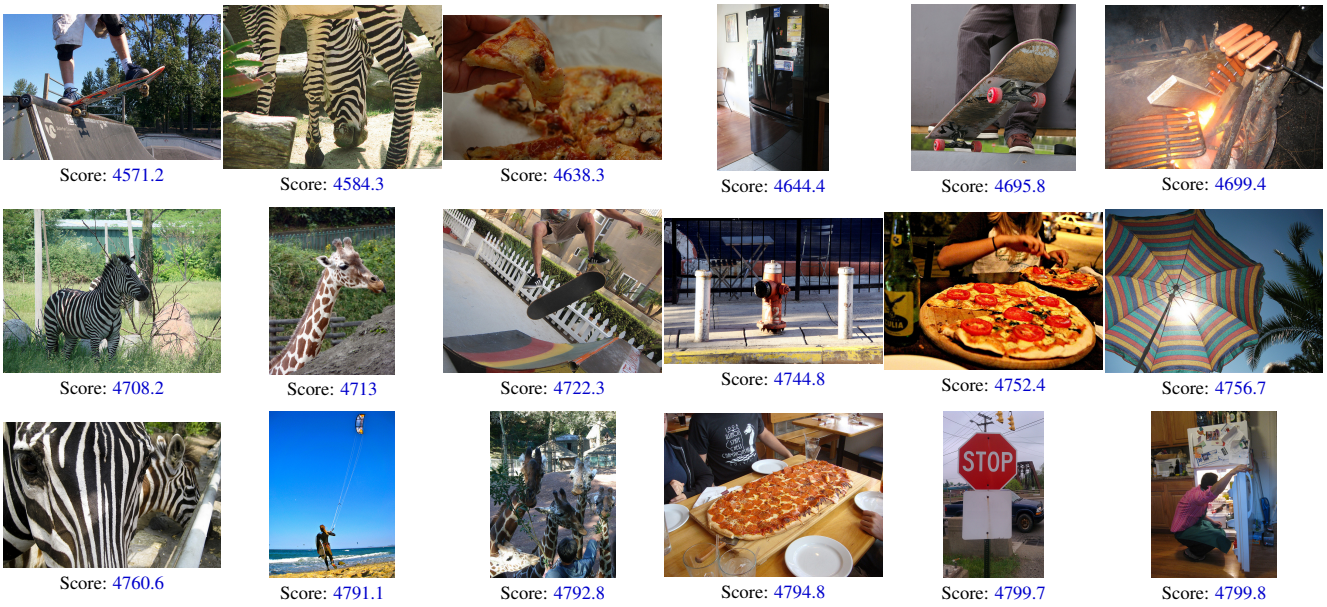
All approximations of the KL divergences need a sample set. Let us consider estimates of the image KL divergence. Right sides of Eqs. (10) and (11) employ sample text set $\mathcal{D}_T$, and Eqs. (18) and (19) use both text set and image set for calculating $\hat{G}_T$, $\hat{G}_I$, $\hat{v}_T$ and $\hat{v}_I$. We evaluated approximation bias and its variance arising from finite-sample approximation using the bootstrap method.

15.1. Bias, Variance and RMSE

To analyze the error of finite sample approximation, we prepared B sample sets of texts and images $\{\mathcal{D}_T^b\}_{b=1}^B, \{\mathcal{D}_I^b\}_{b=1}^B$ where each $\mathcal{D}_T^b$ and $\mathcal{D}_I^b$ are sampled from the original $\mathcal{D}_T$ and $\mathcal{D}_I$ with replacement.

Let $\hat{D}_{\mathrm{KL}}(i;b), \hat{D}_{\mathrm{KLR}}(i;b)$ denote the estimated value of $D_{\mathrm{KL}}(i), D_{\mathrm{KLR}}(i)$ using $\mathcal{D}_T^b$. We also denote the estimated value of $D_{\mathrm{C}}(i), D_{\mathrm{W}}(i)$ using $\mathcal{D}_T^b$ and $\mathcal{D}_I^b$ as

(a) Top 18 images



(b) Bottom 18 images

Figure 6. Top and bottom image examples in MSCOCO captions ranked by D_{C} through cosine similarity of CLIP.

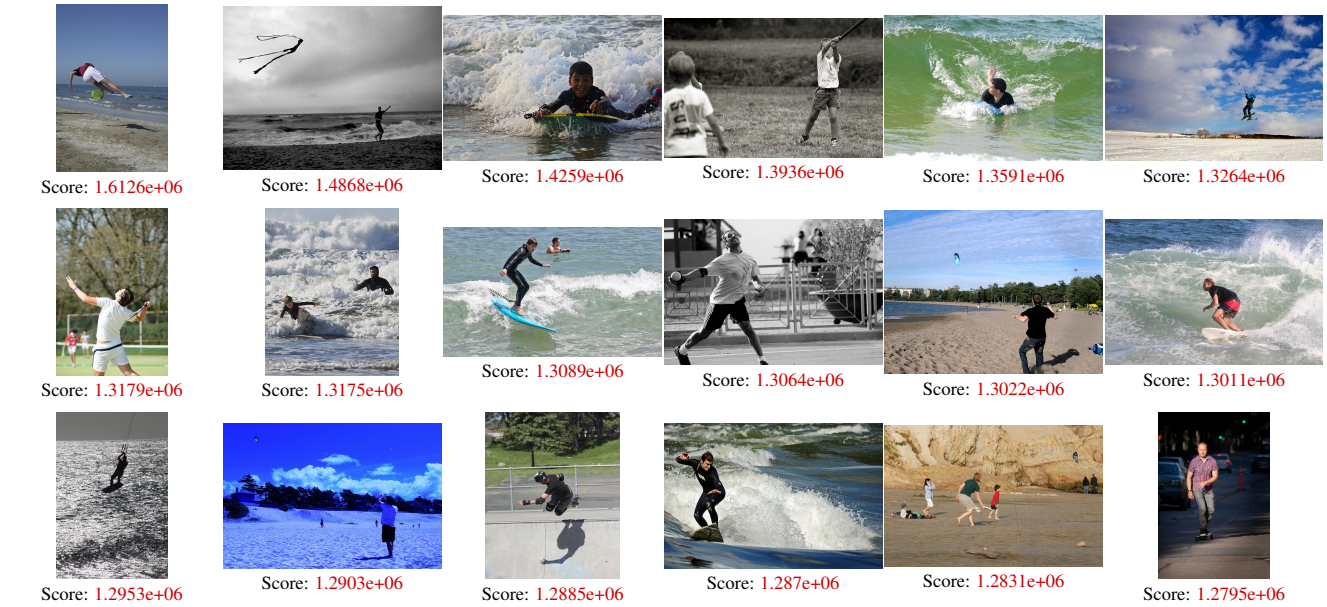
$\hat{D}_{\mathrm{C}}(i;b), \hat{D}_{\mathrm{W}}(i;b)$ We collectively denote these estimations using original samples $\mathcal{D}_T$ (and $\mathcal{D}_I$) as $\hat{D}_*(i)$, and these estimations using $\mathcal{D}_T^b$ (and $\mathcal{D}_I^b$) as $\hat{D}_*(i;b)$.

For a fixed image i, the estimation bias and variance can be calculated as

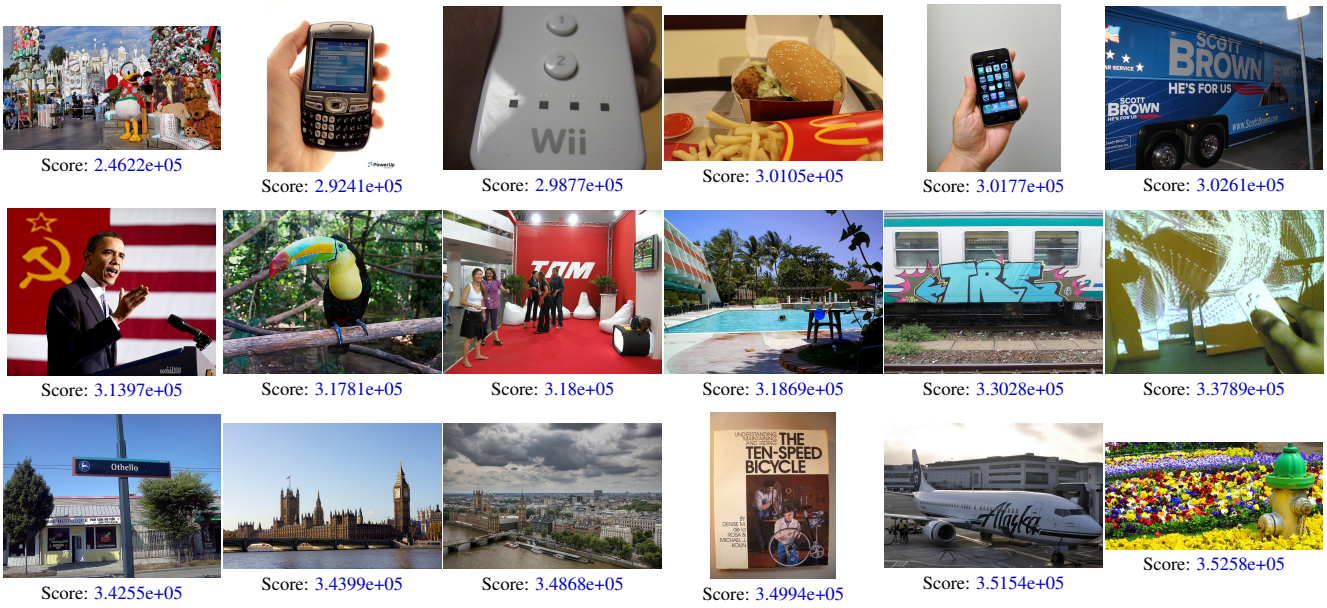
$$\mathrm{Bias}(i) = \frac{1}{B}\sum_{b=1}^{B}\hat{D}_*(i;b) - \hat{D}_*(i) \tag{37}$$

$$\mathrm{Variance}(i) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{D}_*(i;b) - \bar{D}_*(i)\right)^2 \tag{38}$$

where $\bar{D}_*(i) := \frac{1}{B}\sum_{b=1}^{B}\hat{D}_*(i;b)$. Note that Variance(i)



(a) Top 18 images



(b) Bottom 18 images

Figure 7. Top and bottom image examples in MSCOCO captions ranked by D_{W} through cosine similarity of CLIP.

is not the variance of the KL divergence across images, but across sample sets.

Root Mean Square Error (RMSE) is then computed as follows:

$$\begin{aligned}
\mathrm{RMSE}^2(i) &= \frac{1}{B}\sum_{b=1}^{B}\left(\hat{D}_*(i;b) - \hat{D}_*(i)\right)^2 \\
&= \frac{1}{B}\sum_{b=1}^{B}\left(\hat{D}_*(i;b) - \bar{D}_*(i) + \bar{D}_*(i) - \hat{D}_*(i)\right)^2 \\
&= \mathrm{Variance}(i) + \mathrm{Bias}(i)^2 \\
\therefore \mathrm{RMSE}(i) &= \sqrt{\mathrm{Variance}(i) + \mathrm{Bias}(i)^2}. \qquad (39)
\end{aligned}$$

A waitress and restaurant owner showing off a pizza in front of a mural of Venice. Score: 38.591	A group photo of men and boys from the Goodmayes Boys School dated April 1929. Score: 37.113	A green street sign that says Bodacious Drive. Score: 36.416	Sign of Arch Street Fire Department saying "He who laughs last thinks slowest" Score: 36.341	A bottle of beer sits next to a computer at a desk Score: 36.214	A street sign reads "Bacchanalia" above a stop sign. Score: 36.101
a street sign for Library Way and Madison Avenue above a one way arrow Score: 35.955	A ripe tangerine atop a worn wooden picnic table. Score: 35.796	A small cat sitting in the top of a banana tree. Score: 35.741	A bottle of beer sits next to the keyboard and mouse at the computer table. Score: 35.619	A black and white photo of four elderly people sitting in a bench overlooking the ocean. Score: 35.522	A class picture of the Goodmayes Boys' School in April of 1929. Score: 35.281
Two almost identical photos with some minor cropping show two young girls standing on the coast with people in the water in the background as a plane flies overhead Score: 35.094	Assortment of laptop computers displayed on table with backpacks full of electronic cords. Score: 35.062	A picture of a road sign labeled bodacious drive. Score: 35.05	A vintage baby doll with the book "Goody two Shoes" on it's lap Score: 34.898	A red fire hydrant is sitting in the woods near fallen leaves. Score: 34.83	A computer on a desk with a bottle of beer next to it Score: 34.745

(a) Top 18 texts

There is no image to be reviewed on this hit. Score: 1.5582	unable to see this image in this particular hit Score: 1.7586	There is no image here to provide a caption for. Score: 1.8616	There is no image here to provide a caption for. Score: 1.8616	There is no image here to provide a caption for. Score: 1.8616	There is no image here to provide a caption for. Score: 1.8616
no image again there are a lot of these lately Score: 1.9015	There are some appealing support ready to expend. Score: 1.954	I am unable to see the image above. Score: 1.9708	I am not sure what this image is. Score: 1.9813	There is no image to describe for this question. Score: 2.0279	I do not know what this is supposed to be.. Score: 2.2007
I am unable to see an image above. Score: 2.5721	I am unable to see an image above. Score: 2.5721	I am unable to see an image above. Score: 2.5721	I am unable to see an image above. Score: 2.5721	I am unable to see an image above. Score: 2.5721	A being is doing something as of right now that is splendid. Score: 3.137

(b) Bottom 18 texts

Figure 8. Top and bottom captions ranked by D_{KLR}.

Side by side view of two oval plates, one with fork, with chicken salad sandwiches and rosy new potatoes, by an open and an unopened bottle of lager, a pepper mill, paper towel roll, basket behind. Score: 10591	A blue and black ray holding sandwich, coffee and liquor bottles. Score: 10313	Three rectangular bowls with food; Big bowl has nine meat and sesame seed patties with brown sauce, next to it, a bowl of shredded cabbage and carrots with yogurt dollop atop, and behind that is a bowl of cut broccoli and tomatoes with seasoning. Score: 10278	Buildings with large signs "Park Here" with rainy cement Score: 10138	Here is Massachusettes candidate Scott Brown's campaign trailer. Score: 10115	A kale and sweet pea home garden getting the last rays of sunlight. Score: 10037
Sign of Arch Street Fire Department saying "He who laughs last thinks slowest" Score: 10034	Four bowls of snacks: crackers, broccoli and carrots, nuts and dip Score: 9961.7	Bananas, marshmellows, chocolate chips and sprinkles in a bowl Score: 9957.1	Tube of orange and pears, surrounded by boxes of grass and one pineapple Score: 9946.3	Two apples, a bowl of food with berries and a pitcher and spoon on a purple place mat. Score: 9937.1	A woman wearing a blue t-shirt while looking at her cell phone and sitting on a bench next to a bright pink wall. Score: 9937
A view of the street signs "W 122 St.", "Seminary Row", and "Broadway" in front of an old red brick building. Score: 9923.6	United States President Barack Obama gives a speech in front of American and Russian flags. Score: 9913.7	A young redheaded woman in sunglasses and black tank top holds a black leather purse and a white umbrella. Score: 9906.4	A clock tower that is blue reading "Lenox Hill Hospital" Score: 9895.7	A runner poses beneath a "Run for Rights" sign in a green city park. Score: 9895	A bunch of small red flowers in a barnacle encrusted clay vase Score: 9881

(a) Top 18 texts

A lower shot of someone riding their skateboard. Score: 5388.3	A man is performing tricks on a skateboard. Score: 5446.4	A player in action up to bat in a baseball game. Score: 5463.3	A player is up to bat in a baseball game. Score: 5464.1	A train is traveling along a stretch of track. Score: 5464.3	A person that is doing a trick on a skateboard. Score: 5486.2
A baseball player at bat during a baseball game. Score: 5501.7	A man is doing tricks on a skateboard. Score: 5505.4	A man is doing tricks on a skateboard. Score: 5505.4	A picture of a person on a skateboard. Score: 5506.3	A rider is riding a horse at a race track. Score: 5511.6	A person is riding on a skateboard on the street. Score: 5513.1
A person rides a bike on the road. Score: 5514.4	A man who is riding on a skateboard. Score: 5515.6	A man is doing a trick on a skateboard. Score: 5523	A man is doing a trick on a skateboard. Score: 5523	A train that is riding on tracks through a station. Score: 5528.2	A picture of a person playing in a tennis game. Score: 5549.5

(b) Bottom 18 texts

Figure 9. Top and bottom captions ranked by D_{C}.

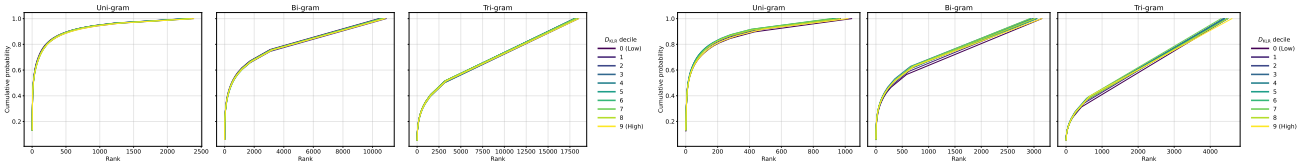
Buses parked on a road outside a large bus station. Score: 3.0478e+05	A city road with buses and people on the sidewalk Score: 2.8882e+05	A city road with buses and bus park Score: 2.8675e+05	A housewife holds a platter of food in the kitchen. Score: 2.8587e+05	A city street with cars and street signs giving directions Score: 2.8556e+05	A bus driving in a city area with traffic signs. Score: 2.8245e+05
Several animals standing in the grass near a lake. Score: 2.8216e+05	Covered food is sitting on a kitchen counter. Score: 2.8135e+05	African animals placidly grazing in a park enclosure. Score: 2.803e+05	People crossing the street near a parked sightseeing bus. Score: 2.7734e+05	many animals in a field with trees and bushes in the background Score: 2.7708e+05	A herd of wild animals walking across a dry grass park. Score: 2.7704e+05
two public transit buses on a city street Score: 2.7531e+05	A herd of animals grazing on a dry grass field. Score: 2.7424e+05	a street filled with older cars, buses and motorcycles Score: 2.7331e+05	A group of animals grazing on a lush green field. Score: 2.7215e+05	A city street with a road sign next to some buildings Score: 2.7193e+05	A bus traveling down the street in front of a large building with a clock tower. Score: 2.7066e+05

(a) Top 18 texts

There is no image to be reviewed on this hit. Score: 18497	A person is taken in this very picture. Score: 25216	There is no image here to provide a caption for. Score: 25437	There is no image here to provide a caption for. Score: 25437	There is no image here to provide a caption for. Score: 25437	There is no image here to provide a caption for. Score: 25437
I am unable to see the image above. Score: 26402	unable to see this image in this particular hit Score: 26929	I am unable to see an image above. Score: 28398	I am unable to see an image above. Score: 28398	I am unable to see an image above. Score: 28398	I am unable to see an image above. Score: 28398
I am unable to see an image above. Score: 28398	no image again there are a lot of these lately Score: 29587	I am not sure what this image is. Score: 30736	I do not know what this is supposed to be.. Score: 31176	An individual is taken in this very picture. Score: 32908	An individual is taken in this very picture. Score: 32908

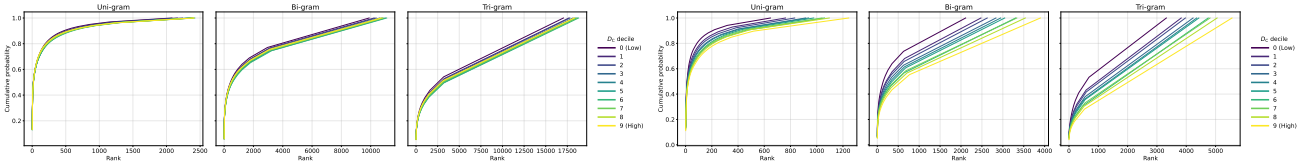
(b) Bottom 18 texts

Figure 10. Top and bottom captions ranked by D_{W}.



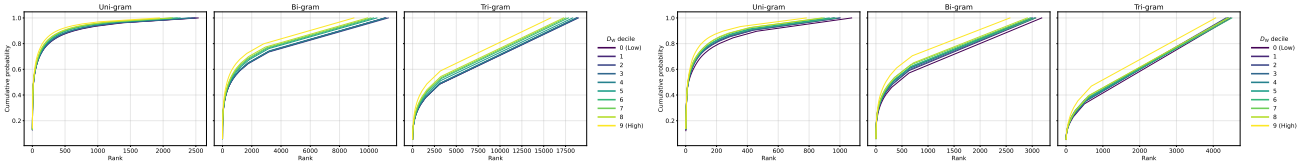
(a) Cumulative N-gram probability across $D_{\mathrm{KLR}}(i)$ deciles

(b) Cumulative N-gram probability across $D_{\mathrm{KLR}}(t)$ deciles



(c) Cumulative N-gram probability across $D_{\mathrm{C}}(i)$ deciles

(d) Cumulative N-gram probability across $D_{\mathrm{C}}(t)$ deciles



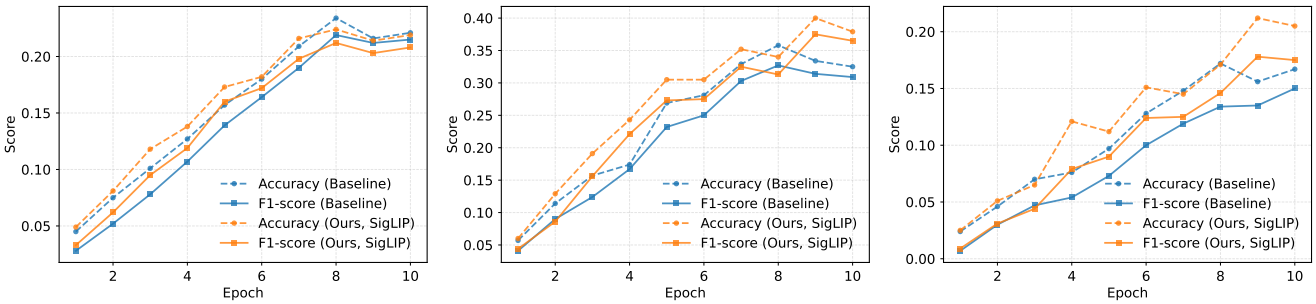
(e) Cumulative N-gram probability across $D_{\mathrm{W}}(i)$ deciles

(f) Cumulative N-gram probability across $D_{\mathrm{W}}(t)$ deciles

Figure 11. N-gram probability coverage across KL deciles. Each decile in Figs. 11a, 11c and 11e is a group of captions corresponding to images i which has the same level of $D_{\mathrm{KLR}}(i)$, $D_{\mathrm{C}}(i)$ and $D_{\mathrm{W}}(i)$. Each decile in Figs. 11b, 11d and 11f is a group of captions t which has the same level of each KL divergence.

We compared each approximation by relative errors; Bias and RMSE are divided by "scale", which is the standard deviation of $\hat{D}_*(i)$ over i, and Variance is divided by squared scale (*i.e.* variance of $\hat{D}_*(i)$ across i). We sampled 100 sets of texts and images from 50,000 image-text pairs in MSCOCO for bootstrapping.

Figures 23 and 24 shows estimation bias, variance, and RMSE of each approximation method. $\hat{D}_{\mathrm{KL}}(i)$ and $\hat{D}_{\mathrm{KLR}}(i)$ have negative estimation bias. The greater the KL divergence, the greater the bias and variance. On the



(a) Food101 (b) Oxford-IIIT Pet Dataset (c) Flowers102

Figure 12. Zero-shot classification performance comparison between baseline and our IWL method using SigLIP similarity scores across three downstream datasets.

other hand, $\hat{D}_{\mathrm{C}}(i)$ and $\hat{D}_{\mathrm{W}}(i)$ have little bias of finite sample approximation. Though the estimation variance of $\hat{D}_{\mathrm{W}}$ linearly grows when $\hat{D}_{\mathrm{W}}$ increases, the relative ratio of estimation error is still much less than that of $\hat{D}_{\mathrm{KL}}(i)$ and $\hat{D}_{\mathrm{KLR}}(i)$. These results suggest that nonlinear operations in these methods cause finite-sample approximation bias. However, all estimation methods have low estimation errors when the KL divergence is not high.

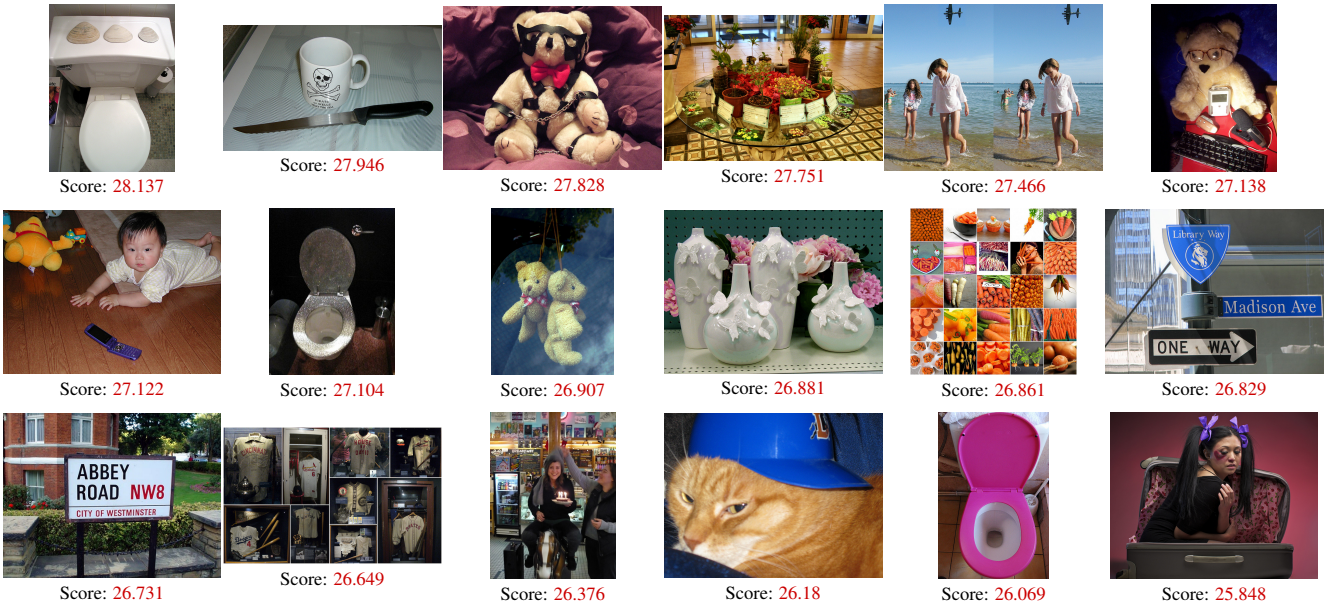
15.2. Sample size and approximation error

We also examined how the sample size n affects the error arising from finite-sample approximation. For each n, we sampled n image-text pairs from 50,000 pairs in MSCOCO 100 times, and calculated estimation RMSE using the bootstrap method. Figure 25 shows RMSE distribution on each sample size. This indicates that around 5000 samples are sufficient for all KL divergence estimation methods.

(a) Top 18 images

(b) Bottom 18 images

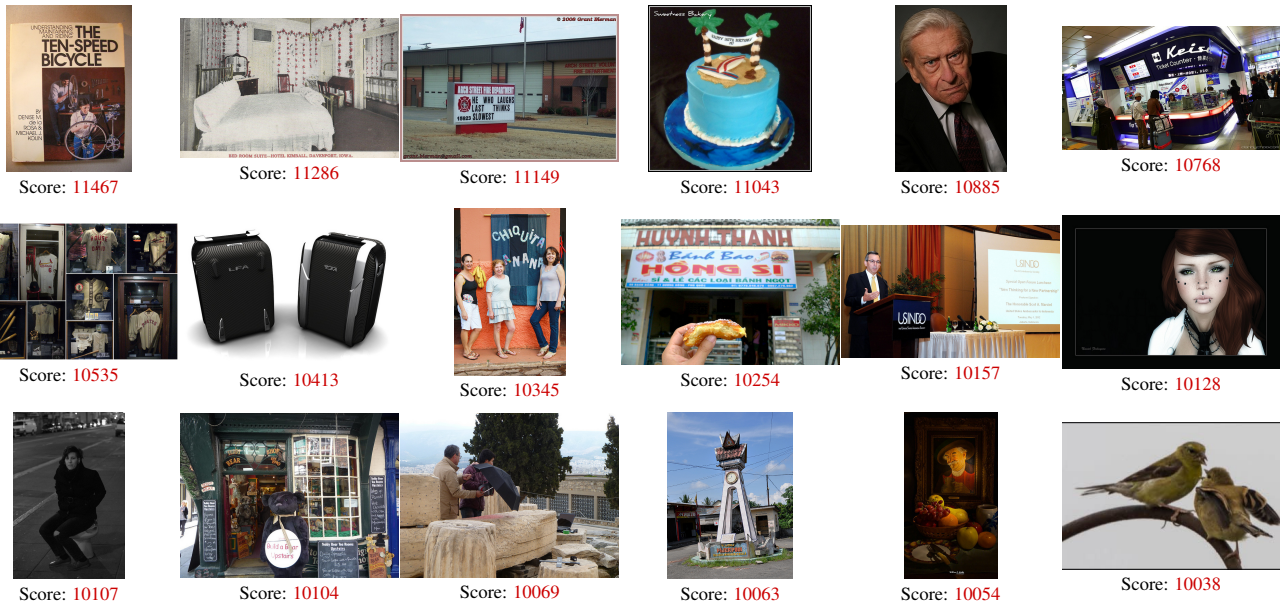
Figure 13. Top and bottom image examples in MSCOCO captions ranked by D_{KL} through similarity score of SigLIP.



(a) Top 18 images

(b) Bottom 18 images

Figure 14. Top and bottom image examples in MSCOCO captions ranked by D_{KLR} through similarity score of SigLIP.

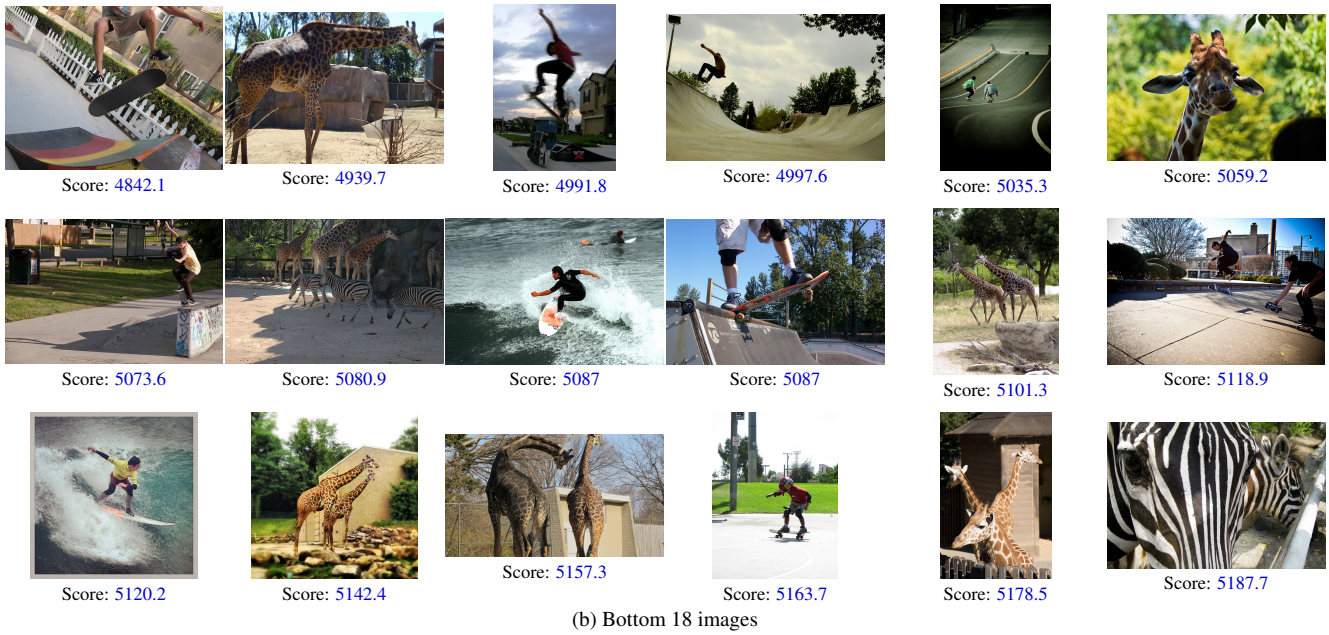


Score: 11467 Score: 11286 Score: 11149 Score: 11043 Score: 10885 Score: 10768

Score: 10535 Score: 10413 Score: 10345 Score: 10254 Score: 10157 Score: 10128

Score: 10107 Score: 10104 Score: 10069 Score: 10063 Score: 10054 Score: 10038

(a) Top 18 images



Score: 4842.1 Score: 4939.7 Score: 4991.8 Score: 4997.6 Score: 5035.3 Score: 5059.2

Score: 5073.6 Score: 5080.9 Score: 5087 Score: 5087 Score: 5101.3 Score: 5118.9

Score: 5120.2 Score: 5142.4 Score: 5157.3 Score: 5163.7 Score: 5178.5 Score: 5187.7

(b) Bottom 18 images

Figure 15. Top and bottom image examples in MSCOCO captions ranked by D_{C} through similarity score of SigLIP.

(a) Top 18 images

(b) Bottom 18 images

Figure 16. Top and bottom image examples in MSCOCO captions ranked by D_{W} through similarity score of SigLIP.

Round glass table full of potted plants and description cards Score: 8.5172	A sheet cake with a lake scene displays the words, "What is the Pickup Line Today?" Score: 8.5172	A heart shaped, glass container filled with origami birds. Score: 8.5172	A clear heart shaped jar filled with many colors of paper cranes. Score: 8.5172	A sign that says "Abbey Road NW8 City of Westminster". Score: 8.5172	A girl in a blue bowtie standing next to a pink Princess Parking Only sign. Score: 8.5172
A woman standing next to a princess parking only sign. Score: 8.5172	Elvis impersonator sitting atop a metal sculpture of a bull. Score: 8.5172	A car covered in dolls and doll parts with people in the background. Score: 8.5172	a white mug showing pirate skull and bones and a large knife on a counter top. Score: 8.5172	Two almost identical photos with some minor cropping show two young girls standing on the coast with people in the water in the background as a plane flies overhead Score: 8.5172	A smiling girl standing beside a sign that says "Princess Parking Only". Score: 8.5172
a home bar with different drink ingredients under a large decorative sign saying, "PUB". Score: 8.5172	A person sits on the floor beside flip flops and a teddy bear. Score: 8.5172	A white frosted donut with sprinkles and a jump rope with leather roping next to it. Score: 8.5172	A mirror image of a red haired doll clock. Score: 8.5172	a woman selling jewelry laid on a blanket on the sidewalk Score: 8.5172	An old truck, painted over blue in the desert Score: 8.5172

(a) Top 18 texts

An individual is taken in this very picture. Score: 1.7682	An individual is taken in this very picture. Score: 1.7682	A being is doing something as of right now that is splendid. Score: 1.8934	There is no image to describe for this question. Score: 1.9816	A thing is in the outline and it shows up like something Score: 2.0293	An individual is in the open view in the picture. Score: 2.1309
There is no image here to provide a caption for. Score: 2.1893	There is no image here to provide a caption for. Score: 2.1893	There is no image here to provide a caption for. Score: 2.1893	There is no image here to provide a caption for. Score: 2.1893	A picture of a person on a skateboard. Score: 2.1898	no image again there are a lot of these lately Score: 2.225
I am unable to see the image above. Score: 2.256	unable to see this image in this particular hit Score: 2.2638	I am unable to see an image above. Score: 2.3005	I am unable to see an image above. Score: 2.3005	I am unable to see an image above. Score: 2.3005	I am unable to see an image above. Score: 2.3005

(b) Bottom 18 texts

Figure 17. Top and bottom captions in MSCOCO captions ranked by D_{KL} through similarity score of SigLIP.

a street sign for Library Way and Madison Avenue above a one way arrow Score: 34.339	A class picture of the Goodmayes Boys' School in April of 1929. Score: 33.828	A sign that says "Abbey Road NW8 City of Westminster". Score: 32.846	Street sign for Anza and 21st Streets in front of building. Score: 32.458	A group photo of men and boys from the Goodmayes Boys School dated April 1929. Score: 32.024	A sheet cake with a lake scene displays the words, "What is the Pickup Line Today?" Score: 31.716
A view of the street signs "W 122 St.", "Seminary Row", and "Broadway" in front of an old red brick building. Score: 31.507	Two almost identical photos with some minor cropping show two young girls standing on the coast with people in the water in the background as a plane flies overhead Score: 31.498	Two cranes behind a stop sign and a street sign indicating Fort York Blvd. Score: 31.297	Outside view of the MGM Grand in Las Vegas with people sitting and walking. Score: 30.966	A key bank sign with a clock on a building Score: 30.957	A waitress and restaurant owner showing off a pizza in front of a mural of Venice. Score: 30.783
Street sign at the corner of Ho'ohu and Pe'e roads on Kauai Score: 30.273	Two planes fly over a bridge in Sydney, Australia, with the Sydney Opera House in the background. Score: 30.253	A street sign at an intersection of Library Way and Madison Avenue Score: 30.237	A bus parked near an Apple sign during the night. Score: 30.107	An green and white overhead street sign on Interstate 278 for Queens and Bronx, showing a truck restriction. Score: 29.965	Street signs for the intersection of Anza and 21st Avenue in front of a large building Score: 29.895

(a) Top 18 texts

There is no image to describe for this question. Score: 1.7863	A being is doing something as of right now that is splendid. Score: 1.9594	An individual is taken in this very picture. Score: 1.9638	An individual is taken in this very picture. Score: 1.9638	A thing is in the outline and it shows up like something Score: 2.0022	There is no image here to provide a caption for. Score: 2.0878
There is no image here to provide a caption for. Score: 2.0878	There is no image here to provide a caption for. Score: 2.0878	There is no image here to provide a caption for. Score: 2.0878	I am unable to see the image above. Score: 2.1347	I do not know what this is supposed to be.. Score: 2.1526	There is no image to be reviewed on this hit. Score: 2.1852
no image again there are a lot of these lately Score: 2.1932	I am unable to see an image above. Score: 2.2464	I am unable to see an image above. Score: 2.2464	I am unable to see an image above. Score: 2.2464	I am unable to see an image above. Score: 2.2464	I am unable to see an image above. Score: 2.2464

(b) Bottom 18 texts

Figure 18. Top and bottom captions in MSCOCO captions ranked by D_{KLR} through similarity score of SigLIP.

Side by side view of two oval plates, one with fork, with chicken salad sandwiches and rosy new potatoes, by an open and an unopened bottle of lager, a pepper mill, paper towel roll, basket behind. Score: 17225	Three rectangular bowls with food; Big bowl has nine meat and sesame seed patties with brown sauce, next to it, a bowl of shredded cabbage and carrots with yogurt dollop atop, and behind that is a bowl of cut broccoli and tomatoes with seasoning. Score: 16587	An green and white overhead street sign on Interstate 278 for Queens and Bronx, showing a truck restriction. Score: 15674	Two teddy bears a pink one, and a tan one. The tan bear is wearing a pink shirt that says the harvey girls. Score: 15485	Two apples, a bowl of food with berries and a pitcher and spoon on a purple place mat. Score: 15061	a cupcake with pink icing in a pink paper cupcake holder next to a spoon. Score: 15027
Several children watch while a girl in a pink sweatshirt plays with a Wii remote while another person with a little girl on their lap snuggle in a chair in the background. Score: 15013	A woman wearing a blue t-shirt while looking at her cell phone and sitting on a bench next to a bright pink wall. Score: 14996	Two women are squatting down and petting brown and white long haired goats. Score: 14939	Four men are standing together behind a group of red chairs. Score: 14833	A styrofoam plate with shredded chicken and a dish of broccoli with cheese sauce. Score: 14825	Two very small yellow bananas on top of a wooden table next to a Malaysian coin. Score: 14822
A spoon, a large carrot, a medium carrot and a small carrot, on blue-green speckled surface Score: 14819	A street sign reading Cambridge st and Norfolk st's crossing lies broken on the ground. Score: 14714	A fork, a pork chop, green beans, an egg and parsley on a white plate. Score: 14679	Four stuffed animals, a leopard and three teddy bears, in a row sitting on a stone ledge with grass and trees behind. Score: 14645	A table with two black trays of food that include raspberries and broccoli. Score: 14492	School boys sit cross legged in front of a chalkboard sign in a vintage black and white photo. Score: 14481

(a) Top 18 texts

A train going down the train track. Score: 4099.8	a train sitting on the tracks at a station Score: 4258.1	A man who is performing a trick on a skateboard. Score: 4303.7	A bus stopped on the side of the road. Score: 4330.1	A train with no cars sitting on the train tracks. Score: 4341.5	a train on a track pulling into a station Score: 4352.5
a man going up a ramp on a skateboard Score: 4374.2	A train traveling down tracks near a train station. Score: 4382	a group of adults playing soccer on a field Score: 4384.1	A train engine is on the railroad tracks. Score: 4396.2	A group of kids playing a game of baseball. Score: 4411.4	A train on the tracks near a train station Score: 4412.1
A train on the train tracks in the daytime. Score: 4414.9	This is a picture of a train coming into the station. Score: 4416.9	A train is moving through a train station. Score: 4423.6	a young boy is playing soccer on a field Score: 4425.7	A man on a skateboard doing a trick. Score: 4435.2	A baseball player preparing to bat during a game. Score: 4436.2

(b) Bottom 18 texts

Figure 19. Top and bottom captions in MSCOCO captions ranked by D_{C} through similarity score of SigLIP.

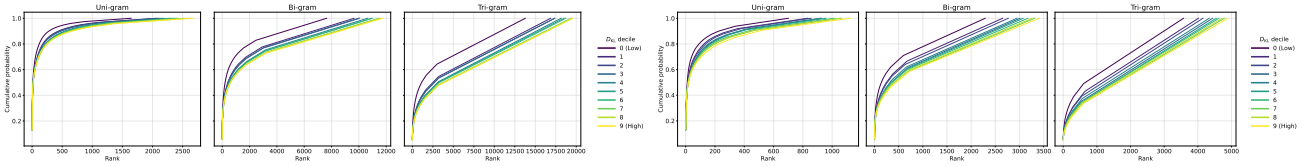
Buses parked on a road outside a large bus station. Score: 2.5771e+05	A photo of a busy intersection with a bus, several cars and a clock tower. Score: 2.4974e+05	An intersection with antique cars and a bus at it. Score: 2.4973e+05	A city road with buses and people on the sidewalk Score: 2.4972e+05	a car and a bus at an intersection in front of a historical looking building Score: 2.477e+05	A street filled with lots of cars next to a boat. Score: 2.4403e+05
a group of ball players playing on a field in front of an audience Score: 2.4109e+05	People walking on the sidewalk of a city with a tour bus in the background. Score: 2.4098e+05	People crossing the street near a parked sightseeing bus. Score: 2.3821e+05	A bus and other cars driving down a multi-laned street . Score: 2.3718e+05	parked cars, motorcycles and buses on a cobblestone parking lot next to a street. Score: 2.3707e+05	Two buses in city with other cars and trees. Score: 2.3516e+05
A busy street with cars, a motorcycle and a passenger bus Score: 2.32e+05	A large bus and other vehicles on a busy street. Score: 2.3187e+05	A city road with buses and bus park Score: 2.3148e+05	An Asian city square, with people, buses, and a McDonald's. Score: 2.3134e+05	Cars and a bus stopped for a train at a crossing. Score: 2.3125e+05	Several men are playing sports on a field near some trees, wall, bus, and several buildings in the background. Score: 2.288e+05

(a) Top 18 texts

A picture of a person on a skateboard. Score: 19561	There is a picture of an outside territory. Score: 23716	There is no image to describe for this question. Score: 25238	An individual is taken in this very picture. Score: 26023	An individual is taken in this very picture. Score: 26023	An individual is in the open view in the picture. Score: 26581
A statue of two people sitting on a bench. Score: 26713	A sign on a sidewalk has a teddy bear on it. Score: 27071	A person sits on top of a bench. Score: 27396	a black and white photo of a train Score: 28376	a double telescope for seeing things that are far away Score: 28425	THERE IS A GAME A BASKETBALL GAME GOING ON Score: 28560
Somebody is in the photograph not certain who that individual is. Score: 28913	The travelers are all traveling with their best luggage. Score: 28980	Their is a hadron and their in this lot, Score: 29067	Individual is doing something at the moment that is intriguing. Score: 29491	Individual is doing something at the moment that is intriguing. Score: 29491	The hand is holding an iPhone for the picture. Score: 29611

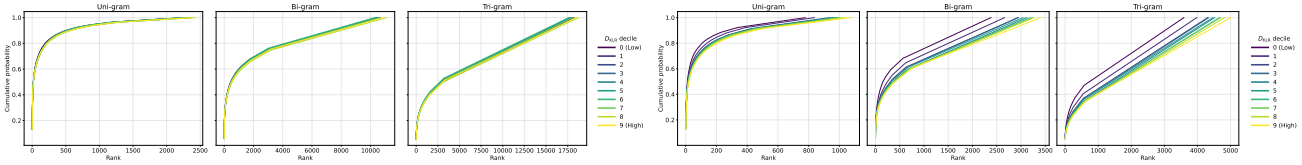
(b) Bottom 18 texts

Figure 20. Top and bottom captions in MSCOCO captions ranked by D_{W} through similarity score of SigLIP.



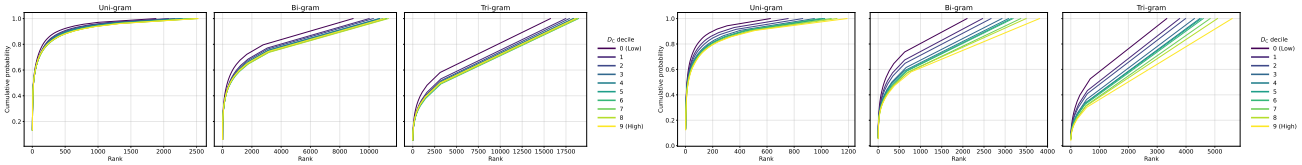
(a) Cumulative N-gram probability across $D_{\mathrm{KL}}(i)$ deciles

(b) Cumulative N-gram probability across $D_{\mathrm{KL}}(t)$ deciles



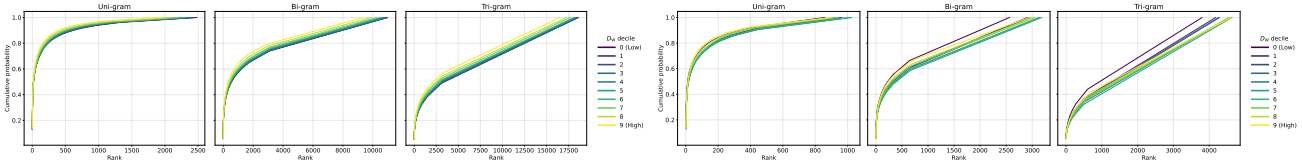
(c) Cumulative N-gram probability across $D_{\mathrm{KLR}}(i)$ deciles

(d) Cumulative N-gram probability across $D_{\mathrm{KLR}}(t)$ deciles



(e) Cumulative N-gram probability across $D_{\mathrm{C}}(i)$ deciles

(f) Cumulative N-gram probability across $D_{\mathrm{C}}(t)$ deciles



(g) Cumulative N-gram probability across $D_{\mathrm{W}}(i)$ deciles

(h) Cumulative N-gram probability across $D_{\mathrm{W}}(t)$ deciles

Figure 21. N-gram probability coverage across KL deciles calculated by SigLIP. Each decile in Figs. 11a, 11c and 11e is a group of captions corresponding to images i which has the same level of $D_{\mathrm{KLR}}(i)$, $D_{\mathrm{C}}(i)$ and $D_{\mathrm{W}}(i)$. Each decile in Figs. 11b, 11d and 11f is a group of captions t which has the same level of each KL divergence.

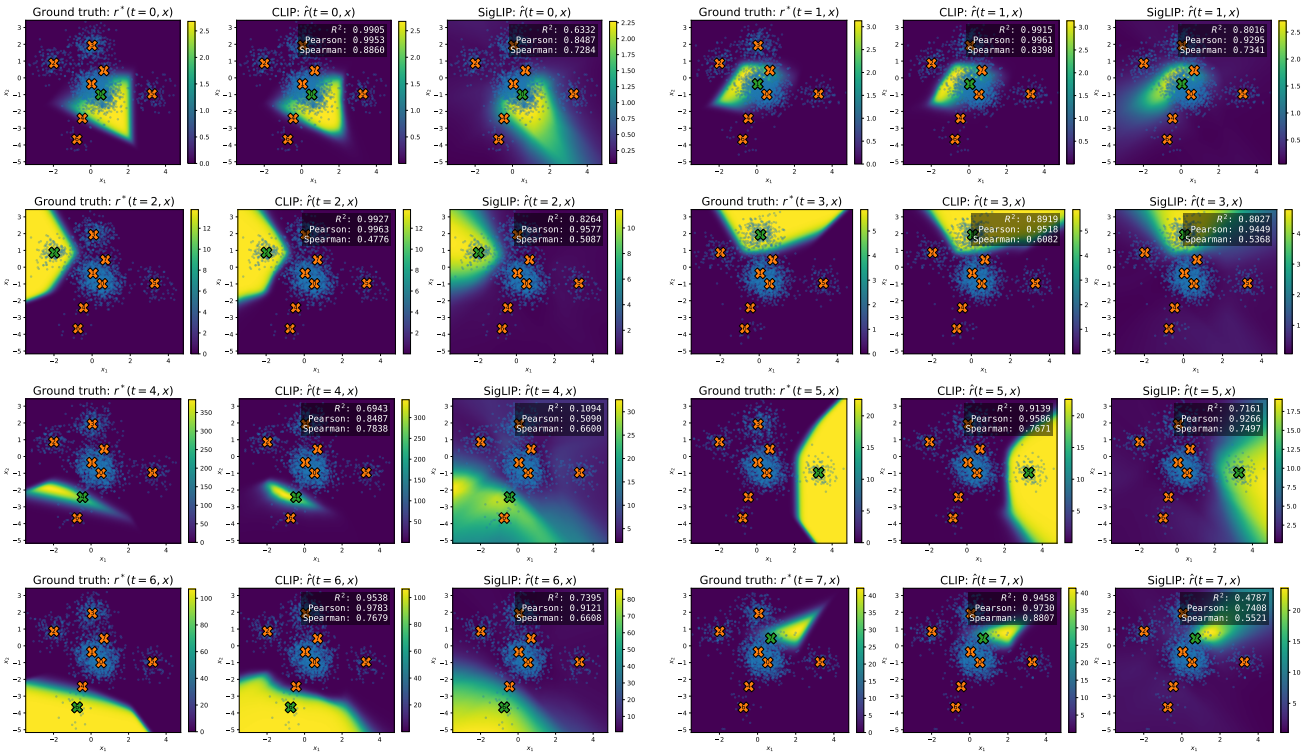


Figure 22. Visualized ground truth and predicted density ratio on image space. Each cross mark is the mean vector corresponding to the text, and blue points are the sample plots. Each metric on the upper right is computed from a grid of density ratios for each mean vector.

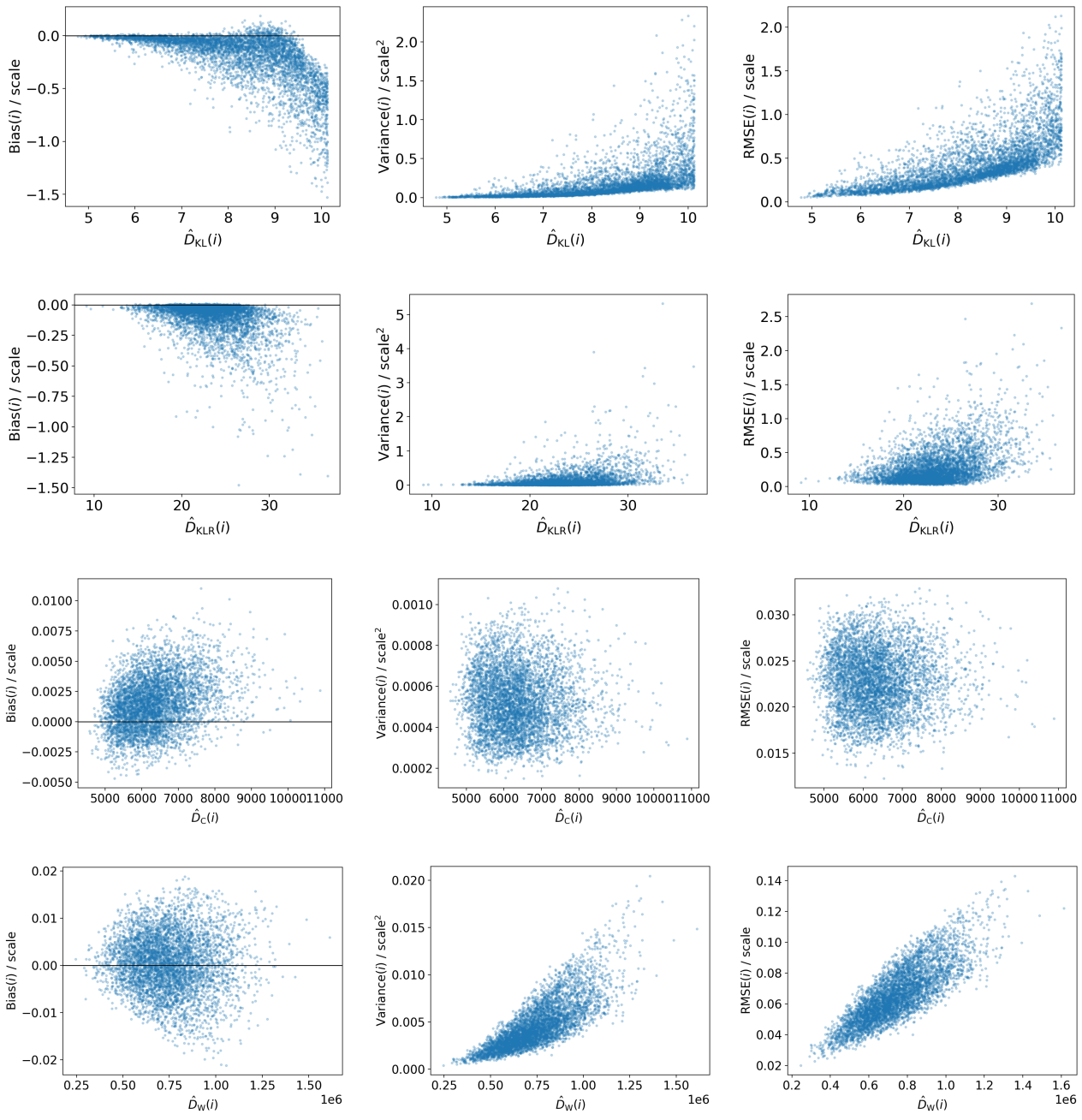


Figure 23. Scatter plots of $\hat{D}_*(i)$ and normalized $\mathrm{Bias}(i)$ (left column), $\mathrm{Variance}(i)$ (center column), $\mathrm{RMSE}(i)$ (right column). Each row represents the errors of each estimation method. KL divergences are computed using CLIP.

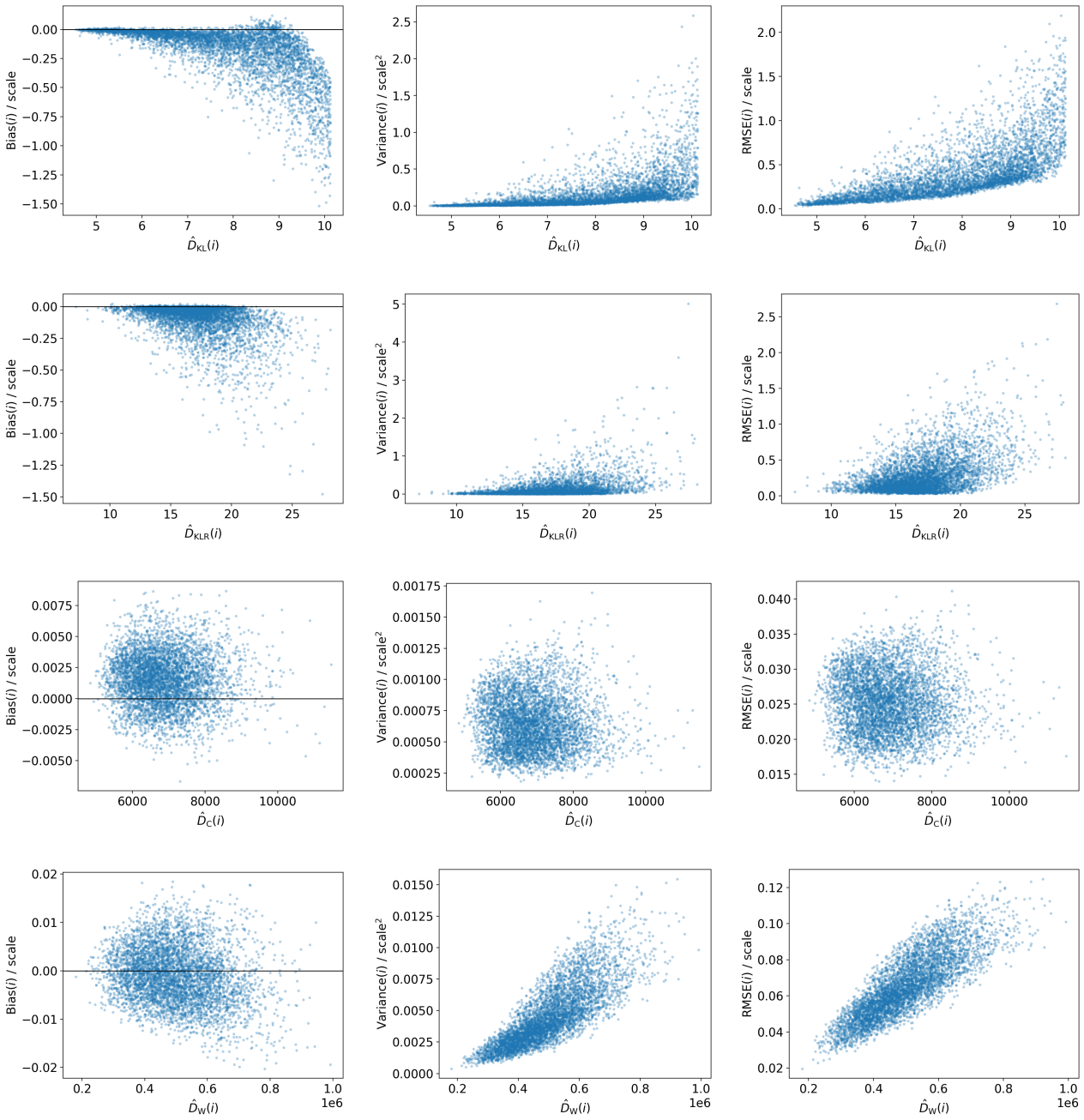
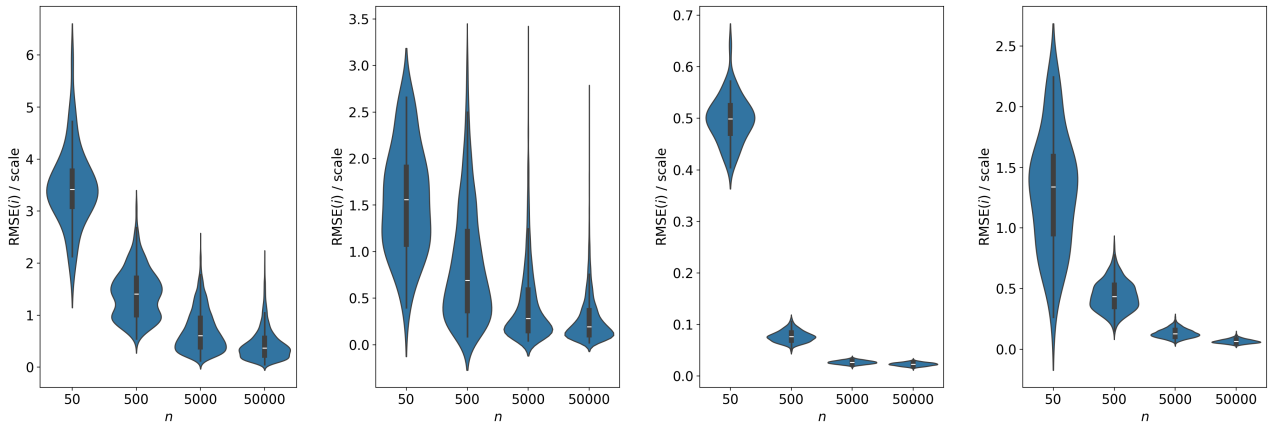
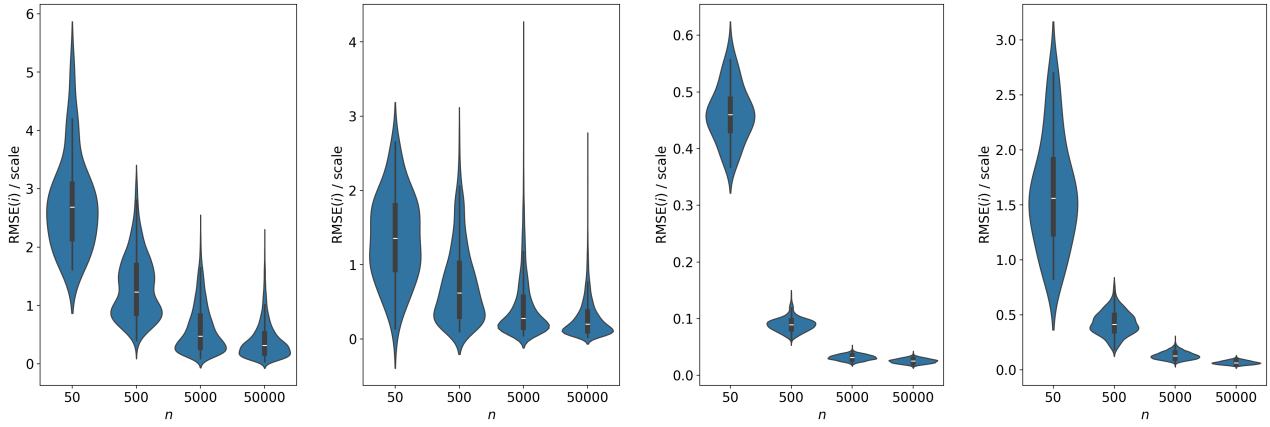


Figure 24. Scatter plots of $\hat{D}_*(i)$ and normalized $\mathrm{Bias}(i)$ (left column), $\mathrm{Variance}(i)$ (center column), $\mathrm{RMSE}(i)$ (right column). Each row represents the errors of each estimation method. KL divergences are computed using SigLIP.



(a) $D_{\mathrm{KL}}(i)$ (b) $D_{\mathrm{KLR}}(i)$ (c) $D_{\mathrm{C}}(i)$ (d) $D_{\mathrm{W}}(i)$

CLIP results



(e) $D_{\mathrm{KL}}(i)$ (f) $D_{\mathrm{KLR}}(i)$ (g) $D_{\mathrm{C}}(i)$ (h) $D_{\mathrm{W}}(i)$

SigLIP results

Figure 25. Violin plot of $\mathrm{RMSE}(i)$ on each estimation method across different sample size n.