Multimodal Representation Learning

Multimodal Generative AI Theories and Applications

Lecture 2

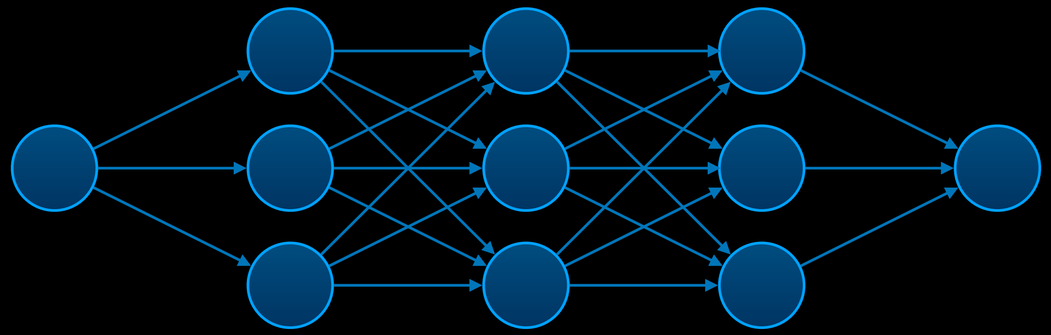
Jin-Hwa Kim and Sangdoo Yun

NAVER AI LAB

Learning multimodal representations

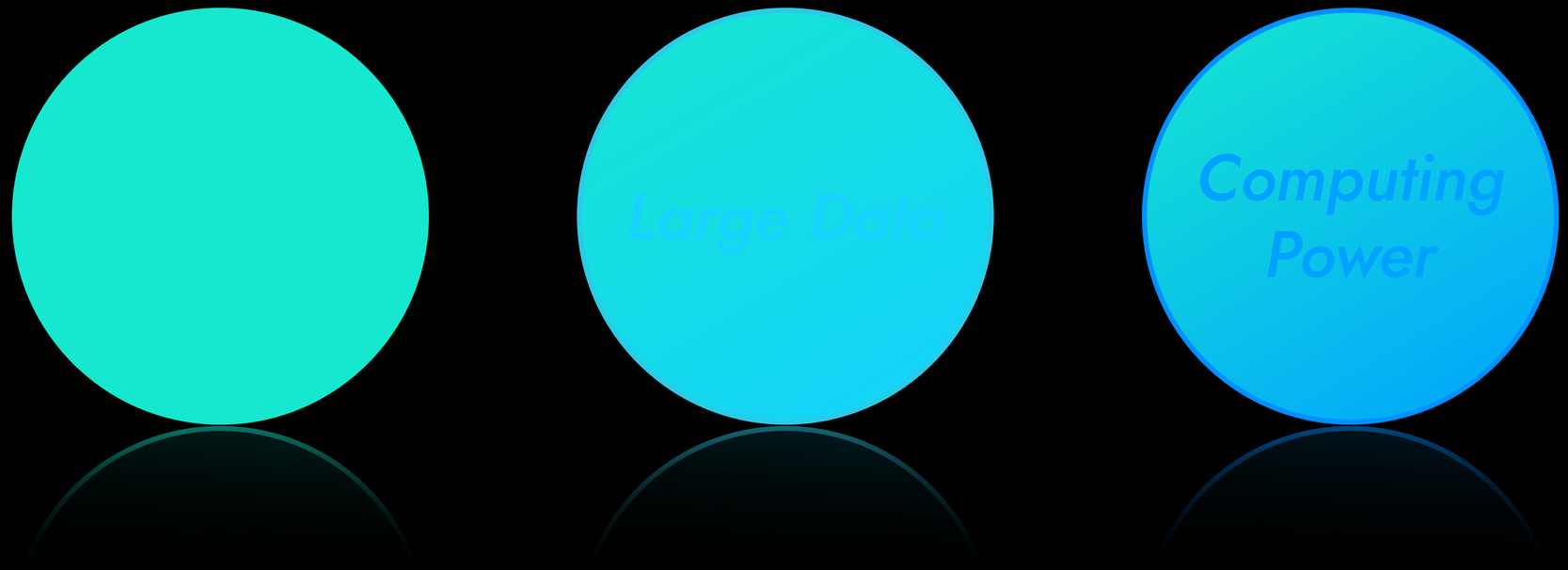
Deep learning

- *Wikipedia* says "Deep learning is part of a broader family of *machine learning* methods based on *artificial neural networks* with representation learning."
- "Artificial neural networks were inspired by *information processing* and distributed communication nodes in *biological systems*."
- Multi-layered and structured neural systems are trained with *large training data*.



Deep learning triarchy

- Advances in deep learning algorithms with large data are enabled by recently surging computational power by cutting edge manufacturers, *e.g.*, NVIDIA, Apple.



Multimodal

- In statistics, a multimodal distribution is a probability distribution with two or more modes. (Wikipedia)
- The use of two or more media in a single artifact
- Vision and language

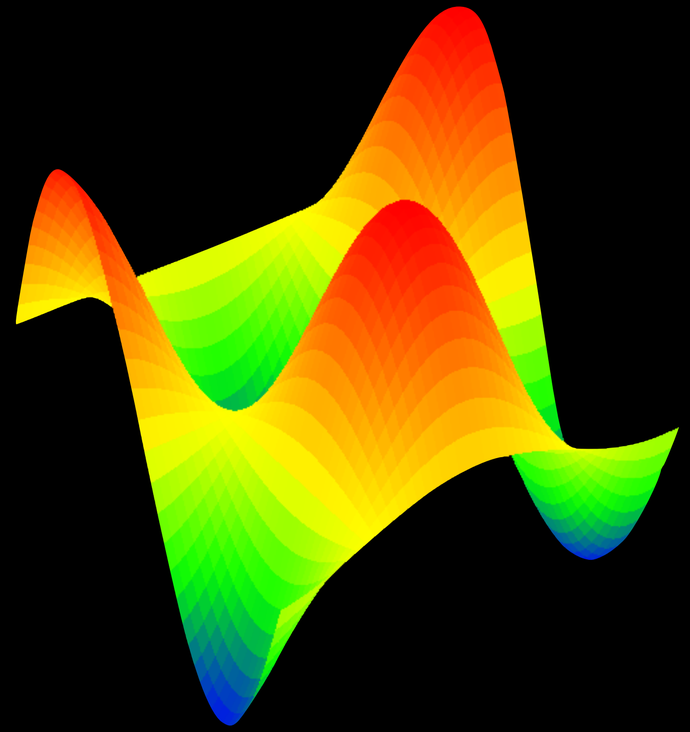


Figure credit: https://www.codeproject.com/Articles/5294893/A-Csharp-3D-Surface-Plot-Control

Multimodal deep learning

- A feature vector represents each modality using modality-dedicated models.
- A joint feature vector should be inferred from the multimodal feature vectors as input to a classifier.
- Then, how do we get the joint representation?
- *What is the limitation of this thought?*

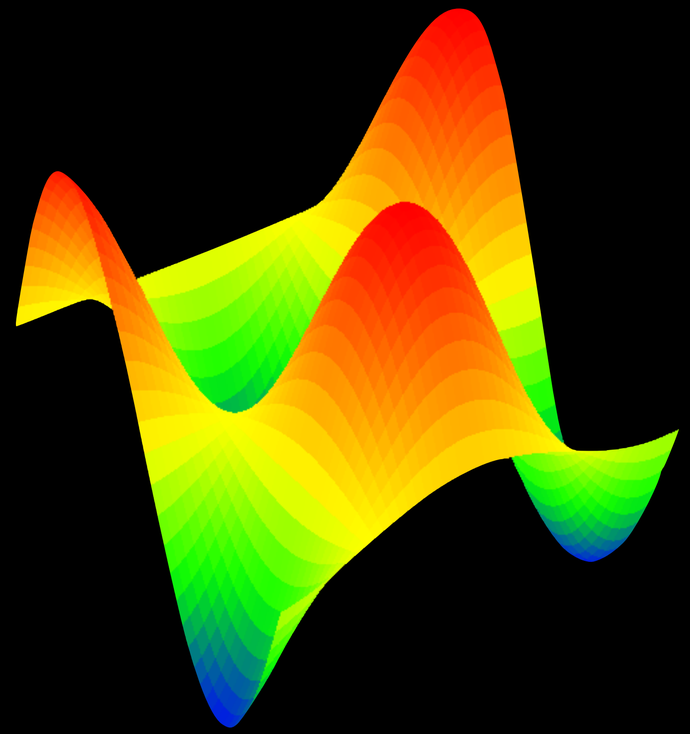
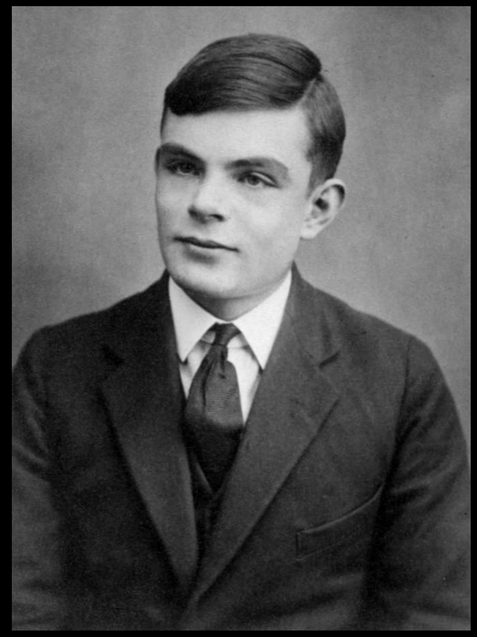


Figure credit: https://www.codeproject.com/Articles/5294893/A-Csharp-3D-Surface-Plot-Control

Visual Turing test

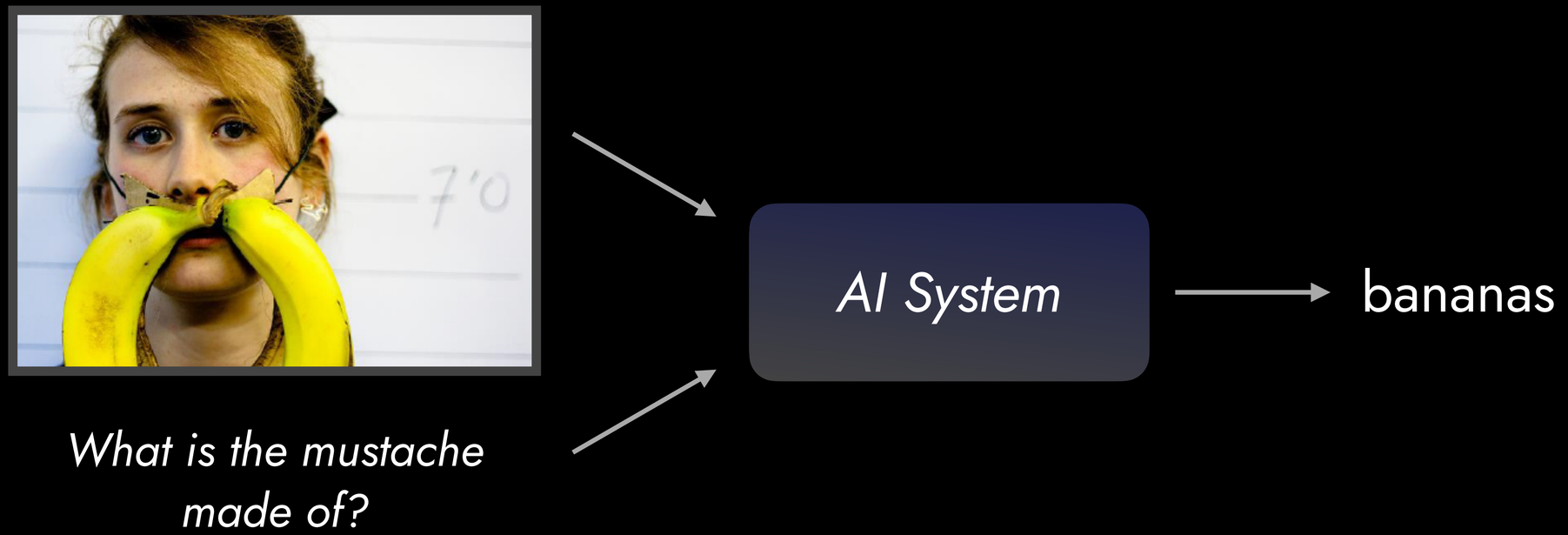
- The *"imitation game"* is a test of a machine's ability to show intelligent behaviors, which is indistinguishable from a human.
- Visual Turing test centers on machine vision and language to more deeply evaluate and interpret (You, Science 2015).



Alan M. Turing
(1912-1954)

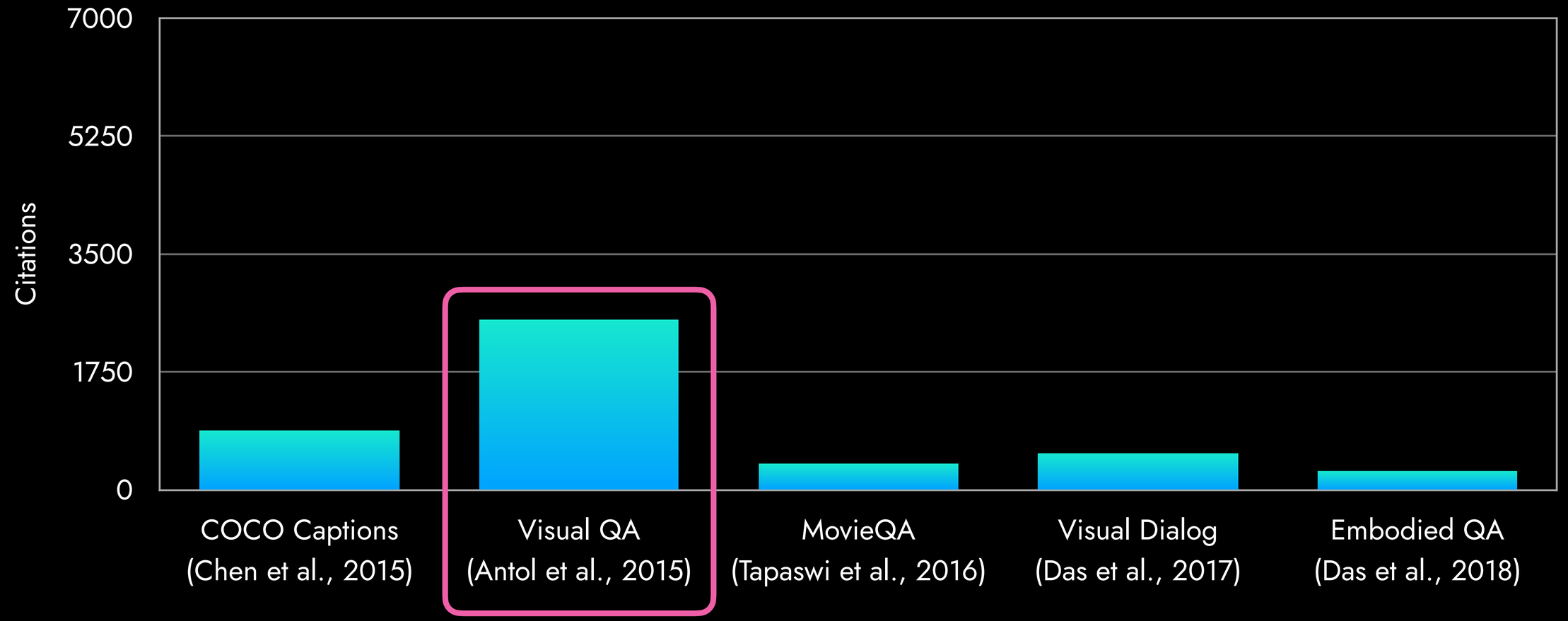
Visual question answering

- Visual Turing test as a generalization of (possibly) all vision tasks
- A representative vision and language benchmark task along with image captioning

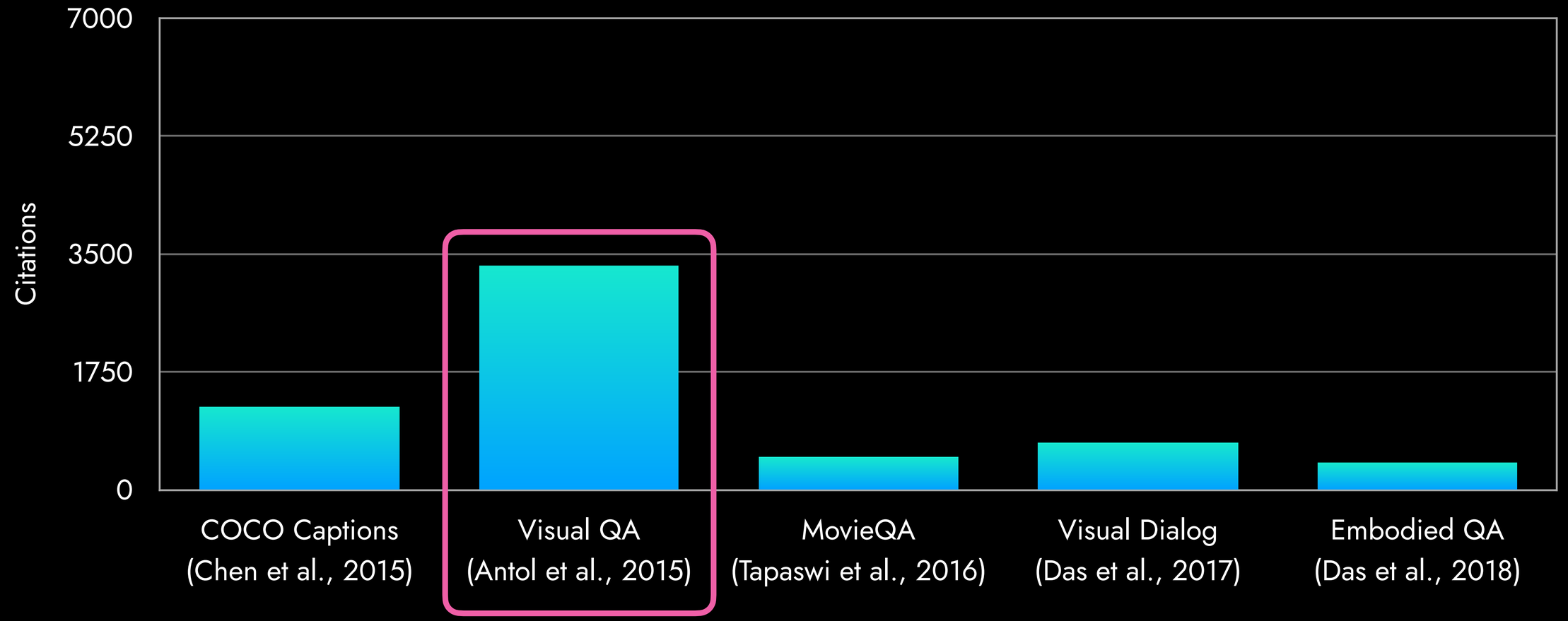


Credit: visualqa.org

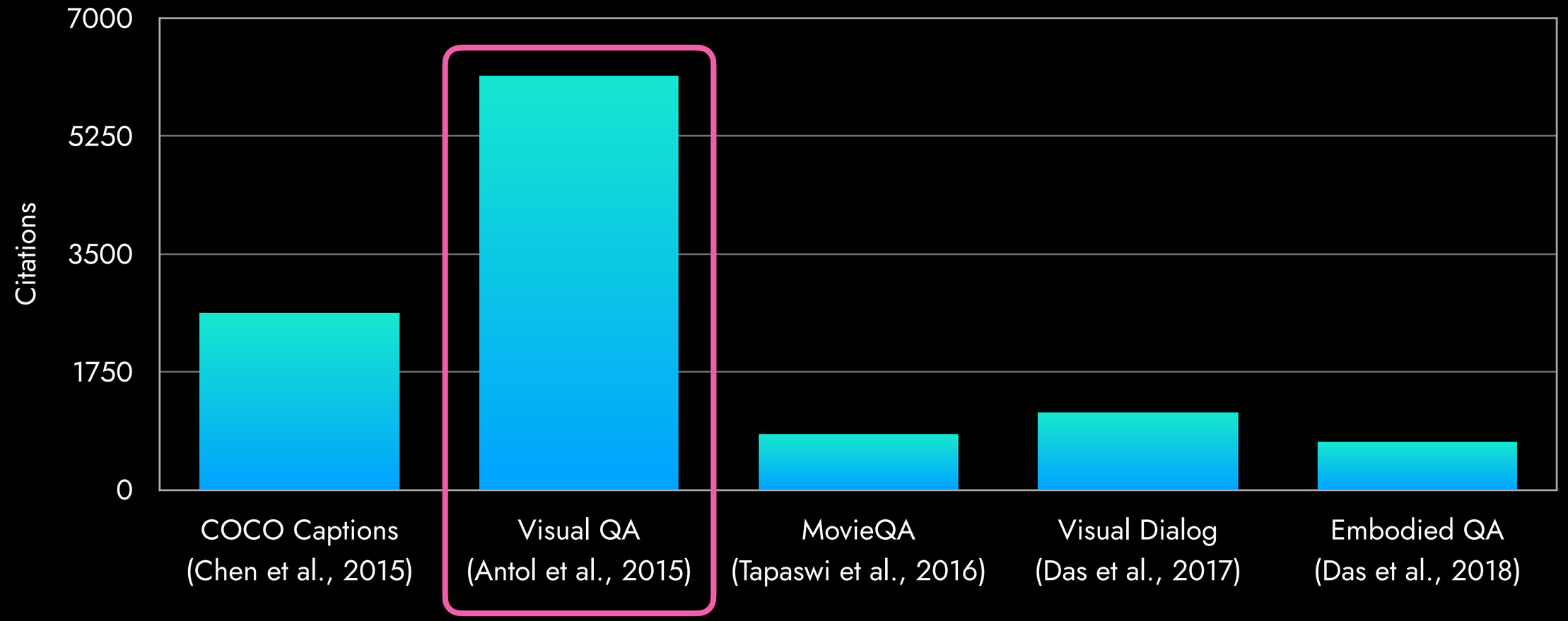
Positioning dataset papers (2021)



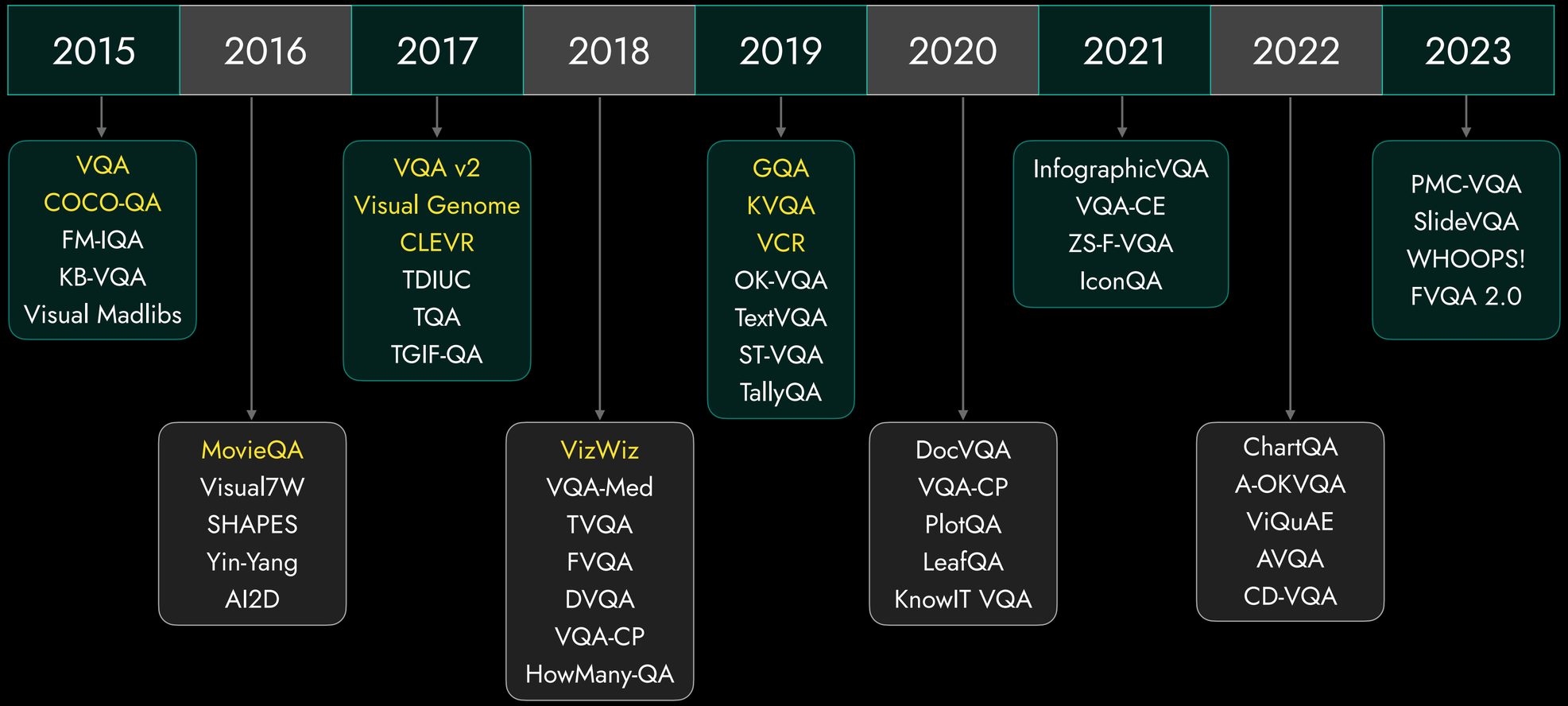
Positioning dataset papers (2022)



Positioning dataset papers (2024.09)



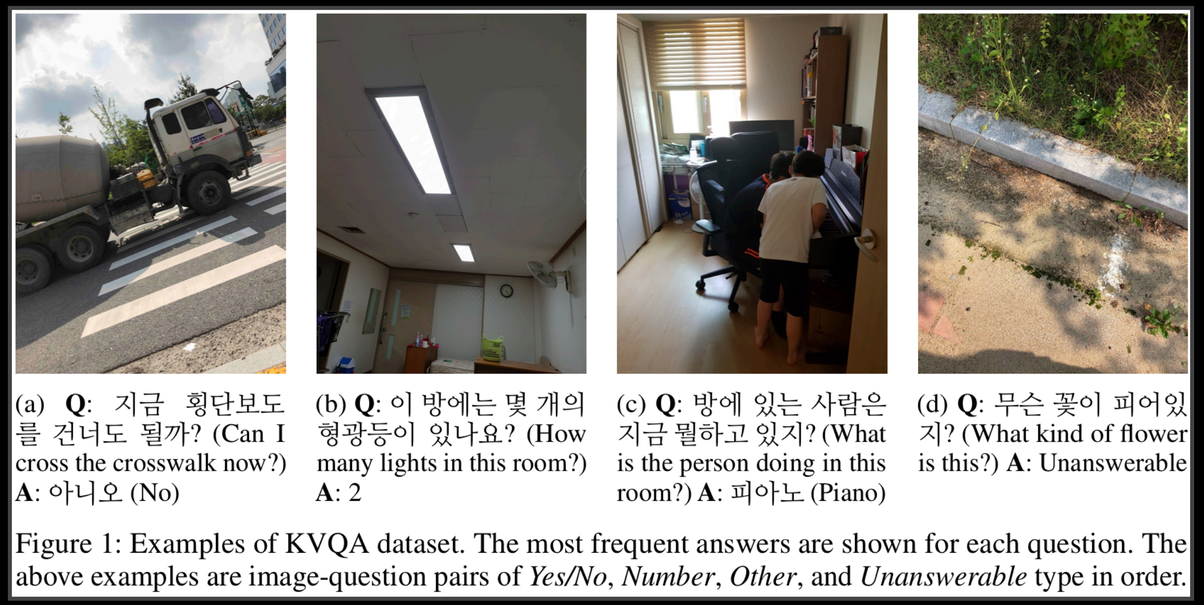
Timeline of popular VQA datasets



Ishmama et al., 2023

Visually-impaired Korean VQA

- To collect data from the blind people who volunteered to participate in this project.
- We translated parts of the published VizWiz dataset, which can be well-suited to the Korean context, and created a complete dataset to train VQA models in Korean.



(a) **Q**: 지금 횡단보도를 건너도 될까? (Can I cross the crosswalk now?) **A**: 아니오 (No)

(b) **Q**: 이 방에는 몇 개의 형광등이 있나요? (How many lights in this room?) **A**: 2

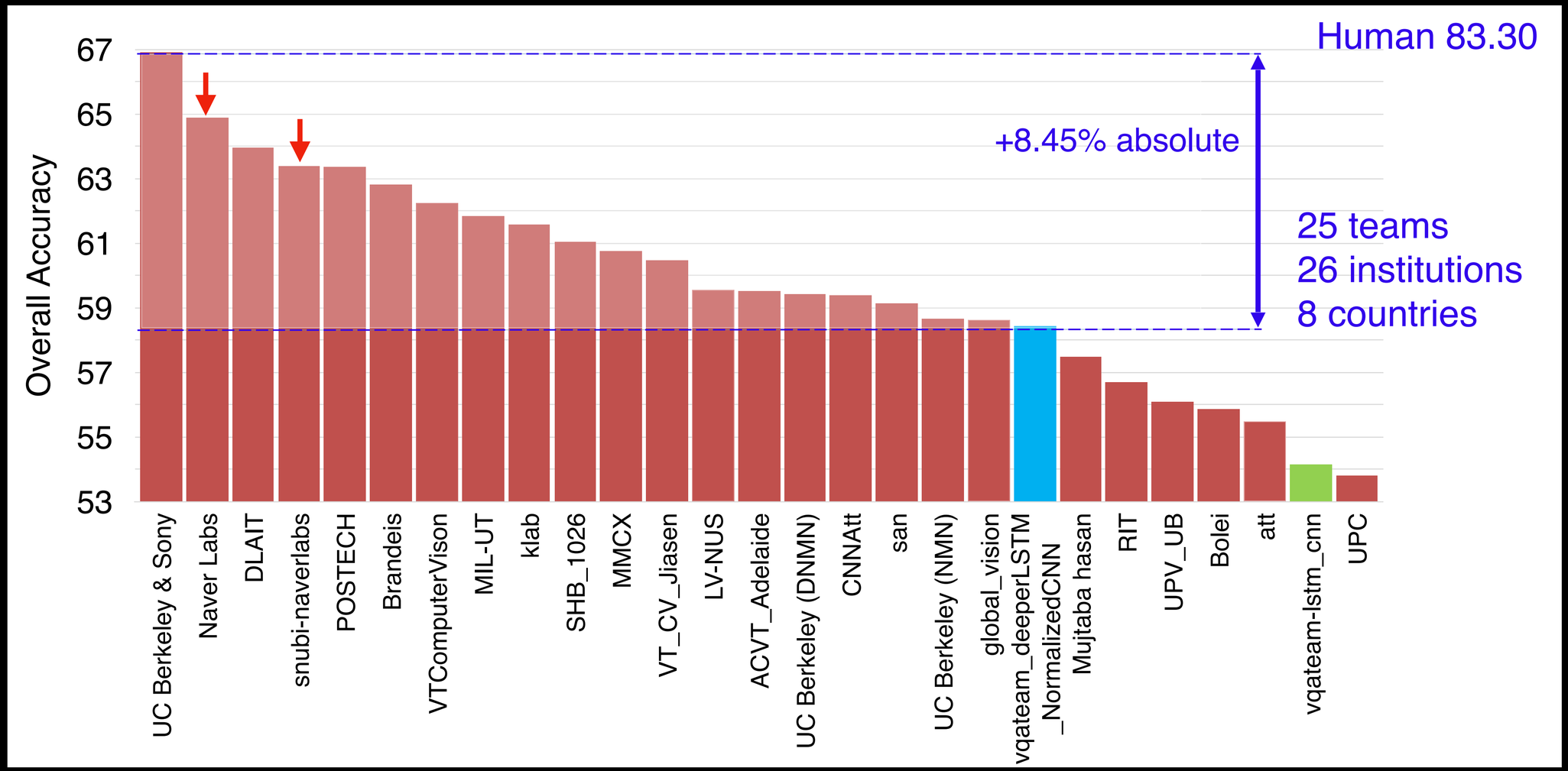
(c) **Q**: 방에 있는 사람은 지금 뭘하고 있지? (What is the person doing in this room?) **A**: 피아노 (Piano)

(d) **Q**: 무슨 꽃이 피어있지? (What kind of flower is this?) **A**: Unanswerable

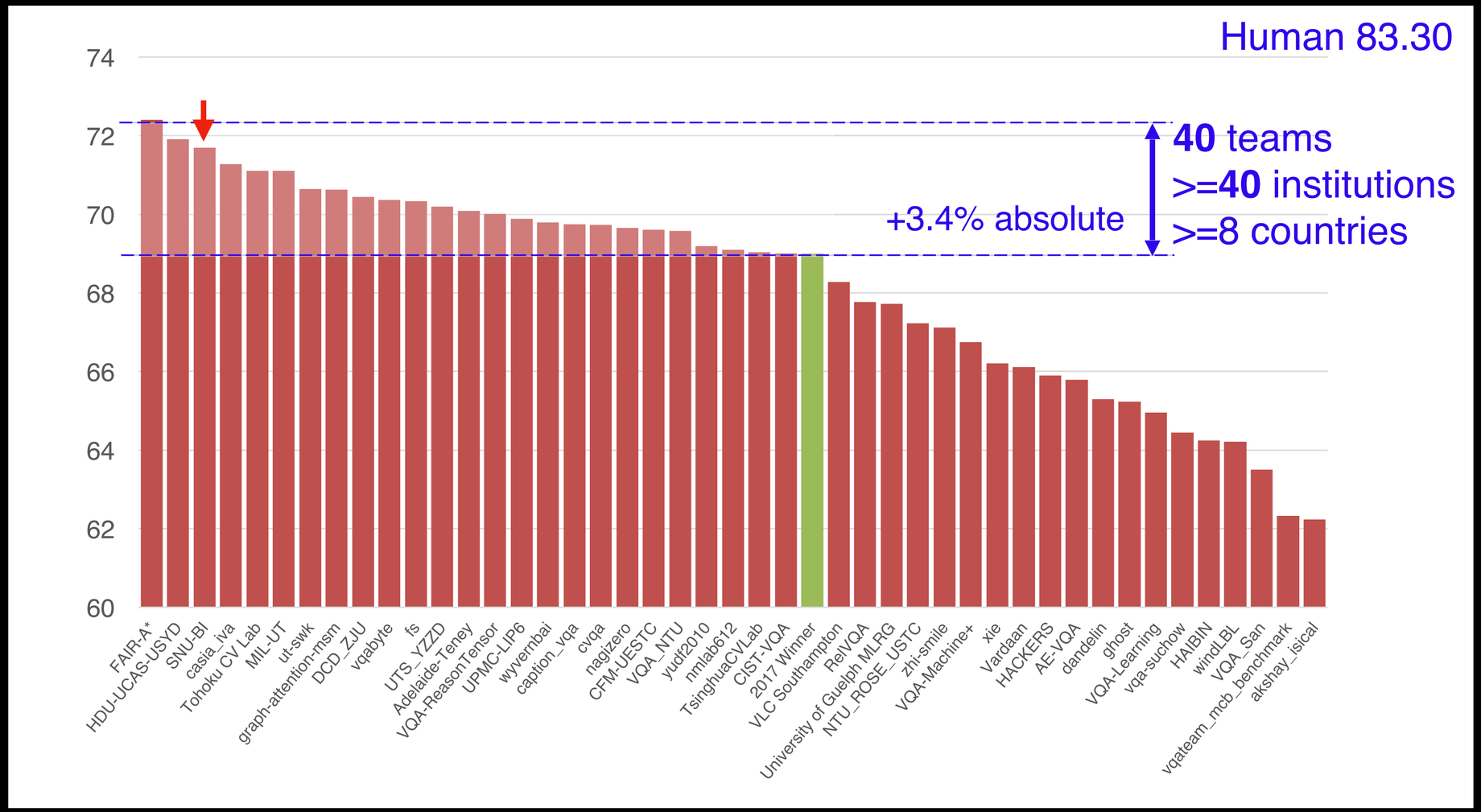
Figure 1: Examples of KVQA dataset. The most frequent answers are shown for each question. The above examples are image-question pairs of *Yes/No*, *Number*, *Other*, and *Unanswerable* type in order.

https://sktbrain.github.io/KVQA

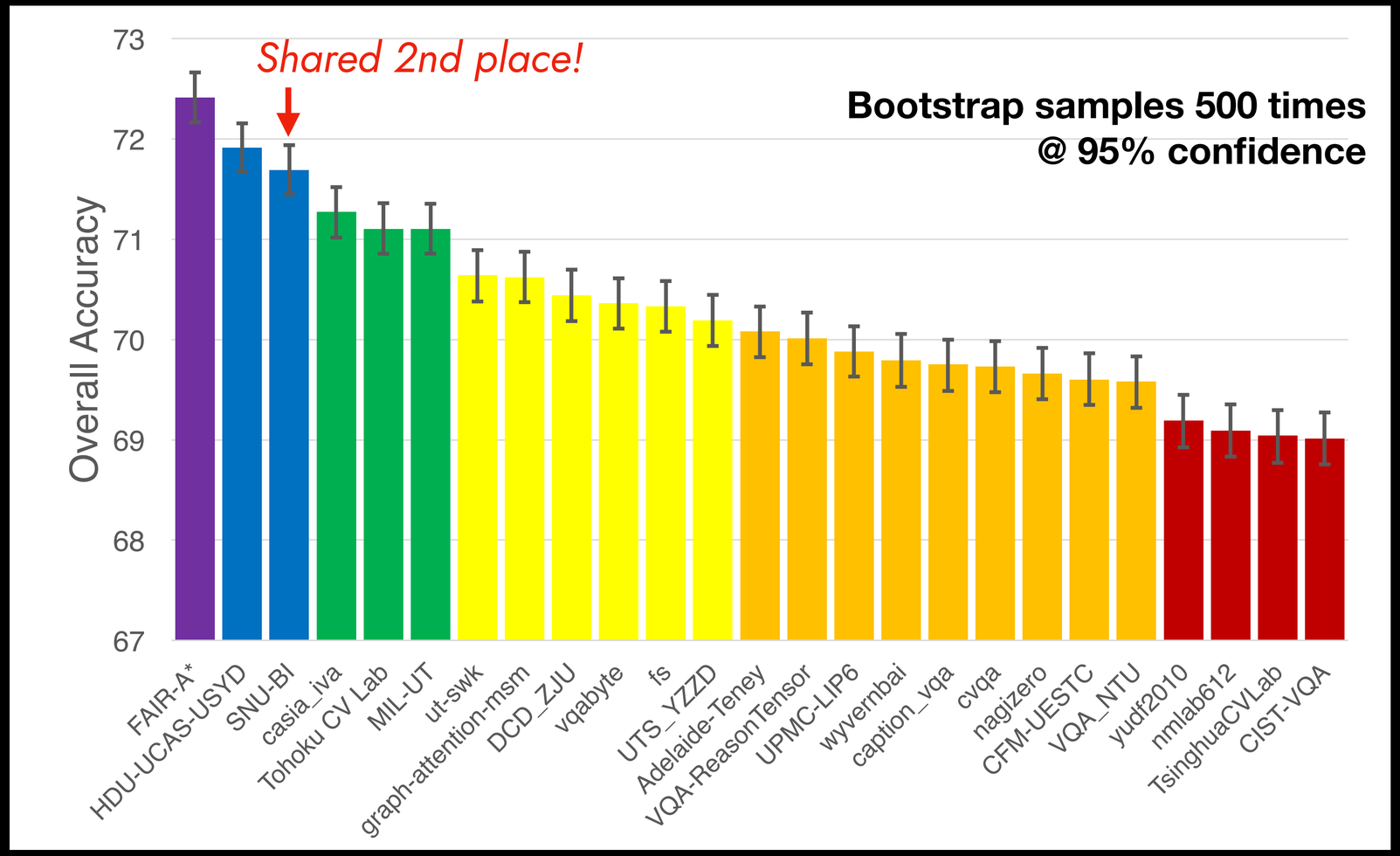
Challenge 2016 (v1.0 real open-ended)



Challenge 2018 (v2.0 real open-ended)



Statistical significance



Challenge 2021 (v2.0 real open-ended)

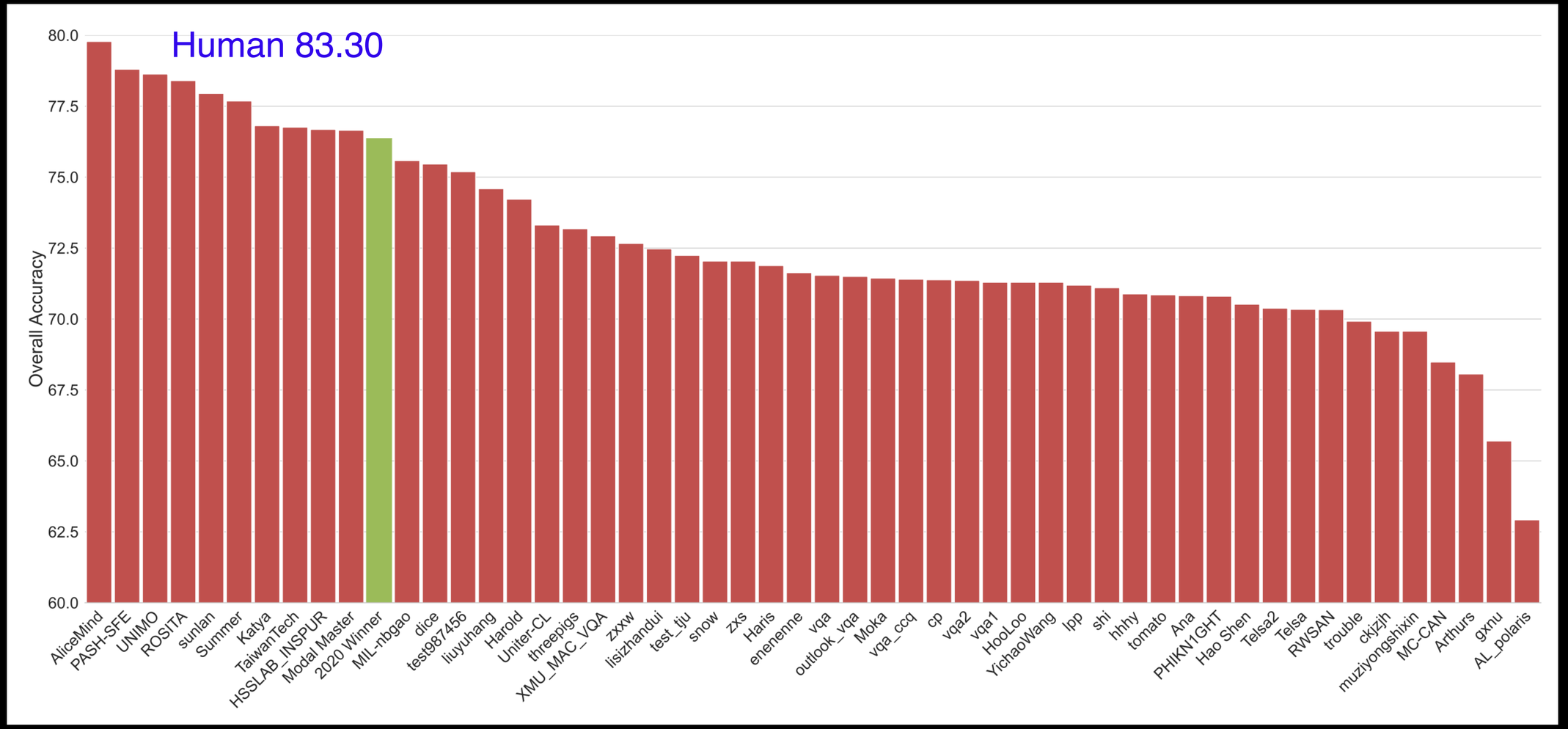


Image captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A horse carrying a large load of hay and two people sitting on it.

Multimodal tasks

	Feature learning	Training	Test
Deep learning	X	X	X
Multimodal fusion	X, Y	X, Y	X, Y
Cross modality	X, Y	X	X
Shared representation	X, Y	X	Y

Ngiam et al., ICML 2011

Multimodal tasks

	Feature learning	Training	Test
Deep learning	X	X	X
Multimodal fusion	X, Y	X, Y	X, Y
Cross modality	X, Y	X	X
Shared representation	X, Y	X	Y

Ngiam et al., ICML 2011

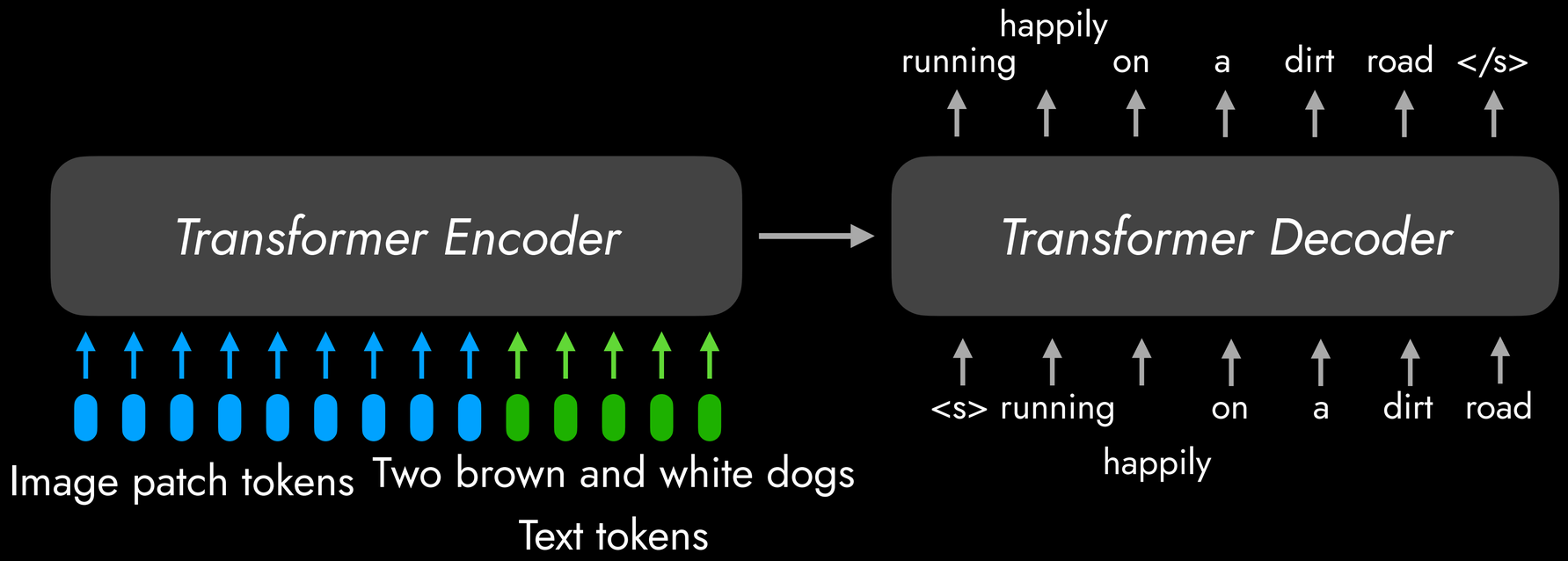
Multimodal tasks

	Feature learning	Training	Test
Deep learning	X	X	X
Multimodal fusion	X, Y	X, Y	X, Y
Cross modality	X, Y	X	X
Shared representation	X, Y	X	Y

Ngiam et al., ICML 2011

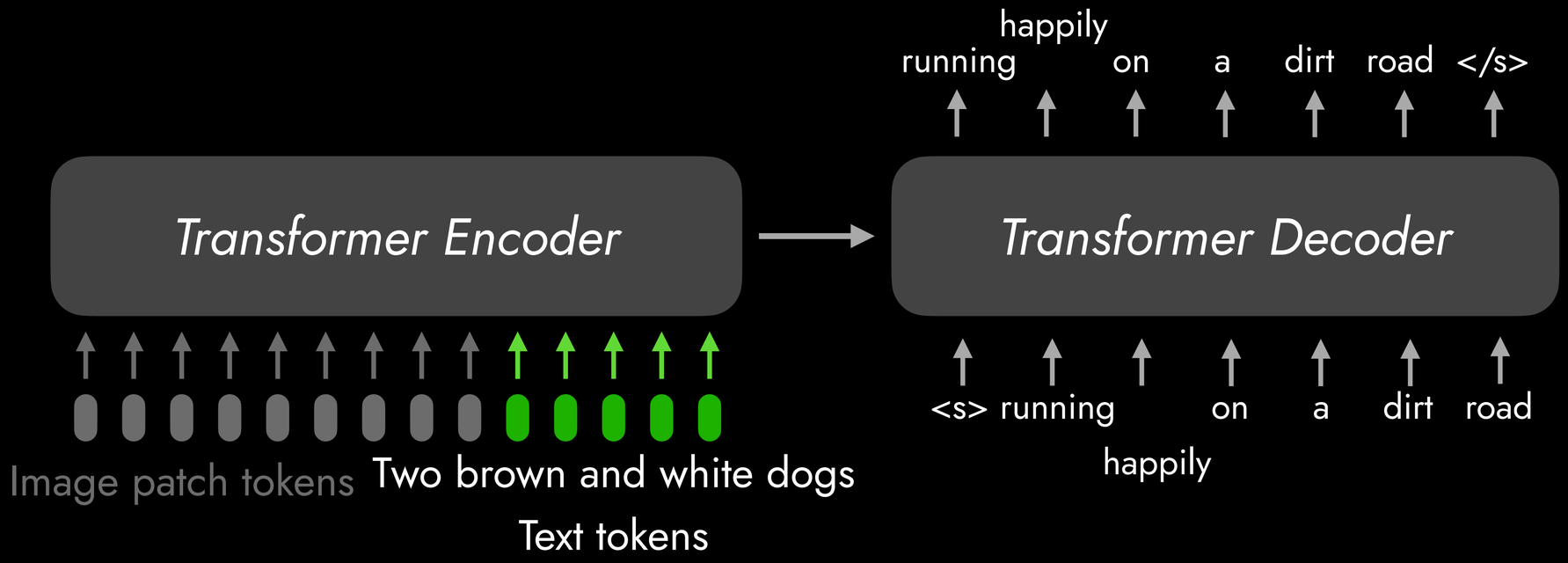
Use-case: SimVLM's versatility

- SimVLM (Wang et al., 2021) "pretrains on large-scale web datasets for both image-text and text-only inputs."
- Their formulation of *PrefixLM* is modality-agnostic, where text-only corpora to compensate for noisy text supervision in web-crawled image-text datasets.



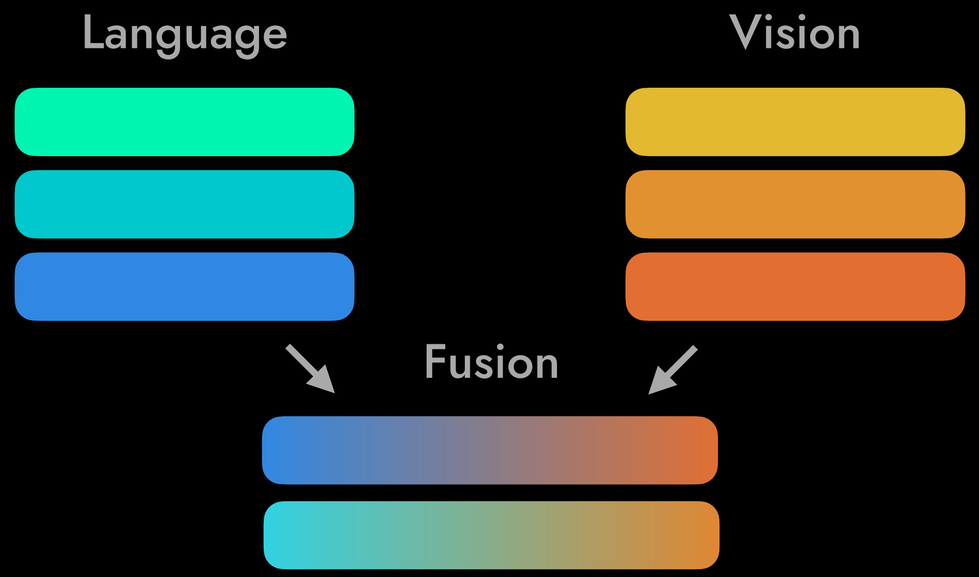
Use-case: SimVLM's versatility

- SimVLM (Wang et al., 2021) "pretrains on large-scale web datasets for both image-text and *text-only* inputs."
- Their formulation of *PrefixLM* is modality-agnostic, where text-only corpora to compensate for noisy text supervision in web-crawled image-text datasets.

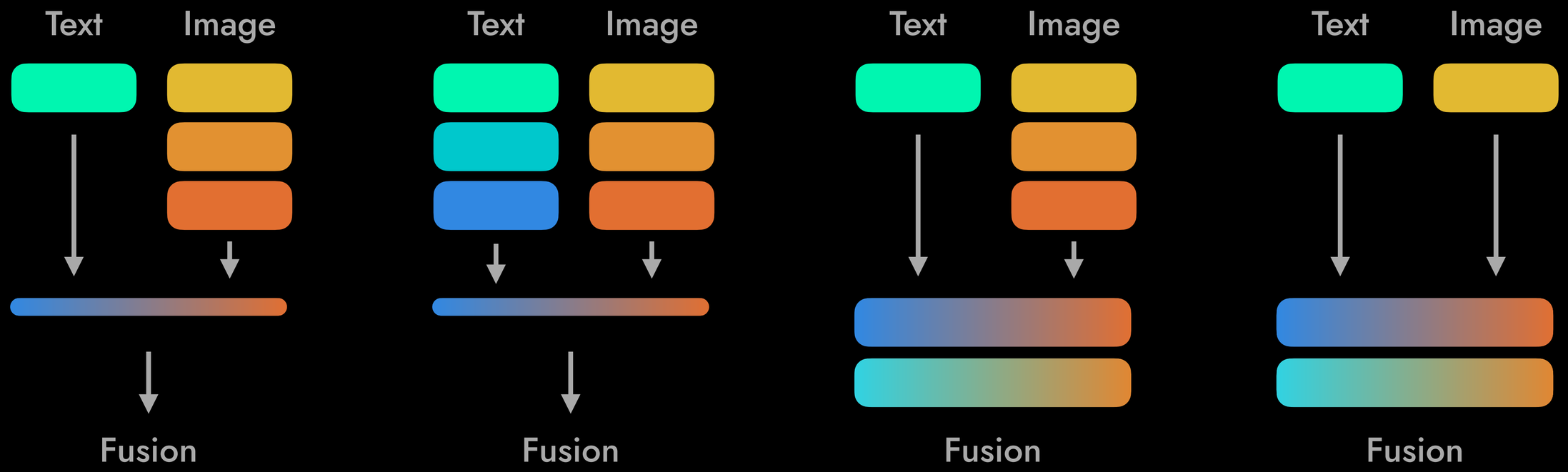


Difficulty of multimodal learning

- Each representation of modality should align with the others
- "Deep" representation helps to learn joint representation (Ngiam et al., 2011)



Four classes of multimodal models



VSE++ (Faghri et al., 2017)
SCAN (Lee et al., 2018)

CLIP (Radford et al., 2021)

Modulation-based
(Perez et al., 2018;
Nguyen et al., 2020)

ViLT (Kim et al., 2021)

Reproduced Figure 2 from ViLT (Kim et al., 2021)

Multimodal fusion

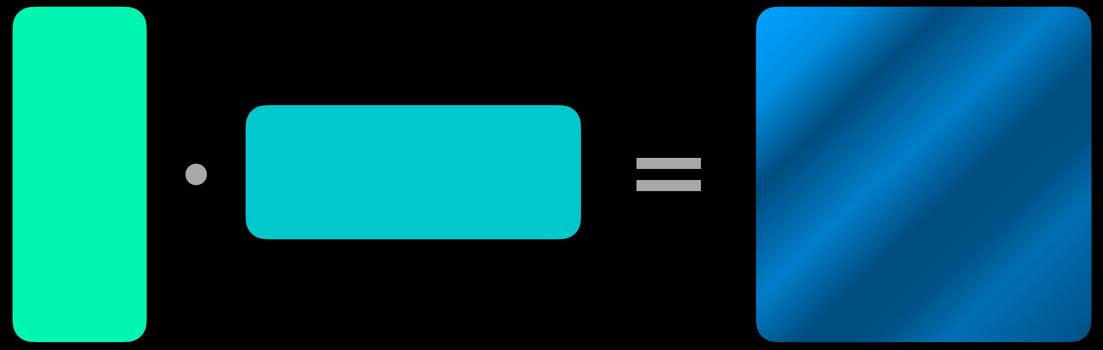
- Traditional methods include addition ($f + g$), concatenation ($f \,|\, g$) and multiply ($f \cdot g$)
  - Hadamard product after linear embedding is low-rank bilinear pooling (Kim et al., 2017)

$$\sum_{ij} w_{ij} x_i y_j = x^\intercal W y \approx x^\intercal U V^\intercal y = 1^\intercal (U^\intercal x \circ V^\intercal y)$$

- *Cross-attention mechanisms* have been developed in multimodal deep learning.
  - Queries and keys are from different modality.
- *Contrastive losses* (*e.g.*, InfoNCE) for self-supervised learning (Oord et al., 2018)

Low-rank approximation

- The *rank* of a matrix A is "the dimension of the vector space generated by its columns," or "the maximal number of linearly independent columns of A."



The rank is at most $d \leq \min(N, M)$. Roughly speaking, no more than d.

Pirsiavash et al., 2009; Kim et al., 2017

Eckart–Young–Mirsky theorem

- Minimize the matrix error $\|\mathrm{A} - \hat{\mathrm{A}}\|_F$ over $\hat{\mathrm{A}}$ subject to $\mathrm{rank}(\hat{\mathrm{A}}) \leq r$
- The optimal $\hat{\mathrm{A}}^\star = \mathrm{U}_r \Sigma_r \mathrm{V}_r^\intercal$ where the singular value decomposition obtains $A = \mathrm{U}\Sigma\mathrm{V}^\intercal$ and the subscript r indicates the slice of the first r columns or rows (*i.e.*, $\mathrm{U}_r \in \mathbb{R}^{m \times r}$, $\mathrm{V}_r \in \mathbb{R}^{n \times r}$, and $\Sigma_r \in \mathbb{R}^{r \times r}$). A diagonal matrix Σ_r contains the largest singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$.
- The error is $\min_{\mathrm{rank}(\hat{\mathrm{A}}) \leq r} \|\mathrm{A} - \hat{\mathrm{A}}\|_F = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_m^2}$ where $m \leq n$.
- *The proof is easy and informative; try to work it out yourself at least once.*

Eckart & Young, 1936; https://en.wikipedia.org/wiki/Low-rank_approximation

Low-rank "tensor" approximation

- *Canonical polyadic (CP) decomposition* extends the previous idea:

$$\underset{\mathrm{a,b,c}}{\operatorname{argmin}} \left\| A - \sum_{i=1}^{r} \mathrm{a}_i \circ \mathrm{b}_i \circ \mathrm{c}_i \right\|$$

  where $\circ$ denotes outer product, $\mathrm{A} \in \mathbb{R}^{m \times n \times k}$, $\mathrm{a} \in \mathbb{R}^{m}$, $\mathrm{b} \in \mathbb{R}^{n}$, and $\mathrm{c} \in \mathbb{R}^{k}$.

- *Tucker decomposition* makes a tensor into a set of smaller matrices and one small core tensor. It is named after Ledyard R. Tucker; but it goes back to Hitchcock (1927).

Hitchcock, 1927

Cross-attention mechanisms

- "Show, Attend and Tell (Xu et al., 2015)" proposed soft and hard attentions:
  - Hard attention uses a multinoulli distribution parameterized by weights.
  - While soft attention is deterministic and a convex combination of values.
- Co-attention or dual-attention (Lu et al., 2016; Nam et al., 2017; Kim et al., 2018)
  - Alternating query and key to get a joint representation from different views
  - BAN (Kim et al., 2018) is proposed to *combine two different views simultaneously*.
- Transformer (Vaswani et al., 2017; Yu et al., 2019)
  - Guided attention for multimodal learning
  - Known for parameter-efficient, multi-layer stackable, and better performance.

Preliminaries on attention networks

- The Trinity of attention: Query, Key, and Value. *Roughly speaking,*
  - Query stores *task* information
  - Key stores information *to search*
  - Value stores information to use *associated with the Key*
- Define a multinoulli distribution to combine values
  - Let $V \in \mathbb{R}^{\phi \times N}$, $Q \in \mathbb{R}^{\rho \times d_k}$ and $K \in \mathbb{R}^{\phi \times d_k}$ where N and d_k are feature sizes. (No batch)
  - ϕ and ρ are the numbers of features (or tokens).
  - $P(V \mid Q, K) := \mathrm{softmax}(\frac{QK^\intercal}{\sqrt{d_k}}) \in \mathbb{R}^{\rho \times \phi}$ [Quiz] what $\sqrt{d_k}$ is for?

Multi-head attention

- Multi-head attention is widely used for its parsimonious.
  - Query and key are embedded by a low-rank matrix where the (at-most) rank is divided by the number of heads.

$$\mathrm{MHAtt}_i(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{W}_{\mathrm{V}_i}^{\intercal}\mathrm{V}\cdot\mathrm{softmax}\left(\frac{\mathrm{K}^{\intercal}\mathrm{W}_{\mathrm{K}_i}\mathrm{W}_{\mathrm{Q}_i}^{\intercal}\mathrm{Q}}{\sqrt{d_k}}\right)$$

$$\mathrm{MHAtt}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{W}_P^{\intercal}\cdot\|_i\,\mathrm{MHAtt}_i(\mathrm{Q},\mathrm{K},\mathrm{V})$$

  - Where || denotes concatenation and $d_k = d_H/g$, hidden size d_H is divided by the number of attention heads g to keep the total number of parameters.
- Roughly speaking, it's a mixture of multinoulli distributions.

Hard, soft-attention and variations

- Hard attention is stochastic and a sampling-based approach
  - Monte Carlo sampling (Xu et al., 2015)
  - Thesholding (Malinowski et al., 2018)
  - Gumbel-softmax (Jang et al., 2017)
- Sparse attention asserts to be a sparse distribution.
  - Sparse transformers (Child et al., 2019)
  - Adaptively sparse transformers (Correia et al., 2019)
- Combining hard and soft attentions (Gang et al., 2022)

Linear attention

- QK computation in the self-attention has $\mathcal{O}(n^2)$ complexity in sequence length n, which becomes prohibitive for long inputs.
- Replace the exponential softmax kernel with a feature map $\sigma(\cdot)$ such that $\mathrm{Att}(Q, K, V) \propto \sigma(Q)\left(\sigma(K)^{\mathsf{T}} V\right)$, reducing the cost to $\mathcal{O}(n)$.
- Performer (Choromanski et al., 2020): Introduced random feature maps to approximate the softmax kernel with variance bounds, setting the foundation for *linear transformers*.
- Recent developments (e.g., SANA, 2024): Explore deterministic kernelizations without random features, integrating Mix-FFN with a new linear DiT, which integrates 3×3 depth-wise convolution into MLP to aggregate the local information of tokens.

Time complexity of linear attention

- Let a positive feature map σ, and omitting embeddings for its simplicity, so that:

$$\text{softmax}(QK^\intercal) \cdot V \approx \frac{\sigma(Q)\big(\sigma(K)^\intercal V\big)}{\sigma(Q)\big(\sigma(K)^\intercal 1\big) + \epsilon}.$$

- The complexity of the numerator is $\mathcal{O}(d_k\phi N + \rho d_k N)$, while the denominator is $\mathcal{O}(d_k\phi + \rho d_k)$ using the *reordering* trick for the row-sum. For the self-attention, we denote $n := \phi = \rho$, and this makes $\mathcal{O}(n)$.
- However, the softmax attention demands $\mathcal{O}(\rho d_k \phi)$, which makes $\mathcal{O}(n^2)$.
- 💁 The implementation of causal attention may need to use `torch.cumsum` for efficient computation of row-sum.

Assumed the sequence length is prohibitively large $\phi, \rho \gg d_k, N$.

Linear vs. softmax

- *Convexity guarantee:* a valid convex combination requires $w_i \geq 0$. Linear attention may exploit ReLU, Softplus, or ELU+1 for σ.
- In linear attention, *stabilize the denominator* by always adding a small amount ϵ.
- Linear attention is *scale-invariant*, while softmax attention is *translation-invariant**. This is why RMSNorm or LayerNorm is used to keep scales tame in the linear attention.
- In softmax, exponential contrast makes them more peaked.
- Softmax never yields exact zeros. When $x_i = 0$ in linear attention, it makes it lie on the simplex boundary (of convex combination). → *winner-take-all*

*The difference originates from the *exponential* operation in softmax.

Bilinear attention networks

- Unlike co-attention, BAN learns a joint representation with two views, simultaneously.
  - A joint probability distribution is defined as:

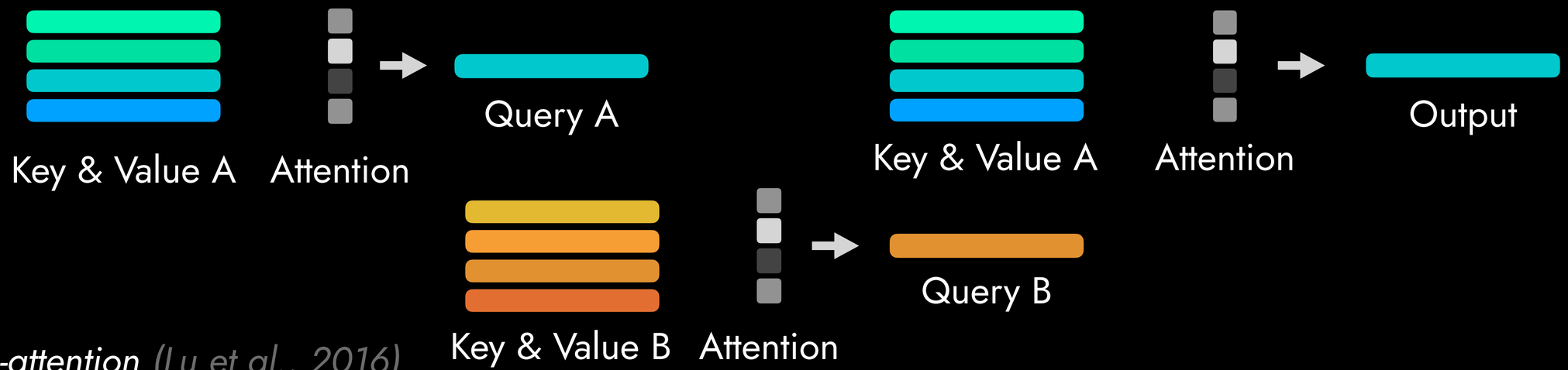
$$\mathcal{A} = \text{softmax}\Big(\big((1 \cdot p^\intercal) \circ X^\intercal U\big)V^\intercal Y\Big) \in \mathbb{R}^{\rho \times \phi}$$

  - A joint representation from two groups of value tokens:

$$f = \text{tr}\Big(\big(X^\intercal U\big)^\intercal \mathcal{A}\big(Y^\intercal V\big)\Big) \in \mathbb{R}^N$$

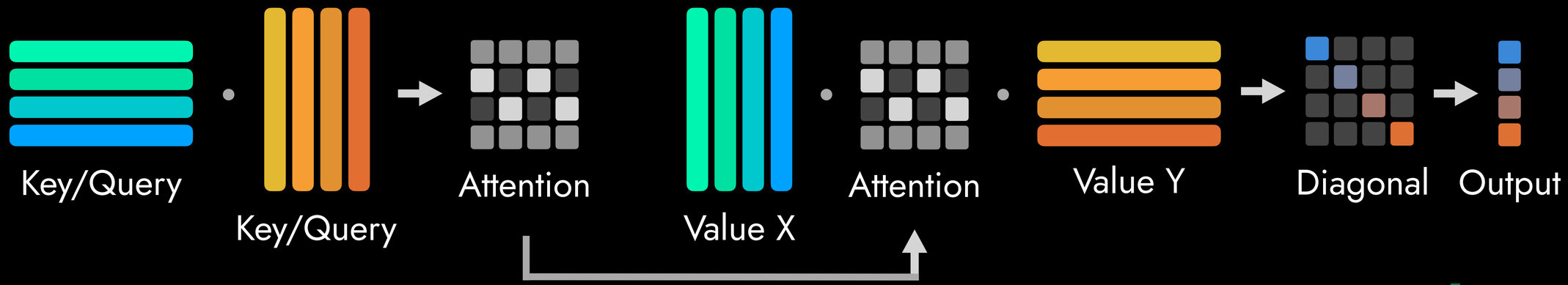
- Nested structure of low-rank bilinear pooling
  - Each attention weight logit is low-rank bilinear pooled as $A_{ij} = p^\intercal\big((U^\intercal X_i) \circ (V^\intercal Y_j)\big)$.
  - Each joint feature is bilinear pooled as $f_k = \sum_{i=1}^{\rho}\sum_{j=1}^{\phi} \mathcal{A}_{ij}(X_i^\intercal U_k)(V_k^\intercal Y_j)$.

Co-attention vs. bilinear attention



Co-attention (Lu et al., 2016)

Bilinear attention (Kim et al., 2018)



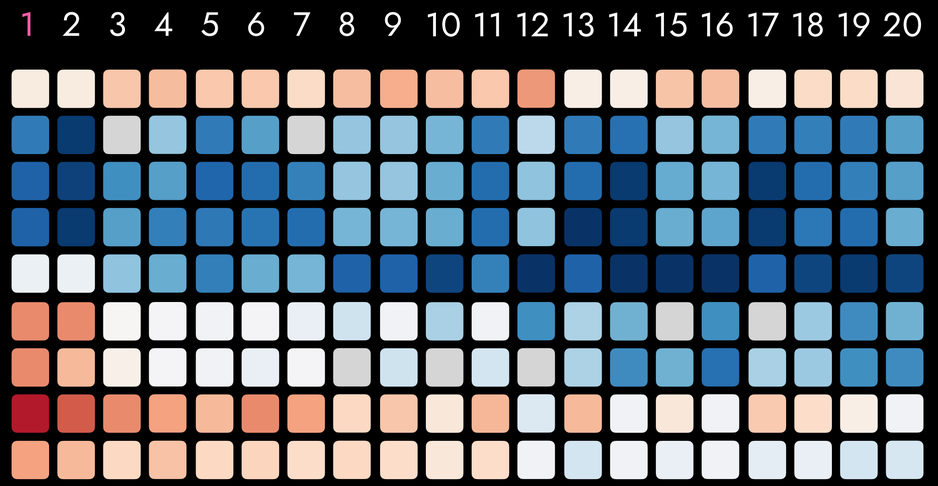
VQA visualization

Q. What color are the pants of the guy skateboarding?

Image

what
color
are
the
pants
of
the
guy
skateboarding

Question

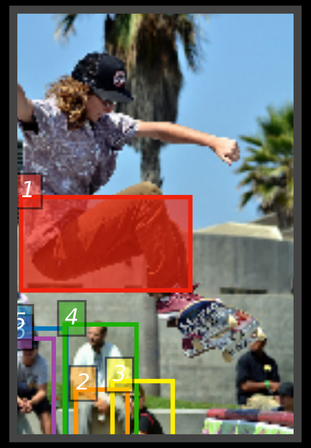


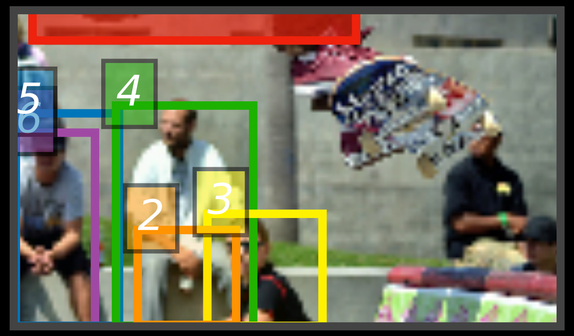
Bilinear attention map $\mathcal{A}$

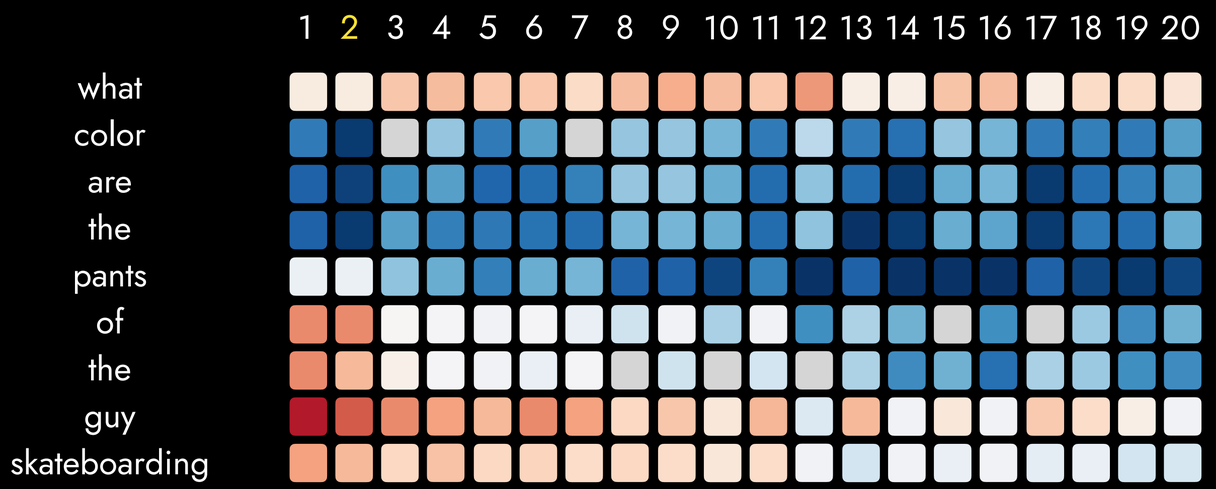
[1] The box order is sorted for visualization.

VQA visualization

Q. What color are the pants of the guy skateboarding?





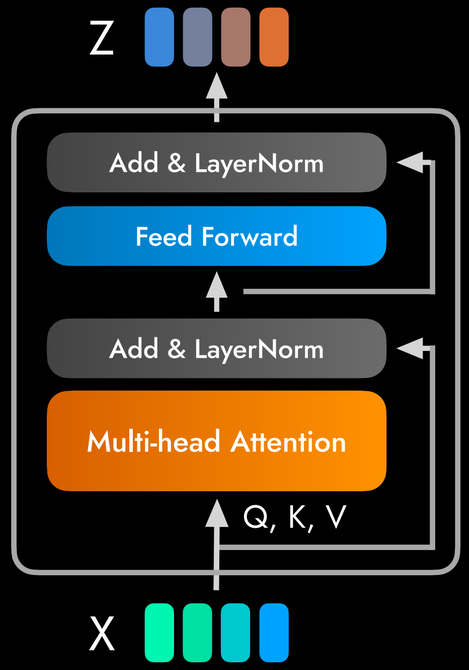


Question

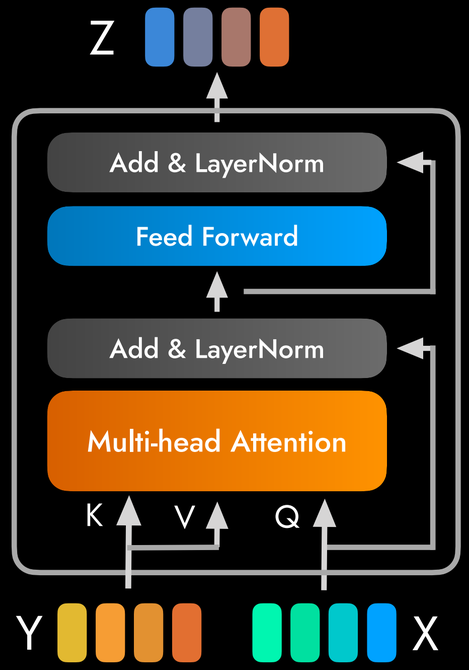
Bilinear attention map $\mathcal{A}$

Transformers with guided-attention

- A curation of self-attention, guided-attention, and MLP with deep learning perks
  - Deep modular co-attention networks (Yu et al., 2019), and *a reference* (Liu et al., 2022)



Self-attention block



Guided-attention block

Layer Normalization

- Ba et al. (2016) proposed a layer that:

$$\mathrm{y} = \frac{\mathrm{x} - E[\mathrm{x}]}{\sqrt{\mathrm{Var}[\mathrm{x}] + \epsilon}} \cdot \gamma + \beta$$

- γ and β are learnable affine transform parameters.
- The normalization step removes the statistical bias in a batch while recovering the global statistics using γ and β.
- The normalization is applied to the last few dimensions (a design choice).

Contrastive learning

- InfoNCE (Oord et al., 2019) as a contrastive loss is defined as:

$$\mathscr{L}_N = -\,\mathbb{E}\left[\log \frac{f(x_i, h)}{\sum_j f(x_j, h)}\right] \quad \text{where} \quad (x_i, h) \text{ is a positive pair.}$$

- Note that regardless of the number of negative samples $N-1$ (see Eqn. 5),

$$f^\star(x_i, h) \propto \frac{p(x_i \mid h)}{p(x_i)} \quad \text{(using Bayes' theorem)}$$

  where * denotes approximation.

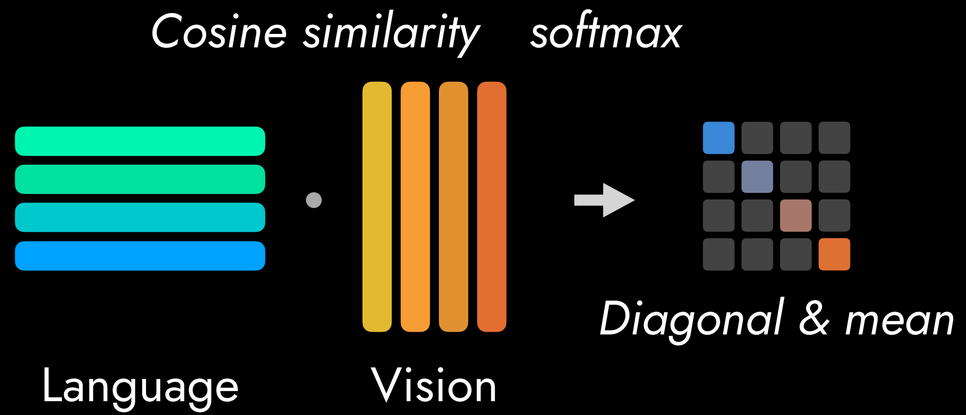
- A negative InfoNCE can *approximate*[1] a lower bound of mutual information:

$$I(x_i; h) \geq \log(N) - \mathscr{L}_N^\star$$

[1]The error is minimizing with a larger N in their proof (see Appendix A.1.)

InfoNCE as contrastive loss

- We choose $f(x, h) = \exp(x^\intercal h)$ as an exponential of cosine similarity.
- Theoretically, a larger N is better. Be careful to multi-gpu implement: 1) *all_gather* does not back-propagate[1], 2) DDP calculates the *mean* of gradients across all processes, not *summation*.



[1]https://amsword.medium.com/gradient-backpropagation-with-torch-distributed-all-gather-9f3941a381f8

GatherLayer in PyTorch

```python
import torch
import torch.distributed as dist

class GatherLayer(torch.autograd.Function):

    """GatherLayer gathers input tensors from all processes.
    """

    @staticmethod
    def forward(ctx, input):
        ctx.save_for_backward(input)
        if dist.is_initialized():
            output = [
                torch.zeros_like(input) for _ in range(dist.get_world_size())
            ]
            dist.all_gather(output, input)
        else:
            output = [input]
        return tuple(output)

    @staticmethod
    def backward(ctx, *grads):
        input, = ctx.saved_tensors

        if dist.is_initialized():
            dist_ops = [
                dist.reduce(grads[i], i, async_op=True)
                for i in range(dist.get_world_size())]

            for op in dist_ops:
                op.wait()

        grad_out = torch.zeros_like(input)
        grad_out[:] = grads[dist.get_rank() if dist.is_initialized() else 0]
        return grad_out
```

forwarding after gathering

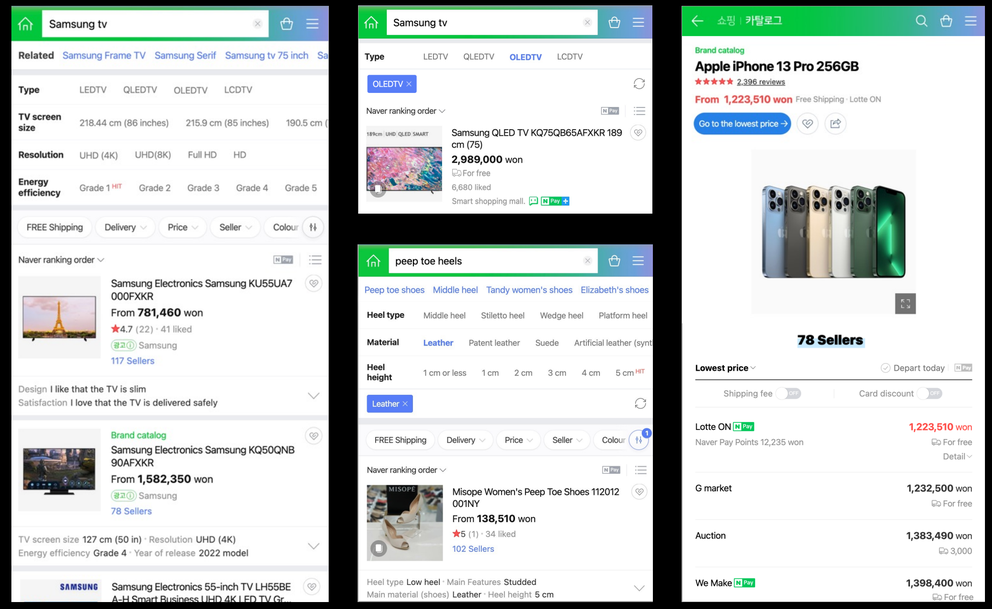
backward with the grad of the current batch

```python
    @staticmethod
    def get_rank_slice(batch_size: int) -> slice:
        """Get a slice to get a portion of gathered tensor corresponding to the
        current rank.

        Args:
            batch_size (int): Local batch size

        Returns:
            slice: slice for the current rank
        """
        rank = dist.get_rank() if dist.is_initialized() else 0
        return slice(rank * batch_size, (rank + 1) * batch_size)
```

The asynchronous operations will maintain the order.

E-CLIP

- Large-scale vision-language representation learning in e-commerce (Shin et al., 2022)
- NAVER Shopping and AI Lab collaborate to build a multimodal backbone for diverse downstream tasks, *i.e., category classification, attribute extraction, product matching, product clustering, and adult product recognition*.

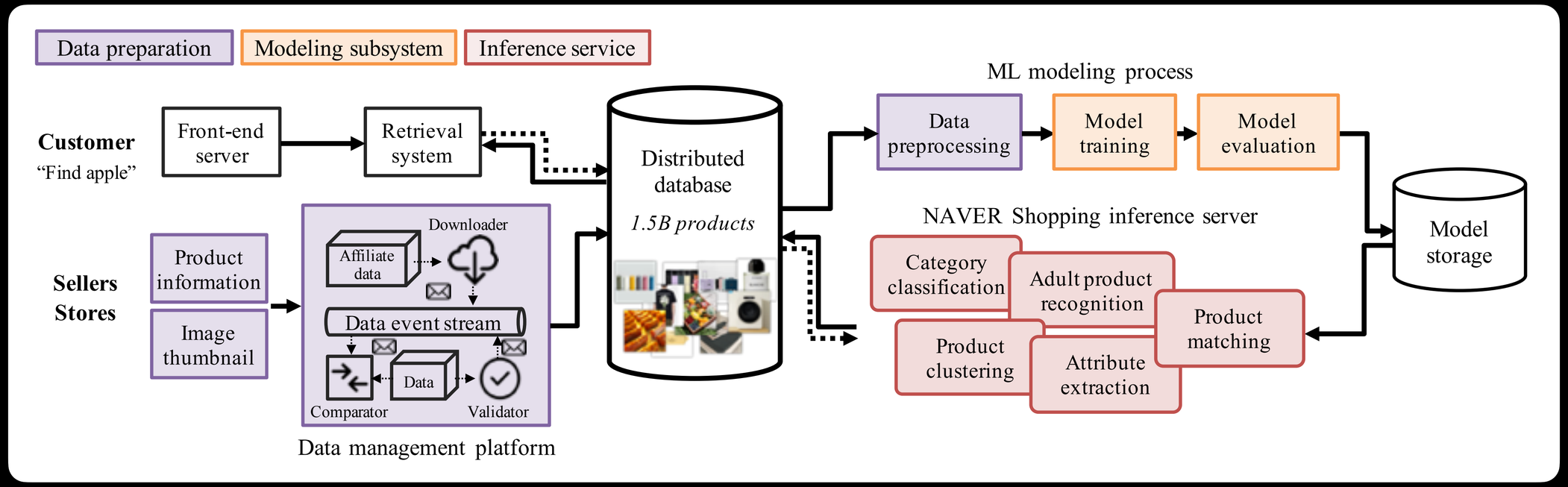


From left, query-based search, filter-based search, and price comparison.

NAVER AI LAB

Data preprocessing

- We filter the 1.5B-scale database to make 330M text-image pairs, removing *invalid*, *duplicated*, and *inappropriate* products to our customer policy.



Multimodal metrics

Multimodal generative models

- Multimodal generative models generate an output in a modality conditioned on the other modal input. (Although there are its variants.)
- Text-to-image generation (text → image) and image captioning (image → text)
- Multimodal representation learning maps two different modalities where the deep neural networks act as a mapping function.

Metrics for text-to-image generation

- Text-to-image generation
  - Inception Score (Salimans et al., 2016)
  - Fréchet Inception Distance (Heusel et al., 2017)
  - R-Precision (Xu et al., 2018), CLIP-R-Precision (Park et al., 2021)
  - Semantic Object Accuracy (YOLOv3) (Hinz et al., 2020)
  - Caption generation (fake → caption → captioning metrics) (Hong et al., 2018)
  - CLIP Similarity

Metrics for image captioning

- Reference-only image caption evaluation
  - BLEU-4 (Papineni et al., 2002): a precision between a candidate and references
  - ROUGE-L (Lin, 2004): a sort of recall
  - METEOR (Banerjee and Lavie, 2005): a word-level alignment
  - CIDEr (Vedantam et al., 2015): n-gram tf-idf weighting and stemming
  - SPICE (Anderson et al., 2016): a semantic parser and scene graph
  - BERTScore (Zhang et al., 2020): a tuned BERT

Metrics for image captioning (Cont'd)

- Reference+image caption evaluation
  - TIGEr (Jiang et al., 2019): a pretrained SCAN model (Lee et al., 2018)
  - ViLBERTScore-F (Lee et al., 2020): a pertained ViLBERT (Lu et al., 2019)
  - RefCLIP-S (Radford et al., 2021)
- Reference-free evaluation
  - Usually for other generation tasks, summarization and dialog
  - VIFIDEL (Madhyastha et al., 2019): an object detector-based for a fixed object vocabulary
  - CLIP-S (Radford et al., 2021)

Rank correlation — metrics for metrics

- Pearson *r* correlation
  - Assume normal distribution, linearity, homoscedasticity[1].
- Spearman rank correlation
  - *Non-parametric*, no assumption about the data distribution, appropriate for ordinal, monotonically related to the other variable. One of factors is the rank difference of corresponding variables
- Kendall rank correlation ✅
  - *Non-parametric*, considering all pairings
  - Spearman and Kendall are not dependent upon the granularity of the integers.
  - "...confidence intervals for Spearman's r_S are less reliable and less interpretable than confidence intervals for Kendall's τ-parameters..." (Kendall & Gibbons, 1990)

[1]Roughly speaking, the same variance or random disturbance is the same across all variables.
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/
Lovie, A. D. (1995). Who discovered Spearman's rank correlation? British Journal of Mathematical and Statistical Psychology, 48(2), 255–269.

Mutual Information Divergence

Kim et al., NeurIPS 2022

Representation measurement

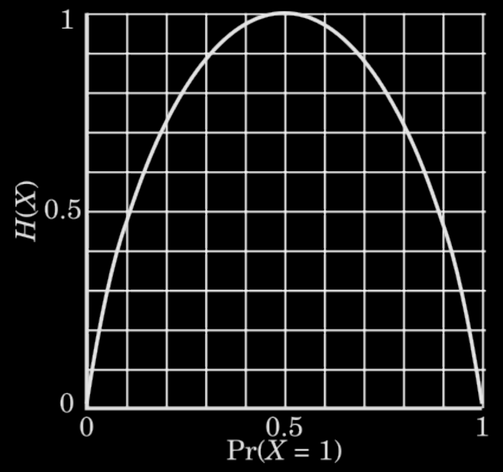
- How to measure the “aligned” multimodal representations?
- In information theory, the *mutual information* of two random variables measures the mutual dependence between the two variables.
  - It quantifies the *information gain* about one random variable by observing the other.
- Viewing the representations as random variables, we measure the quantity of how much they share the multimodal information.

Information entropy

- The *entropy* of a random variable quantifies the *average level of uncertainty* or *information* associated with the variable's potential states or possible outcomes.
- Let a random variable $\mathrm{X} \in \chi$ and the probability $P_X : \chi \to [0,\ 1]$. The entropy is:

$$H(X) = -\sum_{x \in \chi} P_X(x) \log P_X(x) = E_{x \in \chi}\left[-\log P_X(x)\right]$$

- Claude Shannon introduced the concept of information entropy in his 1948 paper *"A Mathematical Theory of Communication."*



Entropy definition

- Information entropy quantifies information defined as:

$$H(X) = -\sum_{x} P_X(x) \log P_X(x)$$

- Mutual information measures the mutual dependence between two variables:

$$I(X;Y) = \sum_{y} \sum_{x} P_{XY}(x,y) \log\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right)$$

- For the continuous random variables,

$$I(X;Y) = \int_{y} \int_{x} P_{XY}(x,y) \log\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right) dxdy$$

Gaussian mutual information

- The general multivariate form of Gaussian distribution for a random D dimensional vector x can be written as:

$$p(x) = \frac{1}{\sqrt{(2\pi)^D \det(\Sigma)}} \exp\left[-\frac{1}{2}(x - \bar{x})^\intercal \Sigma^{-1} (x - \bar{x}) \right]$$

- The Gaussian mutual information is reduced to:

$$I(X; Y) = \frac{1}{2} \log\left(\frac{\det(\Sigma_X) \det(\Sigma_Y)}{\det(\Sigma_Z)} \right)$$

where Z denotes the concatenation of X and Y.

Elements of Information Theory, Cover & Thomas, 2006.

Point-wise mutual information (PMI)

- For sample-wise evaluation, we use PMI defined as:

$$\text{PMI}(x; y) = I(X; Y) + \frac{1}{2}\left(D_M^2(x) + D_M^2(y) - D_M^2(z)\right)$$

  where $D_M^2(x)$ is the squared Mahalanobis distance (SMD) parameterized by μ_X and Σ_X.

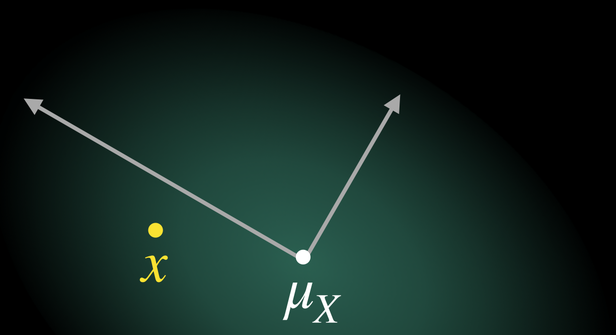
$$D_M^2(x) = (x - \mu_X)^\intercal \Sigma_X^{-1} (x - \mu_X)$$

- The MI is from the normalization terms of the Gaussian distributions, and the SMDs are from the exponential terms in the previous slide's equation.
- Notice that the expectation of the SMD with respect to samples is D for x and y, and $2D$ for $z = [x; y]$.

Mahalanobis distance

- The Mahalanobis distance is a measure of the distance between a point and a probability distribution (Mahalanobis, 1936).
- It is a multivariate generalization of the square of the standard score*: $z = (x - \mu)/\sigma$.

$$D_M^2(x) = (x - \mu_X)^\intercal \Sigma_X^{-1} (x - \mu_X)$$



*Z-score, otherwise.

Expectation of SMD

By the way, the expectation of the squared Mahalanobis distance is the dimension of samples, D.

$$\mathbb{E}_{p(\mathbf{x})} D_M^2(\mathbf{x}) = \frac{1}{N}\mathrm{tr}\left(\mathbf{X}^\intercal \Sigma_{\mathbf{x}}^{-1}\mathbf{X}\right) = \frac{1}{N}\mathrm{tr}\left(\Sigma_{\mathbf{x}}^{-1}\mathbf{X}\mathbf{X}^\intercal\right) = \mathrm{tr}\left(\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{x}}\right) = \mathrm{tr}(\mathbb{I}_D) = D \tag{13}$$

where $\mathbf{X} \in \mathbb{R}^{D\times N}$ is the samples, $\mathbb{I}_D \in \mathbb{R}^{D\times D}$ is the identity matrix. We use the cyclic property of trace where $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$. Therefore, the second term reduces to zero as follows:

$$\frac{1}{2}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[D_M^2(\mathbf{x}) + D_M^2(\mathbf{y}) - D_M^2(\mathbf{z})\right] = \frac{1}{2}(D + D - 2D) = 0. \tag{14}$$

Do CLIP features follow a Gaussian?

- CLIP's embeddings are L2-normalized to lie on the unit hypersphere.
- To separate positives from negatives in the InfoNCE, the model benefits if negatives are spread out evenly (Wang & Isola, 2020).
- In high dimensions, a normalized Gaussian vector is uniformly distributed on the sphere. So, InfoNCE *indirectly* drives embeddings toward behaving like samples from an isotropic Gaussian, normalized.
- Gaussians have the largest entropy among all probability distributions with mean and variance constraints. The most unbiased choice.

Mutual Information Divergence (MID)

- Kim et al. (2022) proposed the expectation of PMI w.r.t. the evaluating sample $(\hat{x}, \hat{y})$, measuring the divergence from the ground-truth or reference samples (X, Y).
- It is the *negative* cross-mutual information (Bugliarello et al., 2020) for Gaussian.

$$\mathbb{E}_{(\hat{x},\hat{y})\sim\mathscr{D}}\text{PMI}(\hat{x};\hat{y}) = I(X;Y) + \frac{1}{2}\mathbb{E}_{(\hat{x},\hat{y})\sim\mathscr{D}}\left[D_M^2(\hat{x}) + D_M^2(\hat{y}) - D_M^2(\hat{z})\right]$$

  where $(\hat{x}, \hat{y})$ denotes a pair of evaluating samples, respectively, and $\hat{z} = [\hat{x}; \hat{y}]$. Here, $D_M^2(\hat{x})$ is parameterized by μ_X and Σ_X, the reference statistics.
- They denoted this as ***Mutual Information Divergence (MID)***, comparable to the *FID*.

*By the way, the cross entropy is defined as $H(p, q) = -\mathbb{E}_p[\log q] = H(p) + D_{KL}(p\|q)$.

Bias and variance decomposition

- The expectation of SMD w.r.t. $(\hat{x}, \hat{y})$ can be decomposed into bias and variance.

The expectation of PMI with respect to evaluating samples needs to calculate the expectation of three terms of the squared Mahalanobis distances (SMD) with respect to the evaluating sample $\hat{\mathbf{x}}$. With a notation of $\hat{\mathbf{X}} \in \mathbb{R}^{D \times N}$ for evaluation samples, we can decompose the expectation of SMD with two terms of bias and variance as follows:

$$\mathbb{E}_{\hat{\mathbf{x}}}\left[D_M^2(\hat{\mathbf{x}})\right] = \frac{1}{N}\mathrm{tr}\left((\hat{\mathbf{X}} - \mu_{\mathbf{x}}\mathbb{1}^\intercal)^\intercal \Sigma_{\mathbf{x}}^{-1}(\hat{\mathbf{X}} - \mu_{\mathbf{x}}\mathbb{1}^\intercal)\right) \tag{16}$$

$$= \frac{1}{N}\mathrm{tr}\left(\Sigma_{\mathbf{x}}^{-1}(\hat{\mathbf{X}} - \mu_{\mathbf{x}}\mathbb{1}^\intercal)(\hat{\mathbf{X}} - \mu_{\mathbf{x}}\mathbb{1}^\intercal)^\intercal\right) \tag{17}$$

$$= \frac{1}{N}\mathrm{tr}\left(\Sigma_{\mathbf{x}}^{-1}\left(\hat{\mathbf{X}}\hat{\mathbf{X}}^\intercal - N\mu_{\hat{\mathbf{x}}}\mu_{\hat{\mathbf{x}}}^\intercal + N(\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}})(\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}})^\intercal\right)\right) \tag{18}$$

$$= \mathrm{tr}\left(\Sigma_{\mathbf{x}}^{-1}\left(\Sigma_{\hat{\mathbf{x}}} + (\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}})(\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}})^\intercal\right)\right) \tag{19}$$

$$= (\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}})^\intercal \Sigma_{\mathbf{x}}^{-1}(\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}}) + \mathrm{tr}(\Sigma_{\mathbf{x}}^{-1}\Sigma_{\hat{\mathbf{x}}}) \tag{20}$$

$$= (\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}})^\intercal \Sigma_{\mathbf{x}}^{-1}(\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}}) + \mathrm{tr}(\Sigma_{\mathbf{x}}^{-1}\Sigma_{\hat{\mathbf{x}}}) - \mathrm{tr}(\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{x}}) + \mathrm{tr}(\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{x}}) \tag{21}$$

$$= (\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}})^\intercal \Sigma_{\mathbf{x}}^{-1}(\mu_{\hat{\mathbf{x}}} - \mu_{\mathbf{x}}) + \mathrm{tr}\left(\Sigma_{\mathbf{x}}^{-1}(\Sigma_{\hat{\mathbf{x}}} - \Sigma_{\mathbf{x}})\right) + D$$

binomial theorem + dummy terms and using $\hat{X}1 = N\mu_{\hat{x}}$

Cyclic property of trace and the trace of a scalar

dummy terms

Trace of an identity matrix

where $\mathbb{1} \in \mathbb{R}^N$ is a vector of ones. Remind that the expectation of SMD is D when the evaluating samples $\hat{\mathbf{x}}$ are following the distribution of $\mathbf{x}$ in Equation 13. However, the above equation shows that if the mean or covariance of $\hat{\mathbf{x}}$ deviates from $\mathbf{x}$, the result may be smaller or larger than D.

Relation to KL divergence

The proposed method MID is related to Kullback-Leibler divergence (or relative entropy). Let $\mathcal{N}_0(\mu_0, \Sigma_0)$ and $\mathcal{N}_1(\mu_1, \Sigma_1)$ are two multivariate normal distributions having the same dimension of D, then the Kullback-Leibler divergence between the distributions is as follows [47]:

$$D_{\mathrm{KL}}\left(\mathcal{N}_0 \parallel \mathcal{N}_1\right) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}(\Sigma_0 - \Sigma_1)\right) + (\mu_1 - \mu_0)^{\mathsf{T}} \Sigma_1^{-1} (\mu_1 - \mu_0) + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right).$$

the previous bias and variance decomposition

Using the above equation and Equation 22, we rearrange Equation 12 as follows:

PMI

$$\begin{aligned}
\mathbb{E}_{(\hat{\mathbf{x}},\hat{\mathbf{y}})\sim\mathcal{D}}\mathrm{PMI}(\hat{\mathbf{x}};\hat{\mathbf{y}}) =& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\left(\log\left(\frac{\det \Sigma_{\mathbf{x}}}{\det \Sigma_{\hat{\mathbf{x}}}}\right) + \log\left(\frac{\det \Sigma_{\mathbf{y}}}{\det \Sigma_{\hat{\mathbf{y}}}}\right) - \log\left(\frac{\det \Sigma_{\mathbf{z}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right)\right) \qquad (23) \\
=& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\log\left(\frac{\det \Sigma_{\mathbf{x}} \det \Sigma_{\mathbf{y}}}{\det \Sigma_{\mathbf{z}}}\right) + \frac{1}{2}\log\left(\frac{\det \Sigma_{\hat{\mathbf{x}}} \det \Sigma_{\hat{\mathbf{y}}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right) \qquad (24) \\
=& I(\hat{\mathbf{X}};\hat{\mathbf{Y}}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \qquad (25)
\end{aligned}$$

Rearrange log terms for the MIs

where $D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) = 0$ since $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the same condition evaluating generations.

Relation to KL divergence

The proposed method MID is related to Kullback-Leibler divergence (or relative entropy). Let $\mathcal{N}_0(\mu_0, \Sigma_0)$ and $\mathcal{N}_1(\mu_1, \Sigma_1)$ are two multivariate normal distributions having the same dimension of D, then the Kullback-Leibler divergence between the distributions is as follows [47]:

$$D_{\mathrm{KL}}\left(\mathcal{N}_0 \parallel \mathcal{N}_1\right) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}(\Sigma_0 - \Sigma_1)\right) + (\mu_1 - \mu_0)^{\mathsf{T}} \Sigma_1^{-1} (\mu_1 - \mu_0) + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right).$$

Using the above equation and Equation 22, we rearrange Equation 12 as follows:

$$\begin{aligned}
\mathbb{E}_{(\hat{\mathbf{x}},\hat{\mathbf{y}})\sim\mathcal{D}}\mathrm{PMI}(\hat{\mathbf{x}};\hat{\mathbf{y}}) =& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\left(\log\left(\frac{\det \Sigma_{\mathbf{x}}}{\det \Sigma_{\hat{\mathbf{x}}}}\right) + \log\left(\frac{\det \Sigma_{\mathbf{y}}}{\det \Sigma_{\hat{\mathbf{y}}}}\right) - \log\left(\frac{\det \Sigma_{\mathbf{z}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right)\right) \qquad (23) \\
=& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\log\left(\frac{\det \Sigma_{\mathbf{x}} \det \Sigma_{\mathbf{y}}}{\det \Sigma_{\mathbf{z}}}\right) + \frac{1}{2}\log\left(\frac{\det \Sigma_{\hat{\mathbf{x}}} \det \Sigma_{\hat{\mathbf{y}}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right) \qquad (24) \\
=& I(\hat{\mathbf{X}};\hat{\mathbf{Y}}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \qquad (25)
\end{aligned}$$

Eliminate the MI of references

where $D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) = 0$ since $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the same condition evaluating generations.

Relation to KL divergence

The proposed method MID is related to Kullback-Leibler divergence (or relative entropy). Let $\mathcal{N}_0(\mu_0, \Sigma_0)$ and $\mathcal{N}_1(\mu_1, \Sigma_1)$ are two multivariate normal distributions having the same dimension of D, then the Kullback-Leibler divergence between the distributions is as follows [47]:

$$D_{\mathrm{KL}}\left(\mathcal{N}_0 \parallel \mathcal{N}_1\right) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}(\Sigma_0 - \Sigma_1)\right) + (\mu_1 - \mu_0)^{\mathsf{T}} \Sigma_1^{-1} (\mu_1 - \mu_0) + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right).$$

Using the above equation and Equation 22, we rearrange Equation 12 as follows:

$$\begin{aligned}
\mathbb{E}_{(\hat{\mathbf{x}},\hat{\mathbf{y}})\sim\mathcal{D}}\mathrm{PMI}(\hat{\mathbf{x}};\hat{\mathbf{y}}) =& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\left(\log\left(\frac{\det \Sigma_{\mathbf{x}}}{\det \Sigma_{\hat{\mathbf{x}}}}\right) + \log\left(\frac{\det \Sigma_{\mathbf{y}}}{\det \Sigma_{\hat{\mathbf{y}}}}\right) - \log\left(\frac{\det \Sigma_{\mathbf{z}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right)\right) \qquad (23) \\
=& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + \cancel{D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y}))} - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\log\left(\frac{\det \Sigma_{\mathbf{x}} \det \Sigma_{\mathbf{y}}}{\det \Sigma_{\mathbf{z}}}\right) + \frac{1}{2}\log\left(\frac{\det \Sigma_{\hat{\mathbf{x}}} \det \Sigma_{\hat{\mathbf{y}}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right) \qquad (24) \\
=& I(\hat{\mathbf{X}};\hat{\mathbf{Y}}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \qquad (25)
\end{aligned}$$

where $D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) = 0$ since $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the same condition evaluating generations.

Relation to KL divergence

The proposed method MID is related to Kullback-Leibler divergence (or relative entropy). Let $\mathcal{N}_0(\mu_0, \Sigma_0)$ and $\mathcal{N}_1(\mu_1, \Sigma_1)$ are two multivariate normal distributions having the same dimension of D, then the Kullback-Leibler divergence between the distributions is as follows [47]:

$$D_{\mathrm{KL}}\left(\mathcal{N}_0 \parallel \mathcal{N}_1\right) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}(\Sigma_0 - \Sigma_1)\right) + (\mu_1 - \mu_0)^{\mathsf{T}} \Sigma_1^{-1} (\mu_1 - \mu_0) + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right).$$

Using the above equation and Equation 22, we rearrange Equation 12 as follows:

$$\begin{aligned}
\mathbb{E}_{(\hat{\mathbf{x}},\hat{\mathbf{y}})\sim\mathcal{D}}\mathrm{PMI}(\hat{\mathbf{x}};\hat{\mathbf{y}}) =& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\left(\log\left(\frac{\det \Sigma_{\mathbf{x}}}{\det \Sigma_{\hat{\mathbf{x}}}}\right) + \log\left(\frac{\det \Sigma_{\mathbf{y}}}{\det \Sigma_{\hat{\mathbf{y}}}}\right) - \log\left(\frac{\det \Sigma_{\mathbf{z}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right)\right) \qquad (23) \\
=& I(\mathbf{X};\mathbf{Y}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) + D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z})) \\
& - \frac{1}{2}\log\left(\frac{\det \Sigma_{\mathbf{x}} \det \Sigma_{\mathbf{y}}}{\det \Sigma_{\mathbf{z}}}\right) + \frac{1}{2}\log\left(\frac{\det \Sigma_{\hat{\mathbf{x}}} \det \Sigma_{\hat{\mathbf{y}}}}{\det \Sigma_{\hat{\mathbf{z}}}}\right) \qquad (24) \\
=& I(\hat{\mathbf{X}};\hat{\mathbf{Y}}) + D_{\mathrm{KL}}(p(\hat{\mathbf{x}}) \parallel p(\mathbf{x})) - D_{\mathrm{KL}}(p(\hat{\mathbf{z}}) \parallel p(\mathbf{z}))
\end{aligned}$$

The MI of evaluating samples

where $D_{\mathrm{KL}}(p(\hat{\mathbf{y}}) \parallel p(\mathbf{y})) = 0$ since $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the same condition evaluating generations.

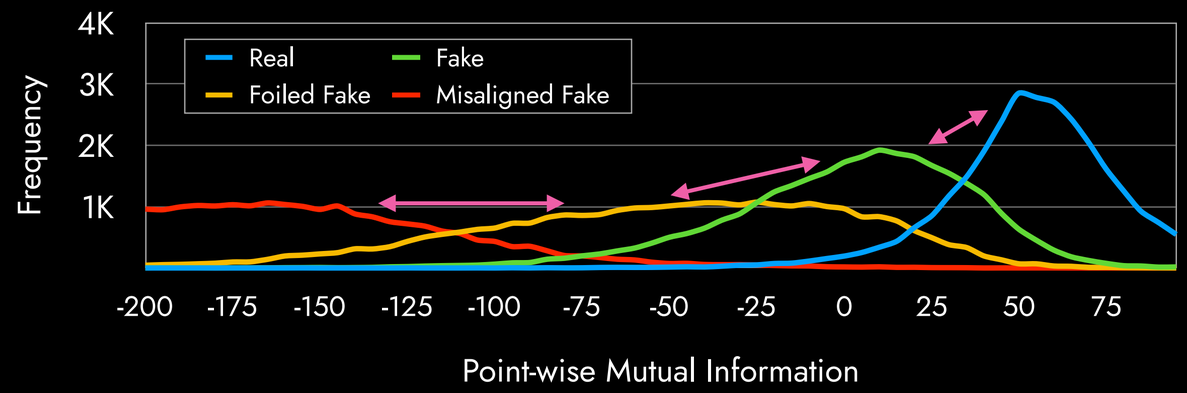
Human-free judgments

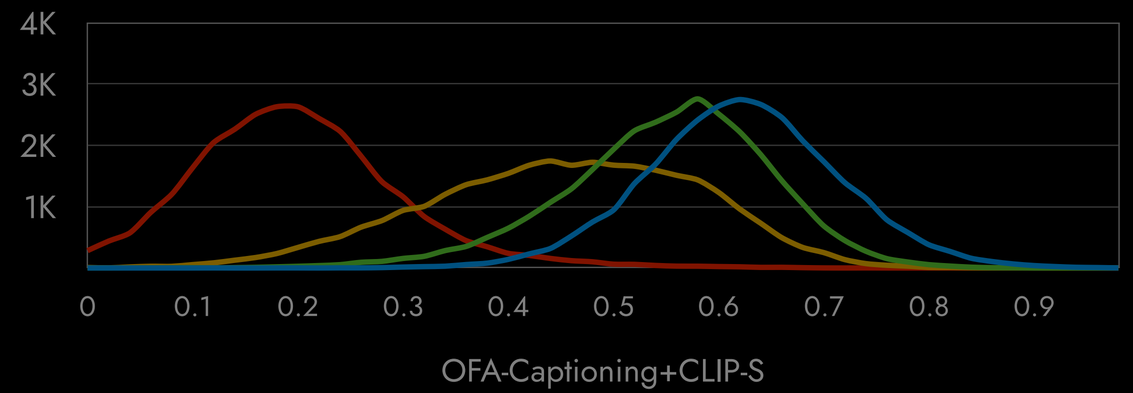
- For a given GT caption, we collect four images with a 1-4 scale judgment.
  - 4: The real image aligned with the given GT caption
  - 3: a generated image from the GT caption
  - 2: a generated image from a FOIL caption
  - 1: a random generated image
- Assumptions
  - Generated image is not better than real image.
  - Generated image from a FOIL caption is not better than the one from the GT caption.
  - A random image is not better than the generated images from the GT or FOIL captions.

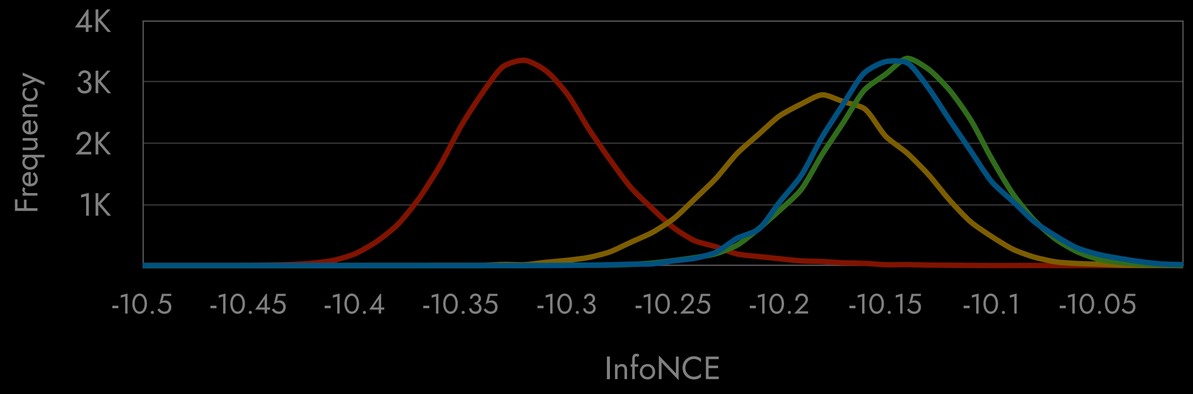
*People riding bicycles down the road approaching a **dog**.*

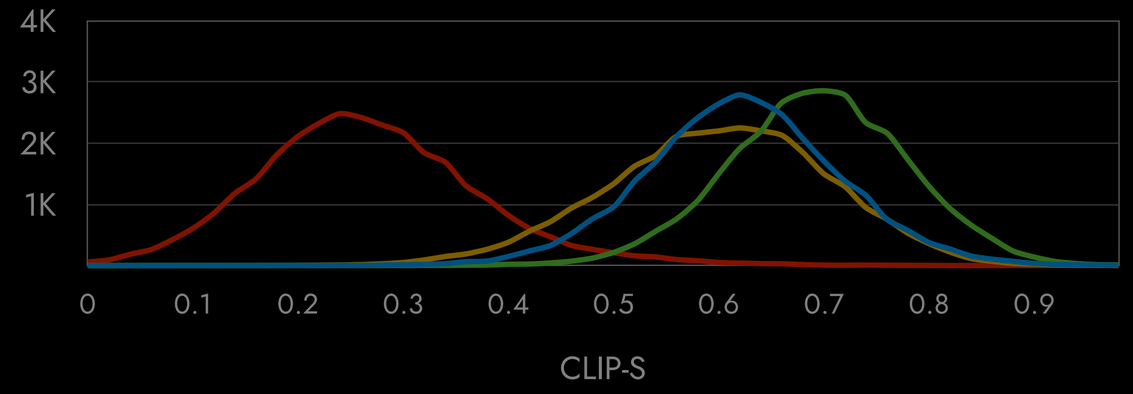
⚠ ***Foiled!***

Score distributions



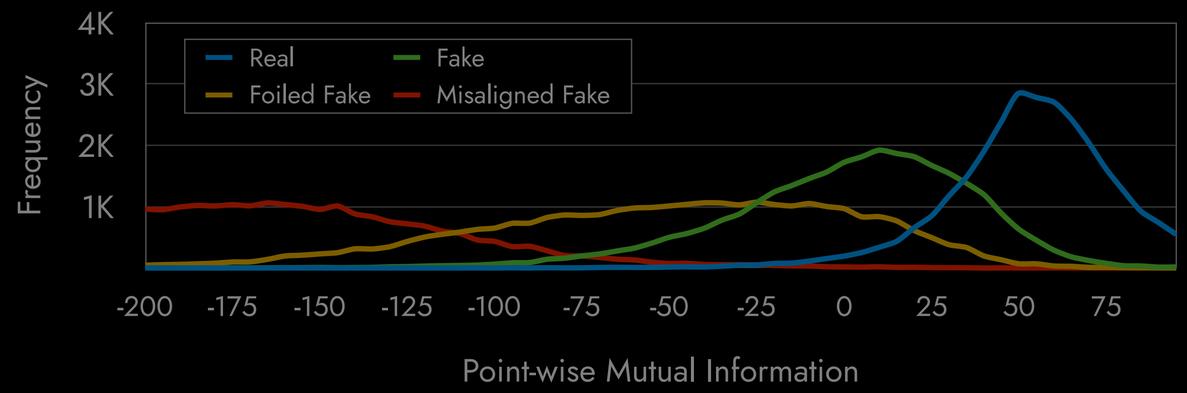


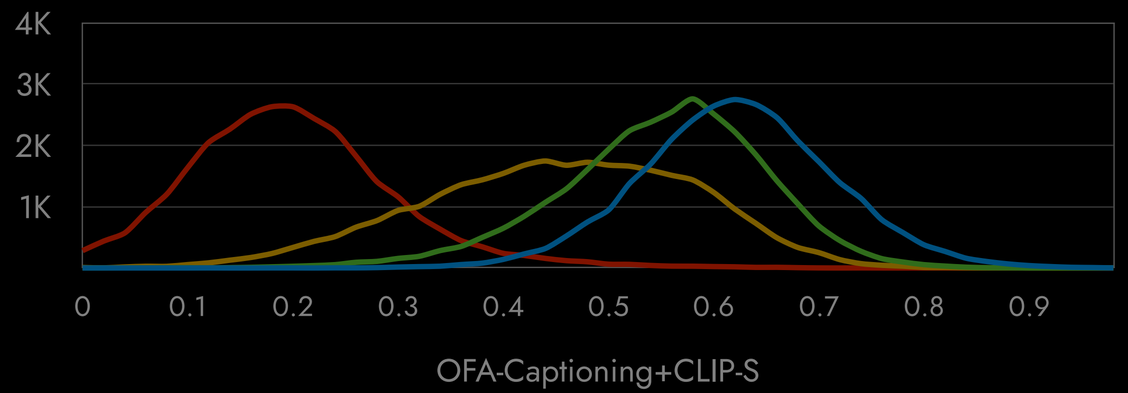


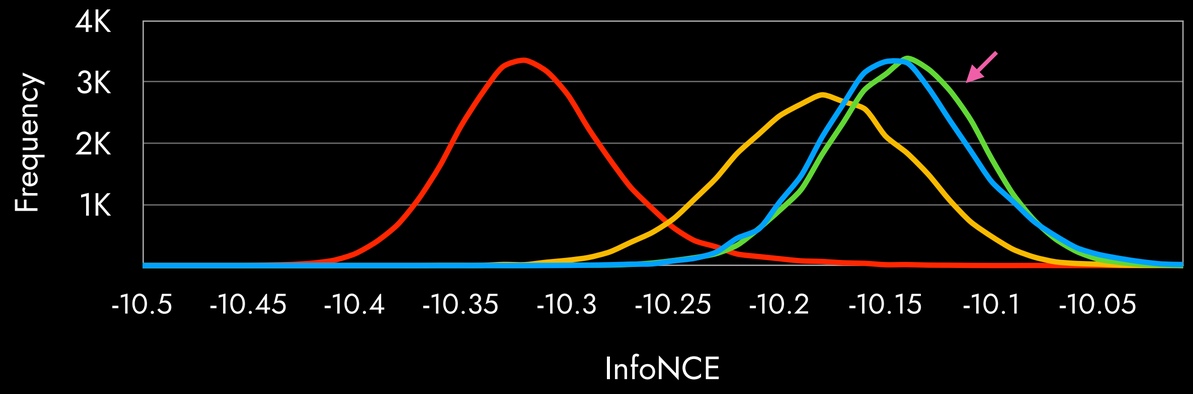


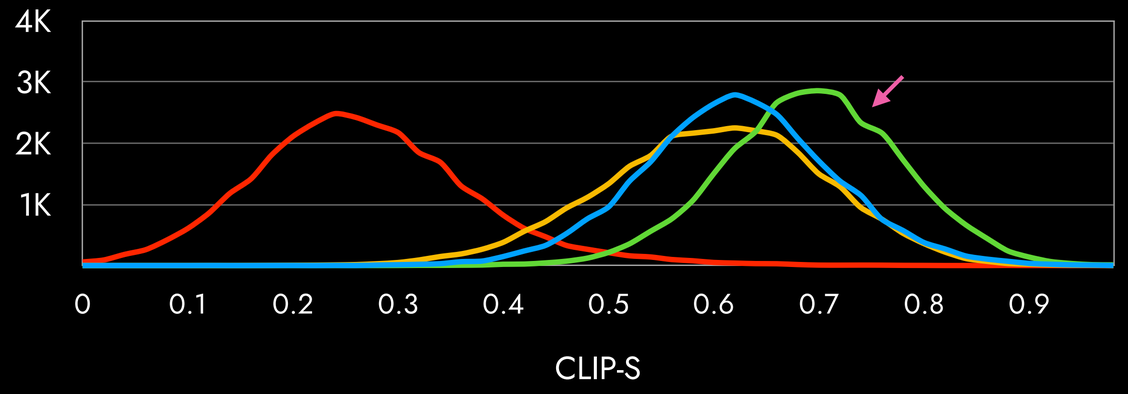
Judgment correlations for each metric can be found in Table 1 (LAFITE) and 8 (VQ-Diffusion).

Score distributions (cont'd)





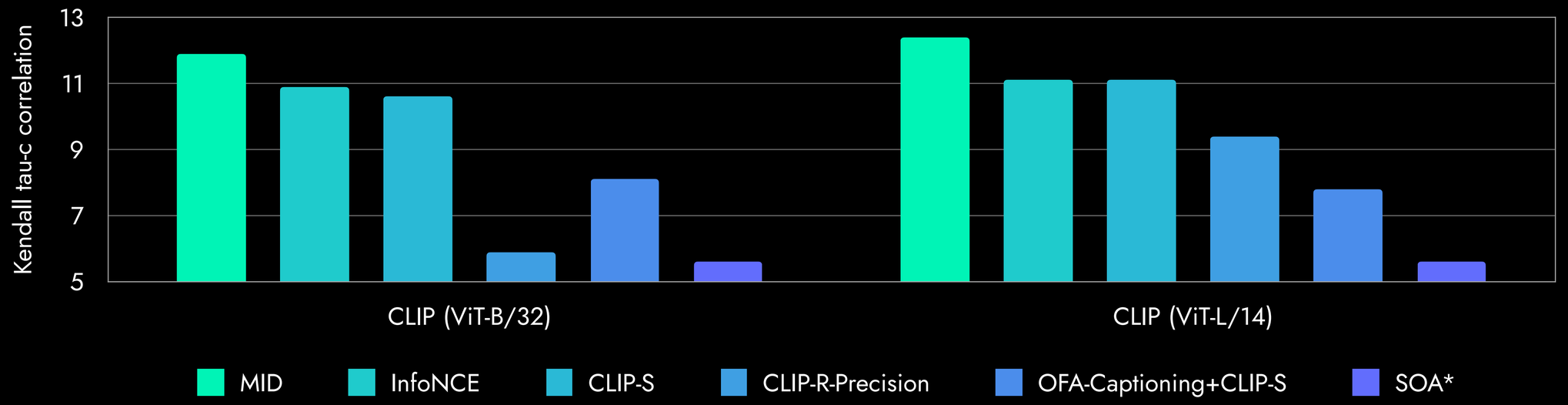




Metrics' overfitting to fake image can be problematic. (Dinh et al., 2021)

Human evaluation for text-to-image models

- The human judgment correlation only for the fake images from GT caption, which is more difficult than the previous generated benchmark.



All results are statistically significant ($p < 0.001$). Table 2 for more details. *SOA used YOLO-v3.

Human evaluations

Flickr8K-Expert Human Judgment Correlation

	Kendall tau_c
BLEU-1	32.3
BLEU-4	30.8
ROUGE-L	32.3
BERT-S (RoBERTa-F)	39.2
METEOR	41.8
CIDEr	43.9
SPICE	44.9
LEIC (τb) (Cui et al. 2018)	46.6
BERT-S++ (Yi et al. 2020)	46.7
TIGEr (Jiang et al. 2019)	49.3
NUBIA (Kane et al. 2020)	49.5
ViLBERTScore-F (Lee et al. 2020)	50.1
CLIP-S (no refs)	51.2
RefCLIP-S	53.0
MID (ours)	**54.9**

Flickr8K-CF Human Judgment Correlation

	Kendall tau_b
BLEU-4	16.9
CIDEr	24.6
METEOR	22.2
ROUGE-L	19.9
SPICE	24.4
BERT-S (RoBERTa-F)	22.8
LEIC	29.5
CLIP-S (no refs)	34.4
RefCLIP-S	36.4
MID (ours)	**37.3**

All metrics use 4-5 ground truth references, except for CLIP-S (which uses none).

Human evaluations

Pascal-50S

	HC	HI	HM	MM	Mean
Length-bias	51.7	52.3	63.6	49.6	54.3
BLEU-4	60.4	90.6	84.9	54.7	72.6
SPICE	63.6	96.3	86.7	68.3	78.7
METEOR	63.8	97.7	93.7	65.4	80.1
ROUGE-L	63.7	95.3	92.3	61.2	78.1
CIDEr	65.1	98.1	90.5	64.8	79.6
BERT-S (RoBERTa-F)	65.4	96.2	93.3	61.4	79.1
TIGEr	56.0	**99.8**	92.8	74.2	80.7
ViLBERTScore-F	49.9	99.6	93.1	75.8	79.6
BERT-S++	65.4	98.1	96.4	60.3	80.1
CLIP-S (no refs)	56.5	99.3	96.4	70.4	80.7
RefCLIP-S	64.5	99.6	95.4	72.8	83.1
MID (ours)	**67.0**	99.7	**97.4**	**76.8**	**85.2**

FOIL

	1-ref	4-ref
Length-bias	50.2	50.2
BLEU-4	66.5	82.6
METEOR	78.8	85.4
ROUGE-L	71.7	79.3
CIDEr	82.5	90.6
SPICE	75.5	86.1
BERT-S (RoBERTa-F)	88.6	92.1
CLIP-S (no refs)	87.2	87.2
RefCLIP-S	**91.0**	**92.6**
MID (ours)	90.5	90.5

Report the average of results with five randomly-sampled references, except for CLIP-S (which uses none).

FOIL visualization

FO: A tall **book** tower with people walking down a city street. (CLIP-S/RefCLIP-S/PMI=.737/.804/16.9)
GT: A tall **clock** tower with people walking down a city street. (.738/.806/20.0)

FO: Some people a **chair** some bananas and plastic cups. (.747/.832/8.63)
GT: Some people a **table** some bananas and plastic cups. (.756/.843/12.0)

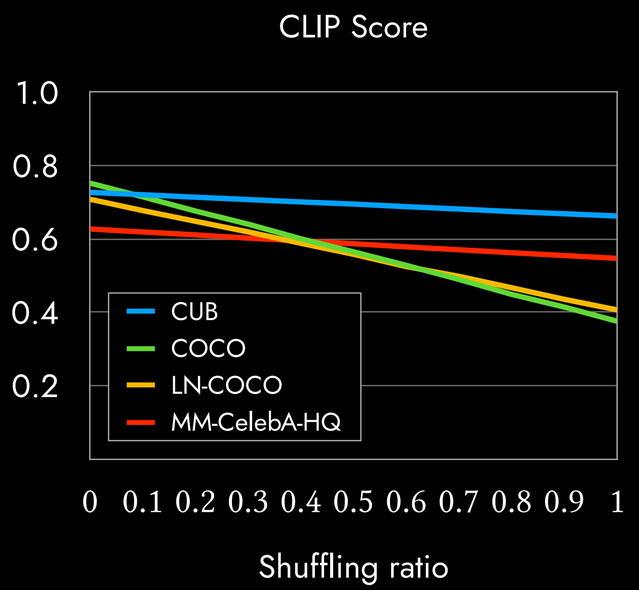
FO: Large bowls of **broccoli** bunches being examined by a female buyer. (.774/.811/11.8)
GT: Large bowls of **banana** bunches being examined by a female buyer. (.761/.815/9.61)

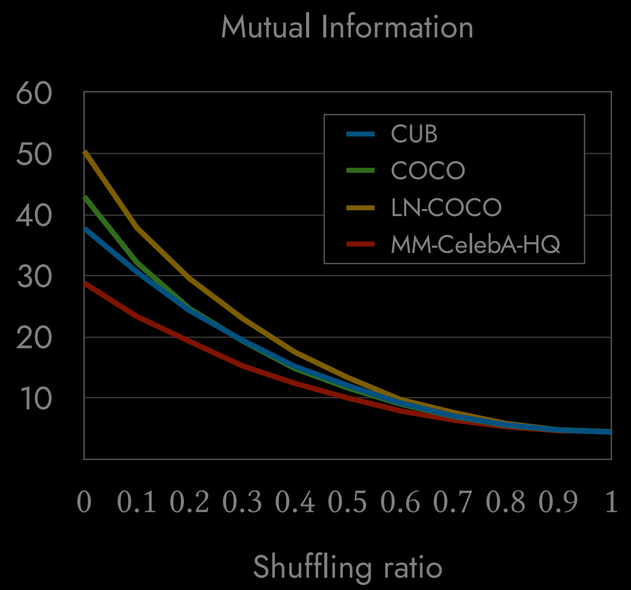
FO: a **giraffe** grazing on grass in an open field. (.707/.747/-4.89)
GT: A **zebra** grazing on grass in an open field. (.718/.821/38.6)

- The first two columns show the corrected examples, while the third column shows an example that CLIP-S and MID failed to detect.
- RefCLIP-S directly exploits the reference captions where the counterpart of the foiled word appears. One of the references of the third example was that ``a woman is picking *bananas* from a basket.''
- The fourth example shows that MID can be negative for unlikely samples since it is based on the definition of differential entropy.

Robustness toward domains

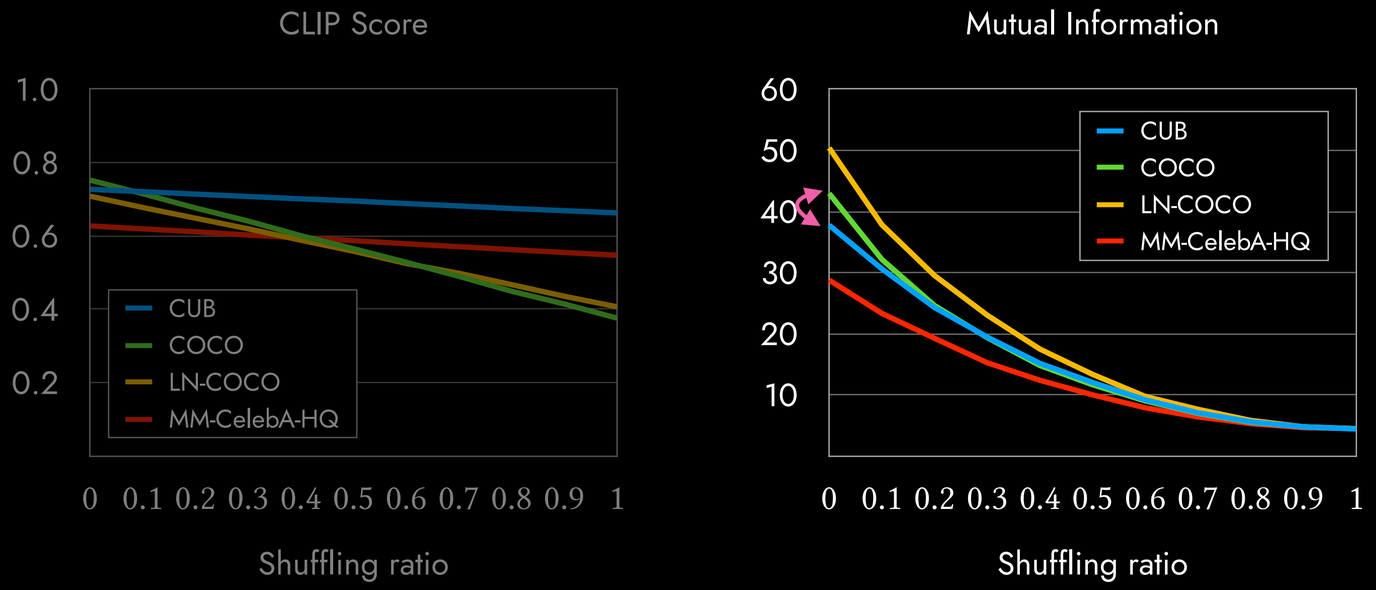
- Randomly shuffled text-image alignments to see the change of scores. Narrow domains (*e.g.*, CUB and CelebA) have small changes for the CLIP scores.





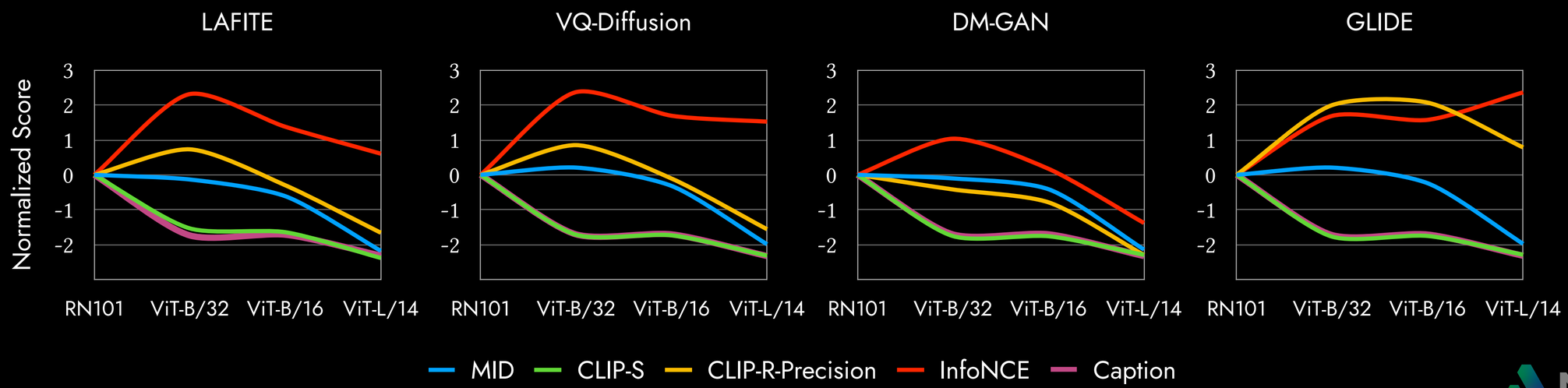
Robustness toward domains

- Randomly shuffled text-image alignments to see the change of scores. Narrow domains (*e.g.*, CUB and CelebA) have small changes for the CLIP scores.
- However, mutual information gives the better normalized scores across various domains (see CUB and COCO curves on the right).



Robustness toward CLIPs

- LAFITE (Zhou et al., 2021) used textual and visual encoders of the pre-trained CLIP ViT-B/32, VQ-Diffusion (Gu et al., 2021) used text encoder, while DM-GAN (Zhu et al., 2019) used none. GLIDE (Nichol et al., 2021) used their noised ViT-L CLIP.
- Notice that InfoNCE and CLIP-R-Precision are related to contrastive training loss.
- Although CLIPs impact on the performance of metrics, MID is the most stable metric.



Representation diversity

- The covariance matrix of features can be the ground for representation diversity.
- If the representations $X \in \mathbb{R}^{F \times N}$ are unbiased (mean is zero) and l2-normalized, the sum of all eigenvalues $\{\lambda_i\}$ of the covariance matrix is one. The proof is as follows:

$$\sum_i \lambda_i = \text{tr}(\Sigma) = \text{tr}(XX^\intercal / N) = 1$$

  where $\Sigma_{ii} = 1/N$ due to the definition of l2-normalization.

- And, since Σ is (semi-)positive definite, $\lambda_i \geq 0$. So, using the eigenvalues, we can define the *probability distribution over dimensions* (Friedman & Dieng, 2022).

Trace of a product: $\text{tr}(\text{A}^\intercal\text{B}) = \text{tr}(\text{AB}^\intercal) = \text{tr}(\text{B}^\intercal\text{A}) = \text{tr}(\text{BA}^\intercal)$.

Representation diversity (Cont'd)

- The entropy of eigenvalues gives you the sense of feature diversity.
  - The higher entropy is the more diversity, vice versa.
- For the multimodal representation pairs, we may use the joint (concatenated) features to get the covariance matrix.
- The Vendi Score (Friedman & Dieng, 2022) measures the feature diversity using the below definition:

$$\mathrm{VS}(x_1, \cdots, x_N) = \exp\left(-\sum_{i=1}^{F} \lambda_i \log \lambda_i\right)$$

Remarks

- Cross-modality zero-shot capability is worth to explore.
- Vision-language joint representation learning is moving from multimodal fusion to vision-language pre-training thanks to computing power and big data.
- We can measure the multimodal representations using Gaussian feature assumption in aspect of multimodal alignment and their diversity.

“The only thing that overcomes hard luck is hard work.”

—Harry Golden