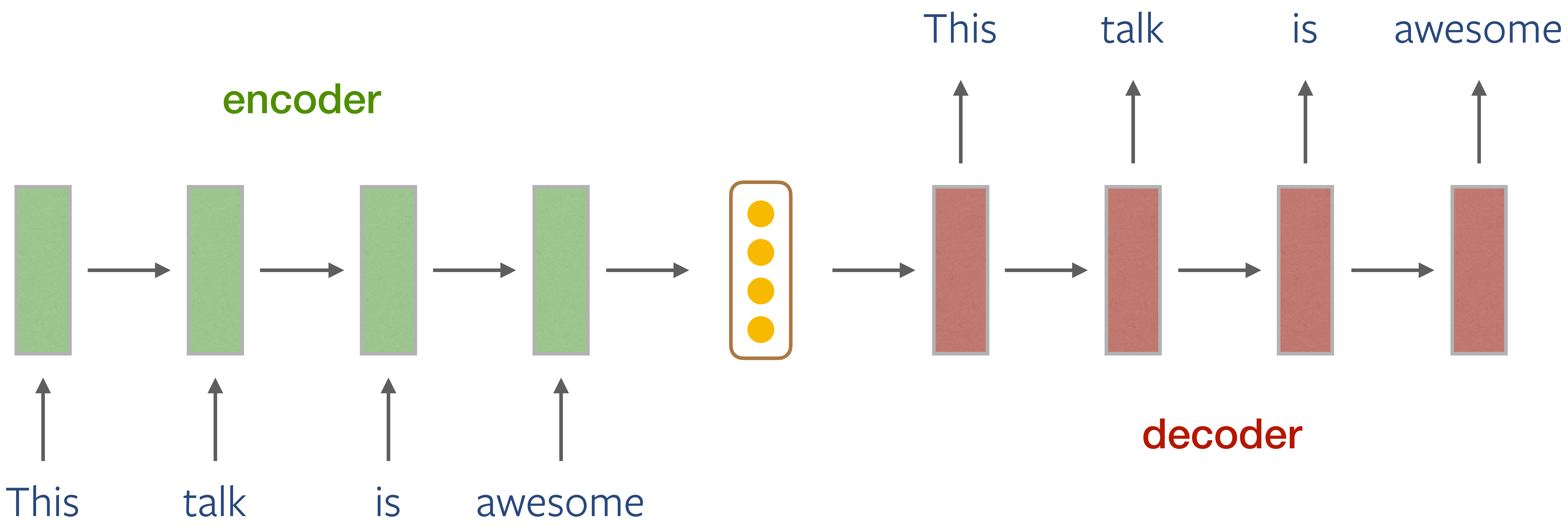
Educating Text Autoencoders: Latent Representation Guidance via Denoising

Tianxiao Shen   Jonas Mueller   Regina Barzilay   Tommi Jaakkola

Text Autoencoders

Represent sentences as vectors in a latent space

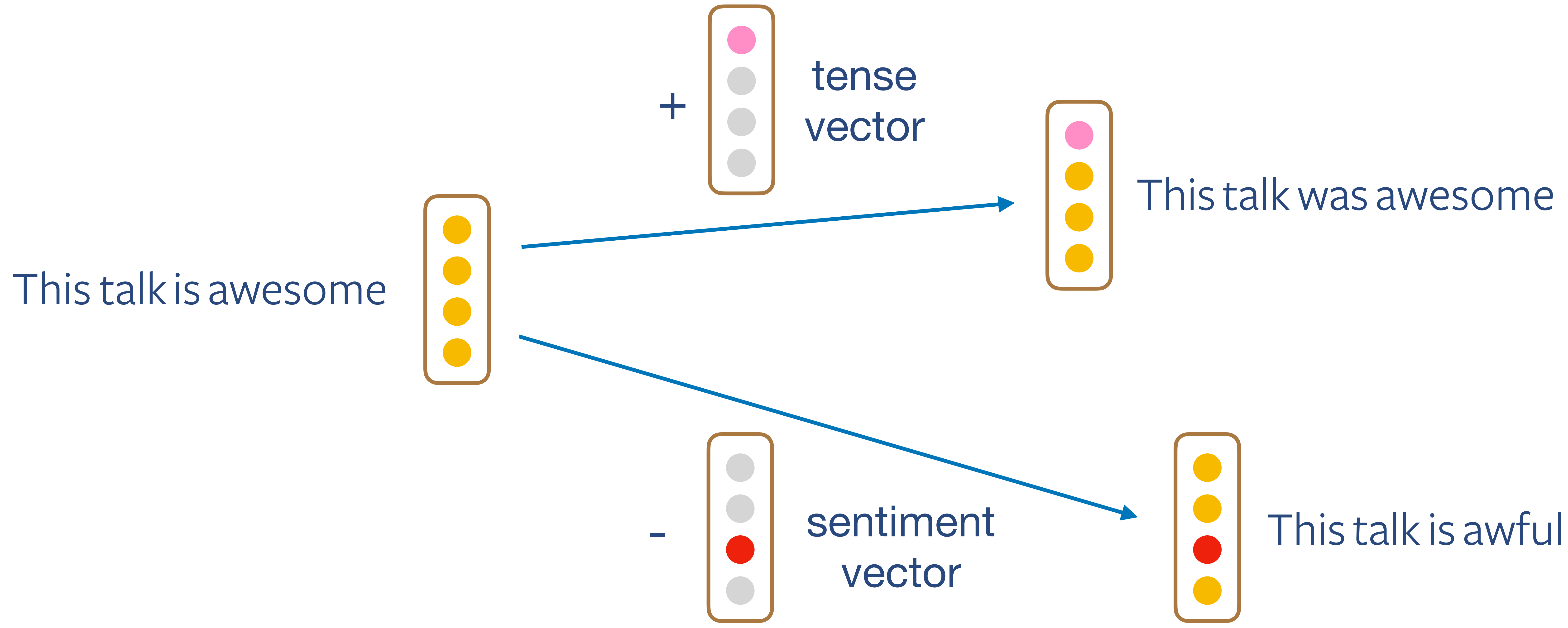


MIT CSAIL aws

Text Autoencoders

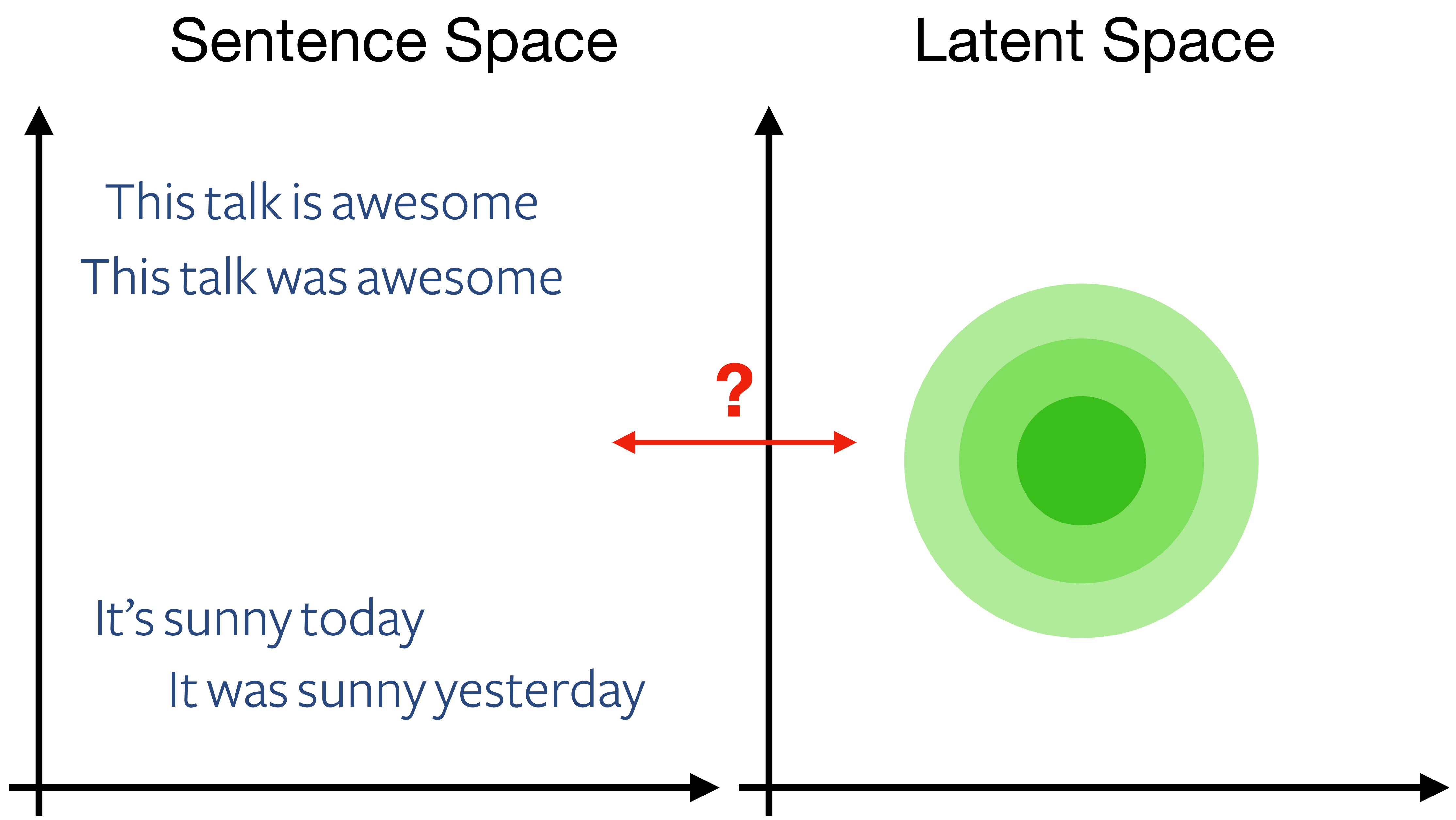
Represent sentences as vectors in a latent space

Manipulate sentences via modifying their latent representation



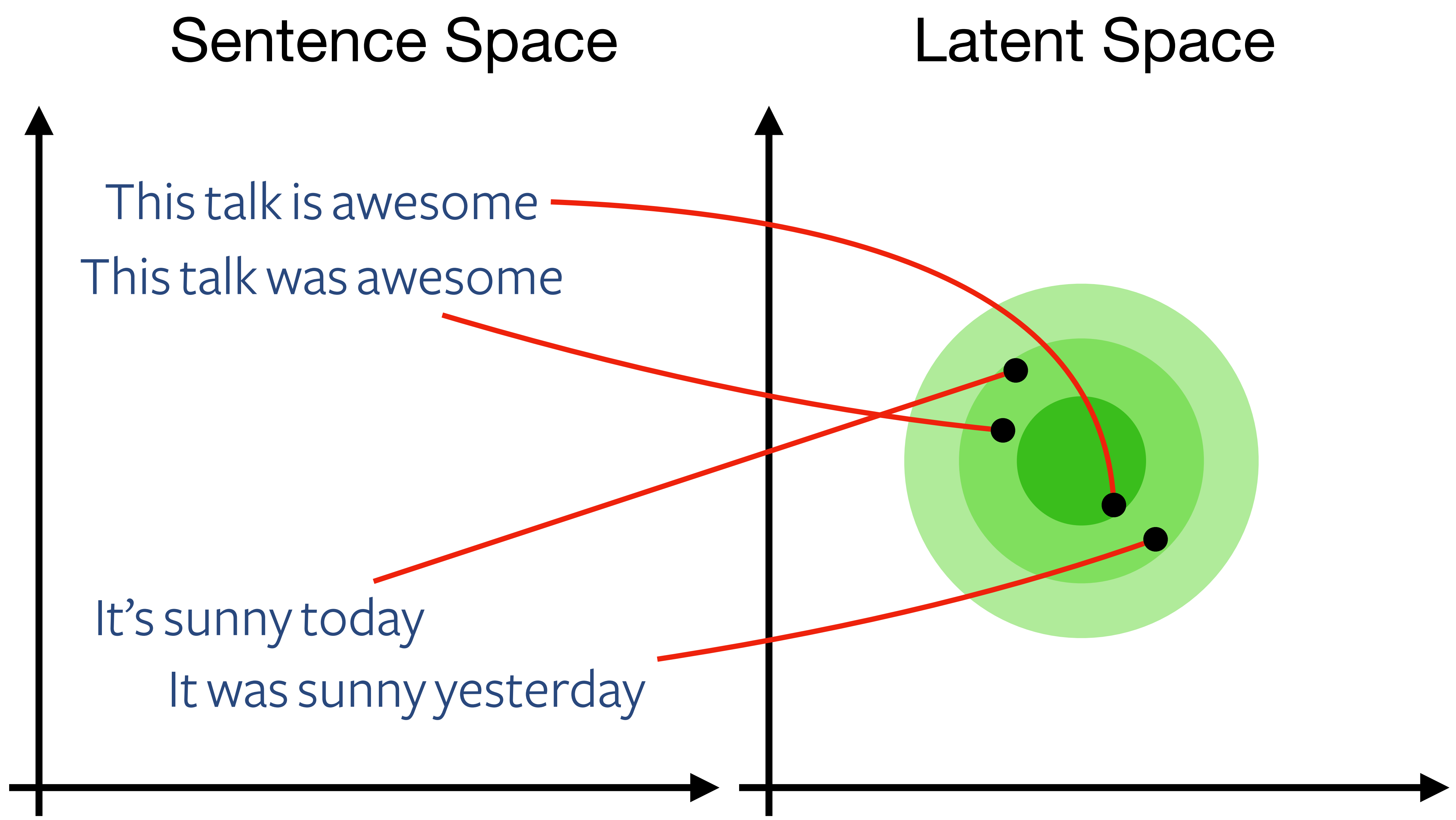
Latent Space Geometry

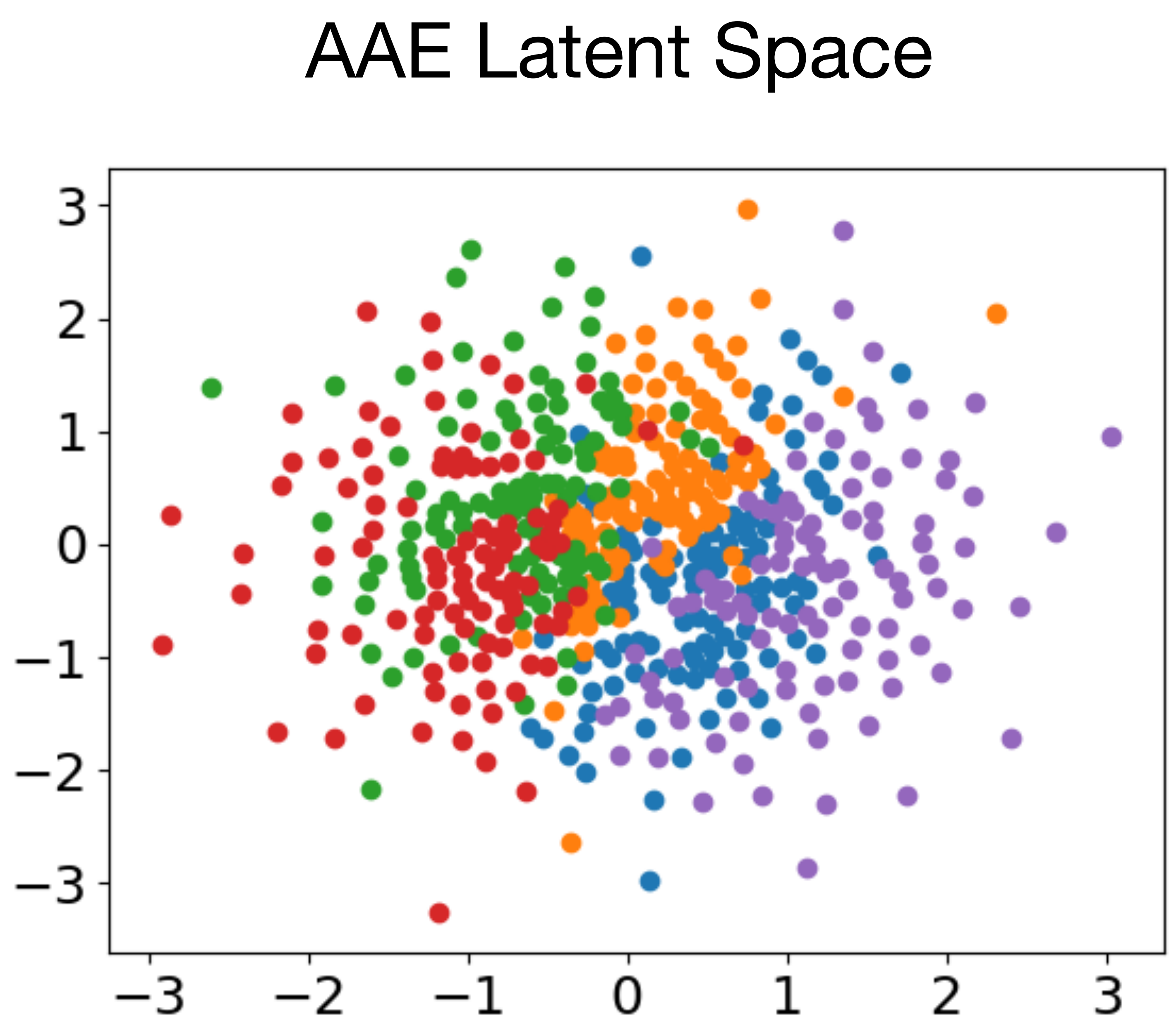
Which mapping between sentences and latent vectors will be learned?



Latent Space Geometry

Fortuitous geometry that captures sentence semantics is unlikely to arise





MIT CSAIL aws

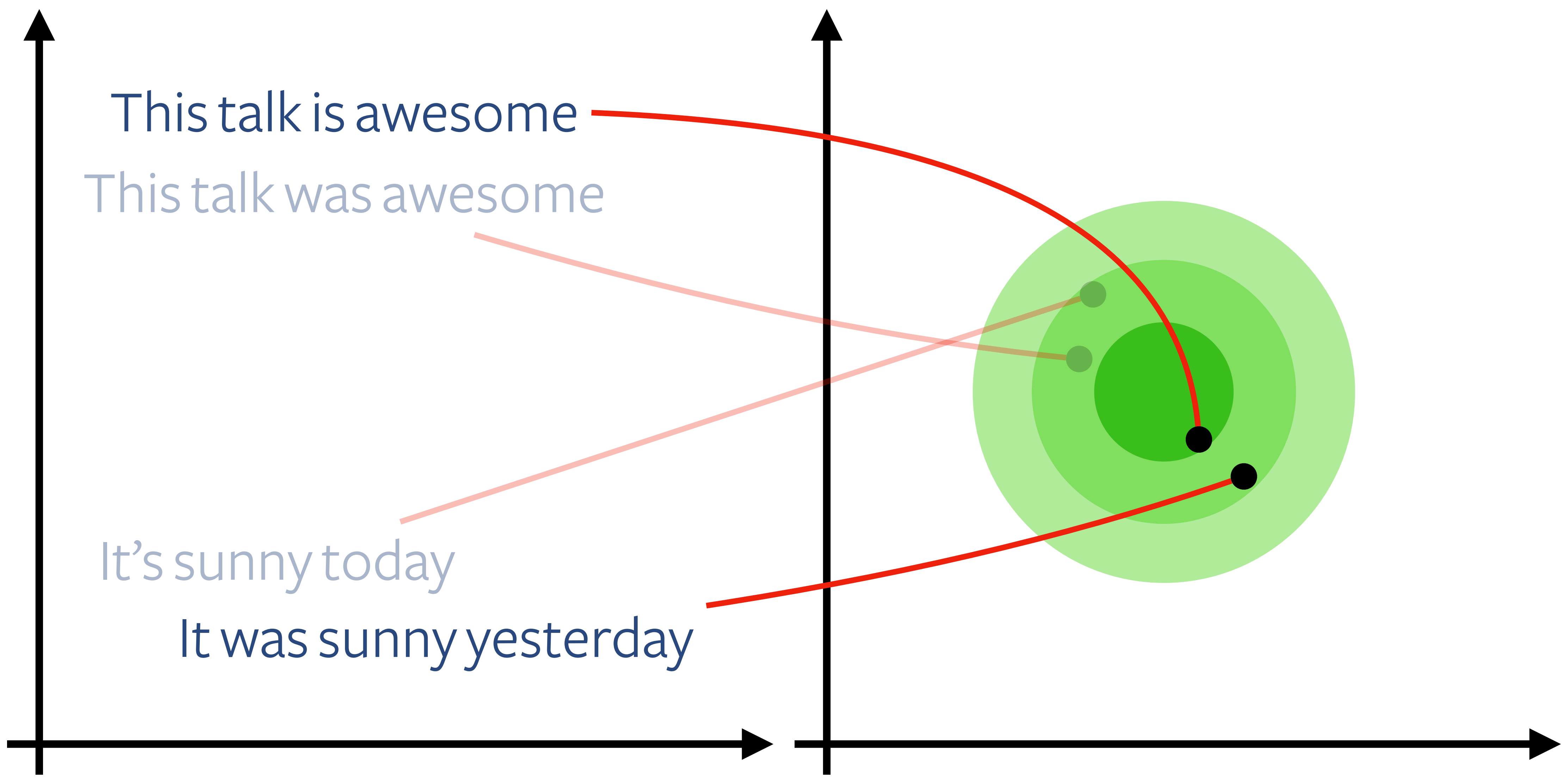
Latent Space Geometry

Fortuitous geometry that captures sentence semantics is unlikely to arise

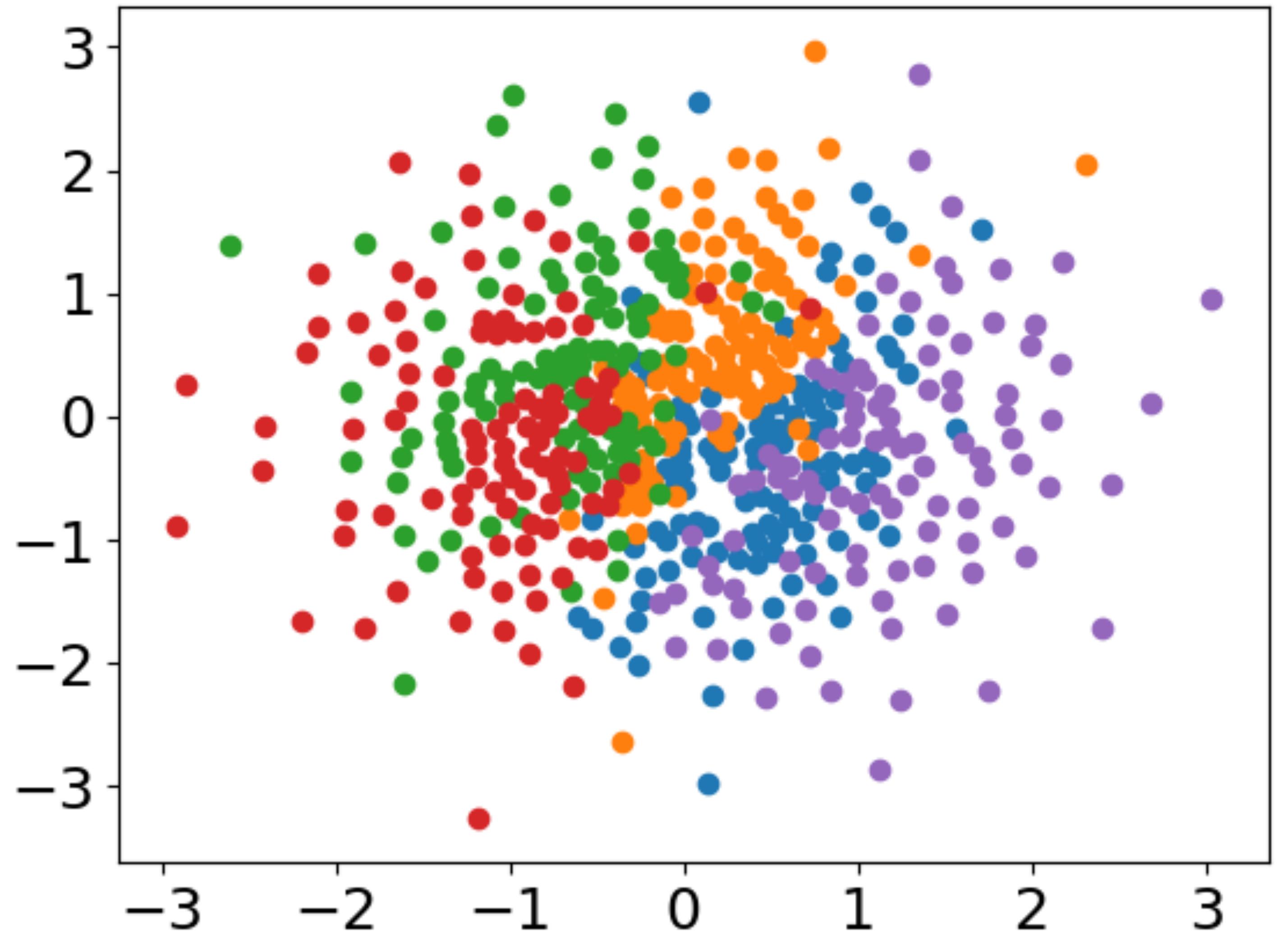
Minimal latent space manipulations can yield random, unpredictable changes in the resulting text

Sentence Space

Latent Space



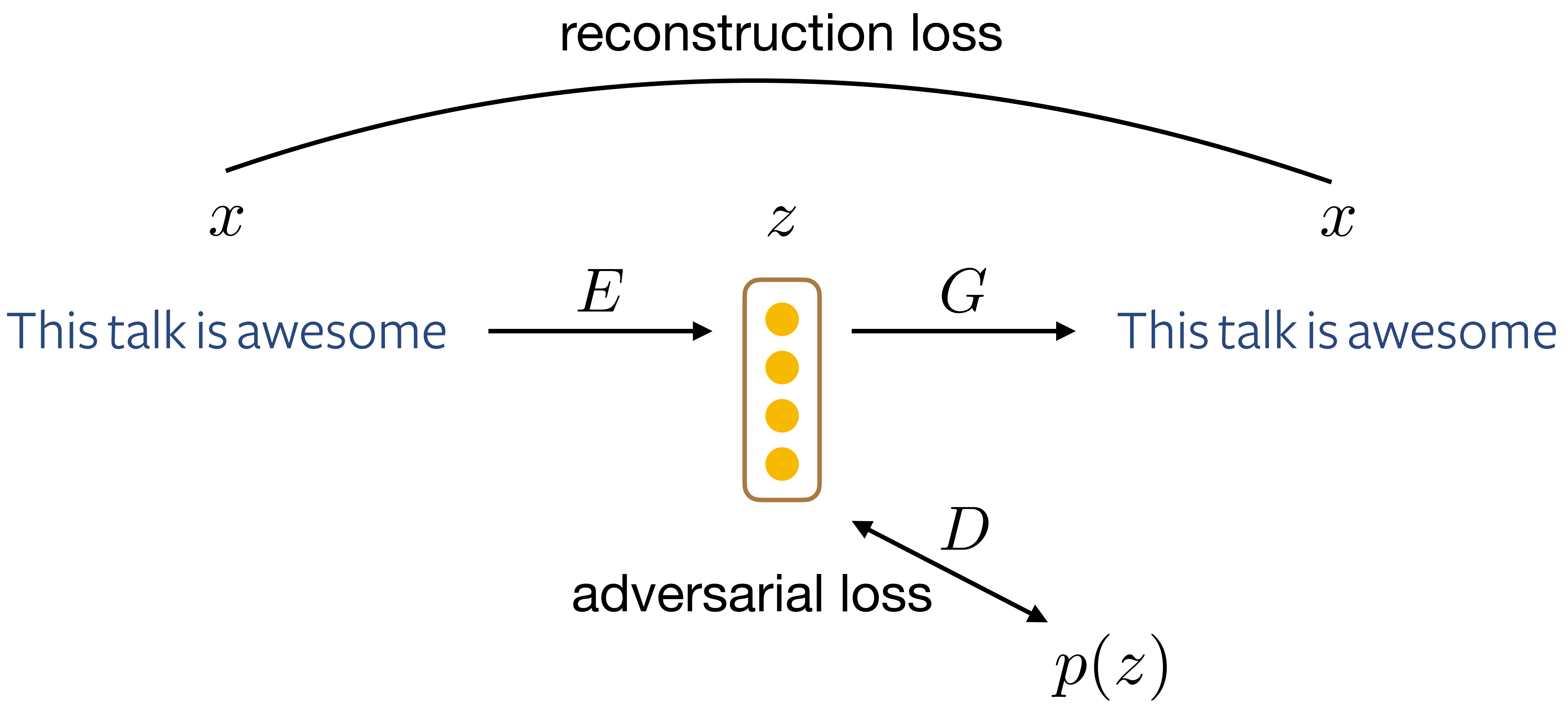
AAE Latent Space



Adversarial Autoencoder (AAE)

encoder E, decoder G, discriminator D

sample $z \sim p(z), x \sim p_G(x|z)$ to generate new data



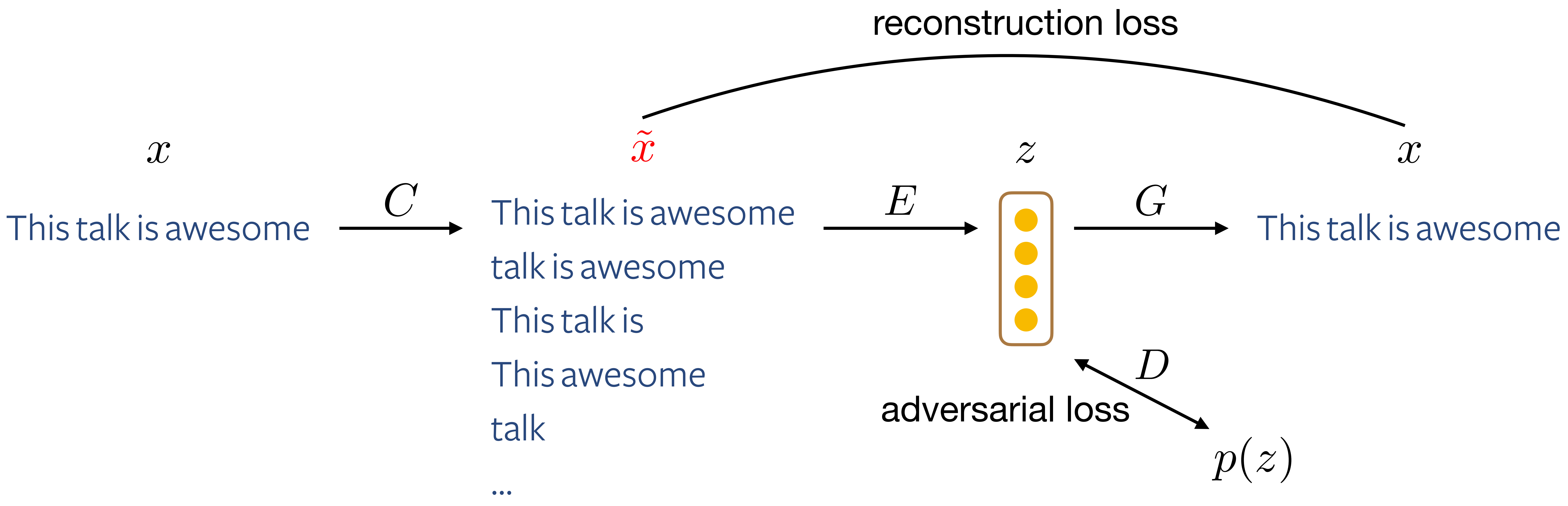
$$\min_{E,G} \max_{D} \; \mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{adv}}$$

MIT CSAIL aws

[Makhzani et al., 2015]

Our Model: Denoising AAE (DAAE)

Introduce a perturbation process C that maps x to nearby $\tilde{x}$ (e.g., randomly drop each word with probability p), and ask the model to reconstruct x from $\tilde{x}$ [Vincent et al., 2008]

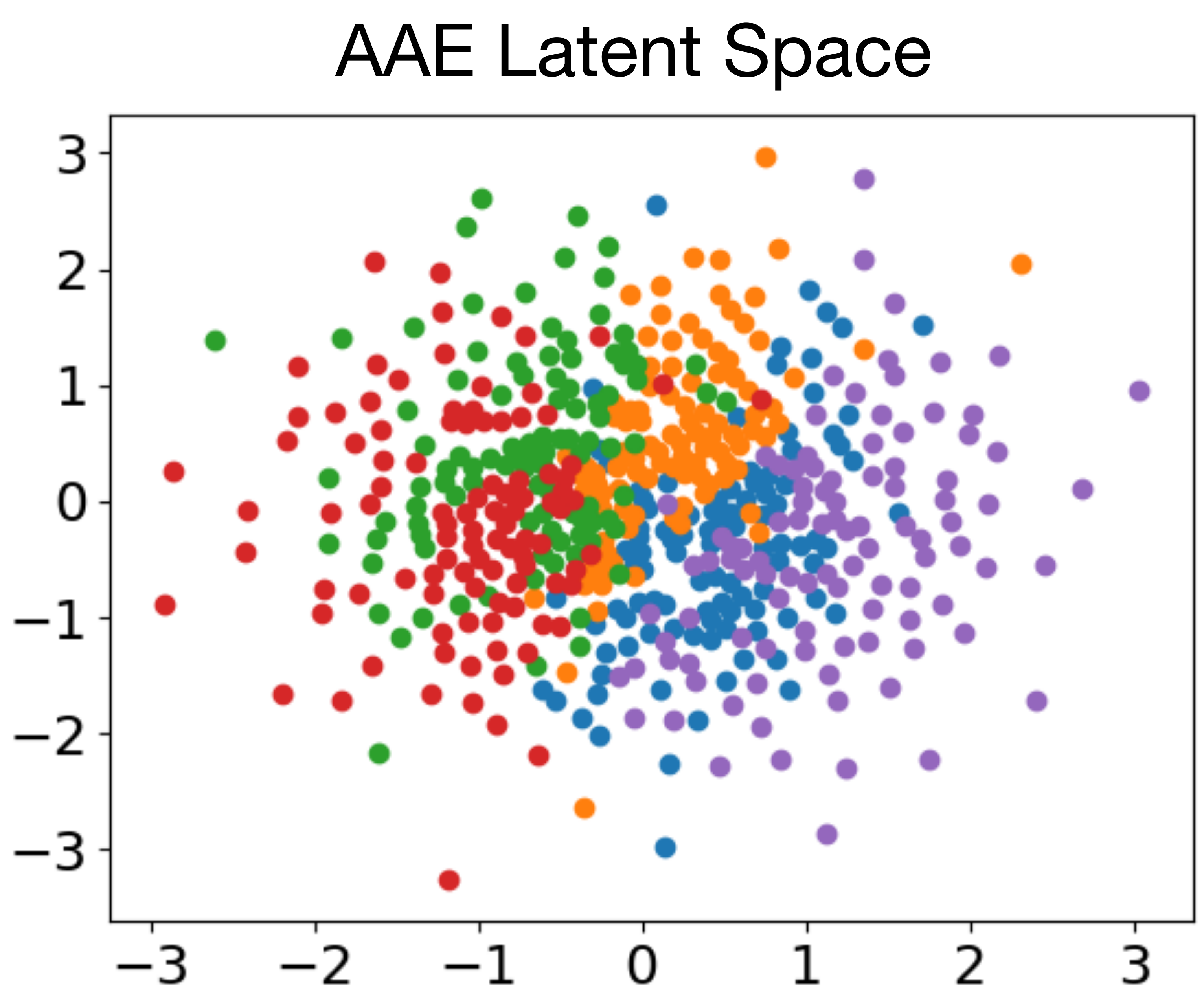


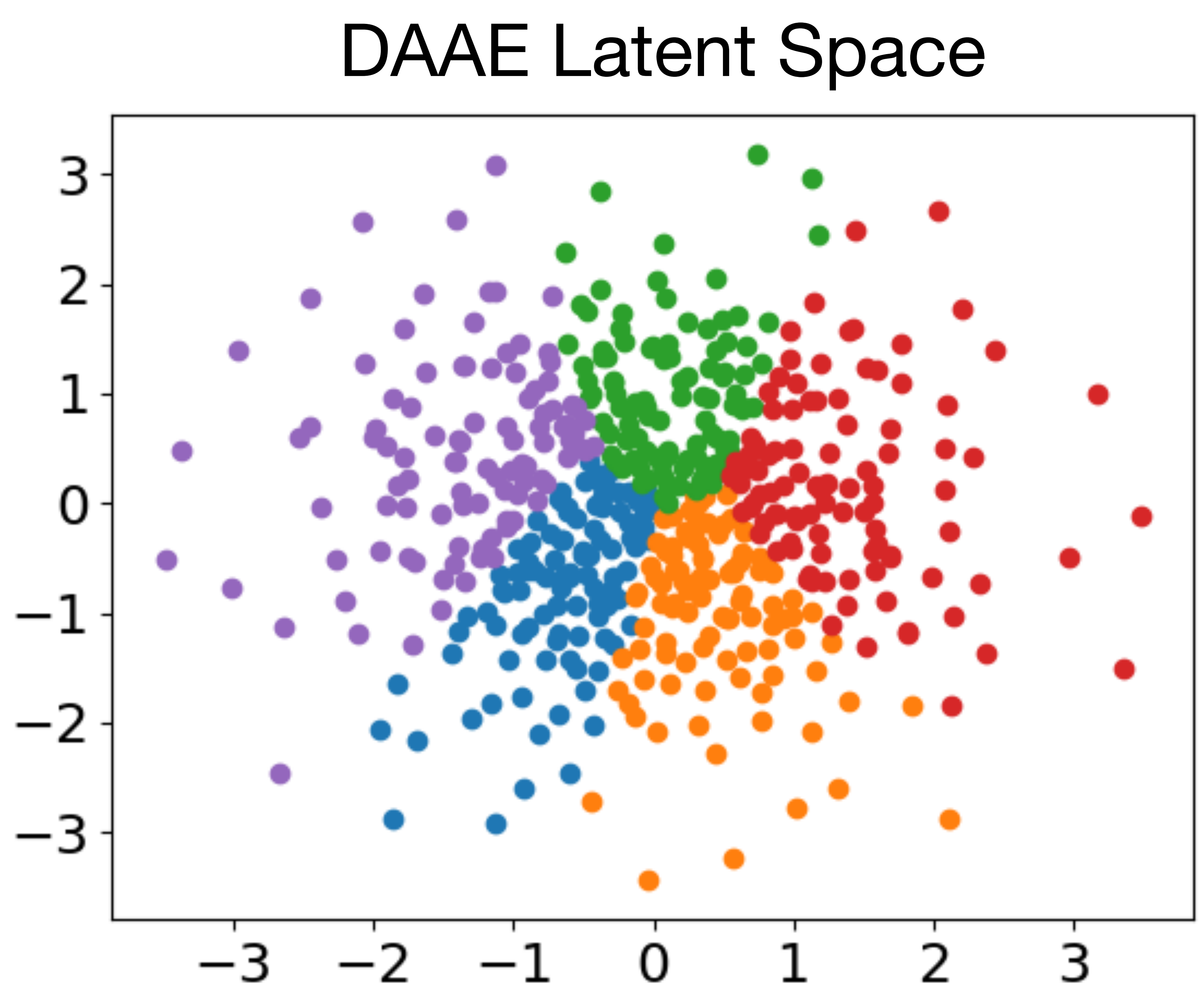
$$\min_{E,G} \max_{D} \; \mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{adv}}$$

MIT CSAIL aws

Toy Experiment

$\mathcal{X} = \{0,1\}^{50}, \ \mathcal{Z} = \mathbb{R}^2$ Data stem from 5 clusters, with 100 sequences sampled from each





similar sequences → distant representations

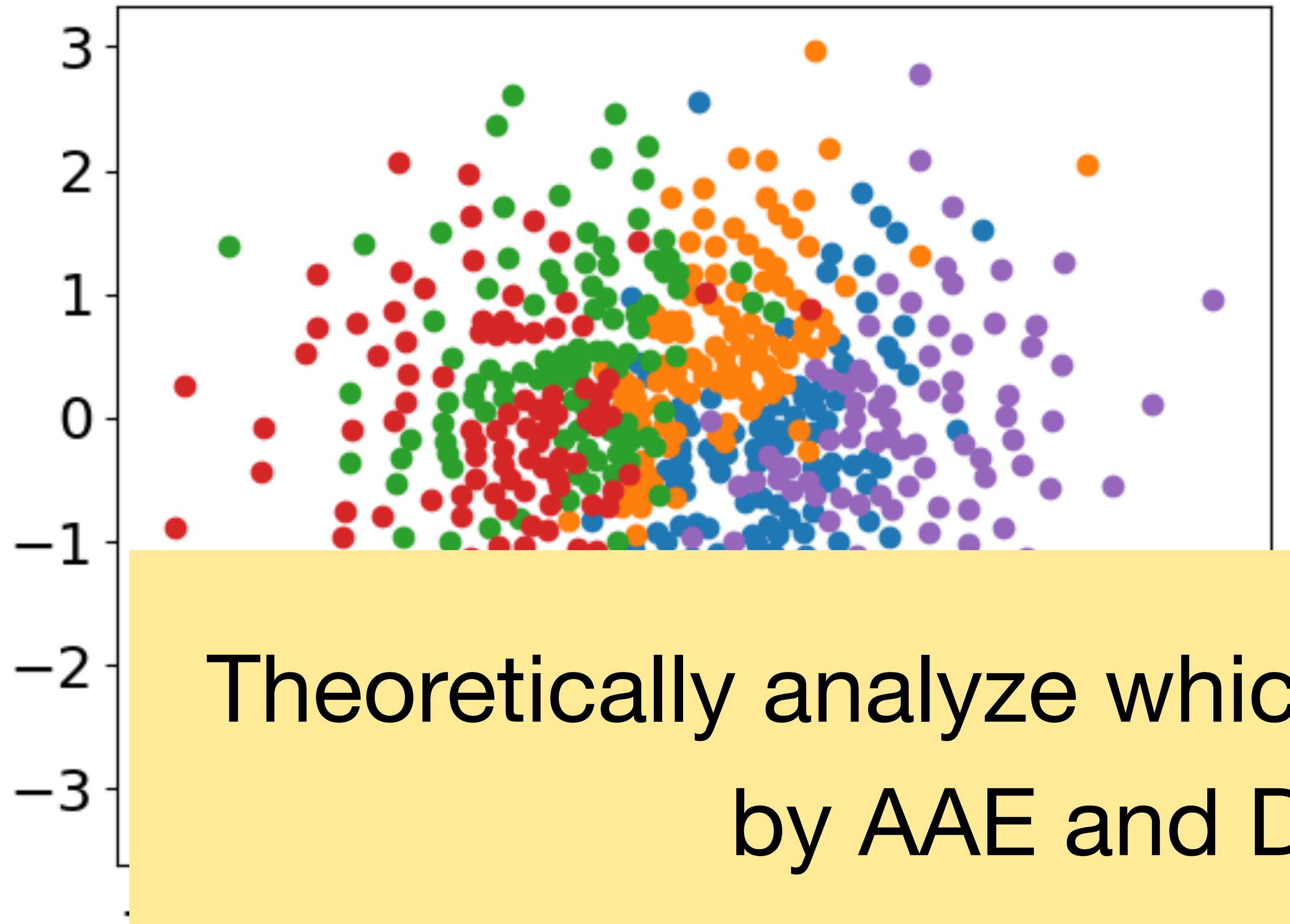
similar sequences → similar representations

MIT CSAIL aws

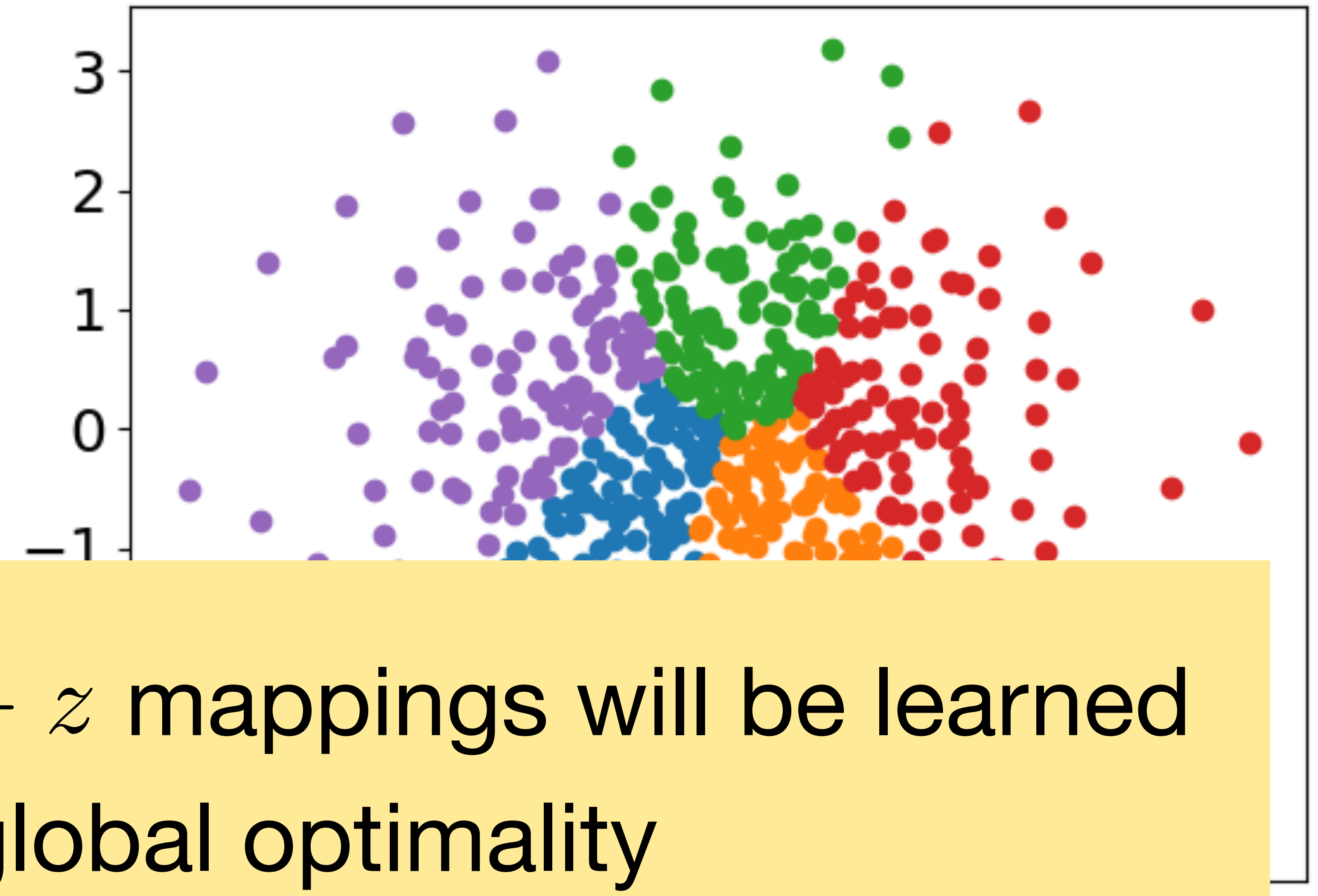
Toy Experiment

$\mathcal{X} = \{0,1\}^{50}, \ \mathcal{Z} = \mathbb{R}^2$ Data stem from 5 clusters, with 100 sequences sampled from each

AAE Latent Space



DAAE Latent Space



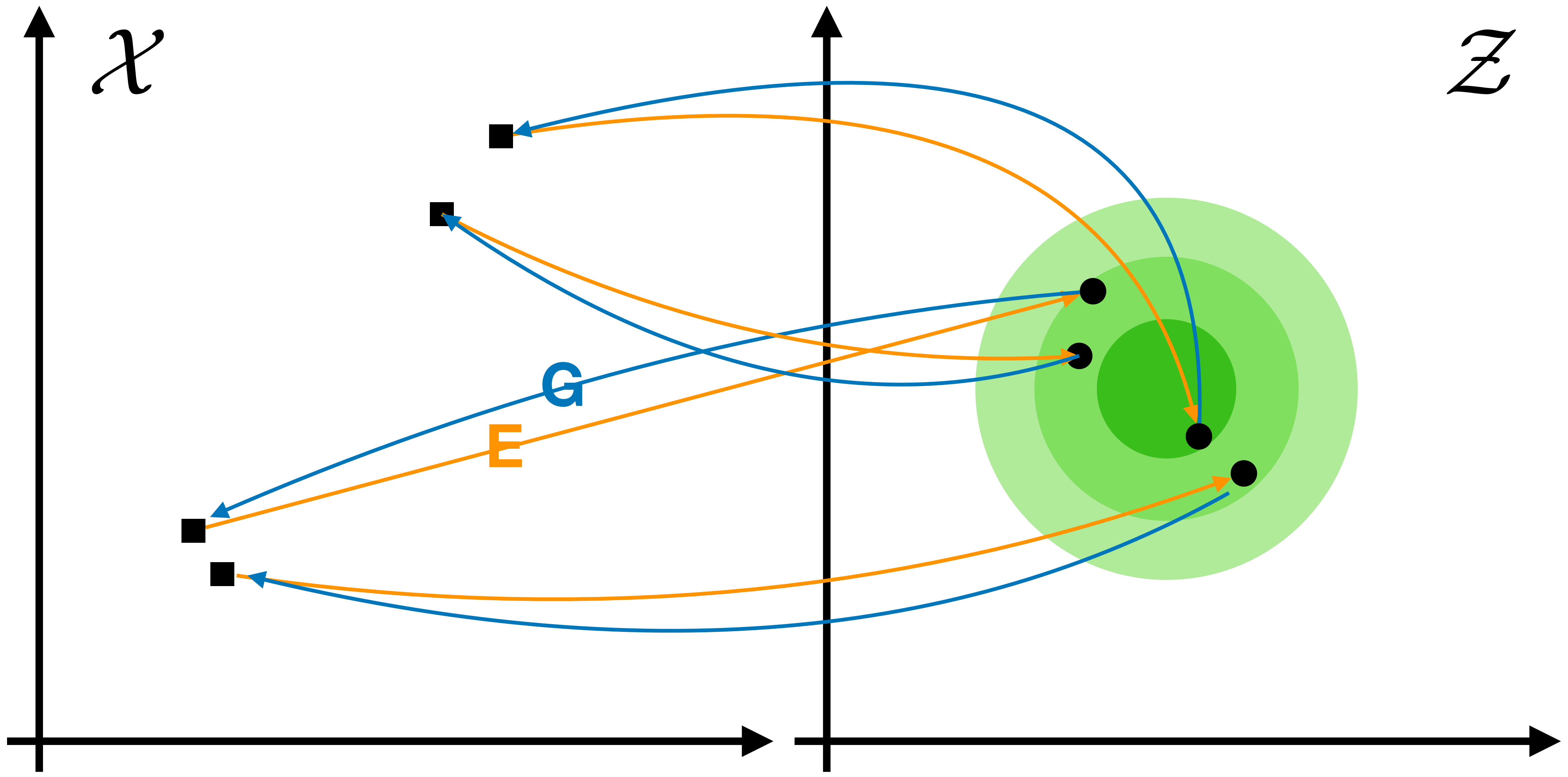
Theoretically analyze which type of $x - z$ mappings will be learned by AAE and DAAE under global optimality

similar sequences → distant representations

similar sequences → similar representations

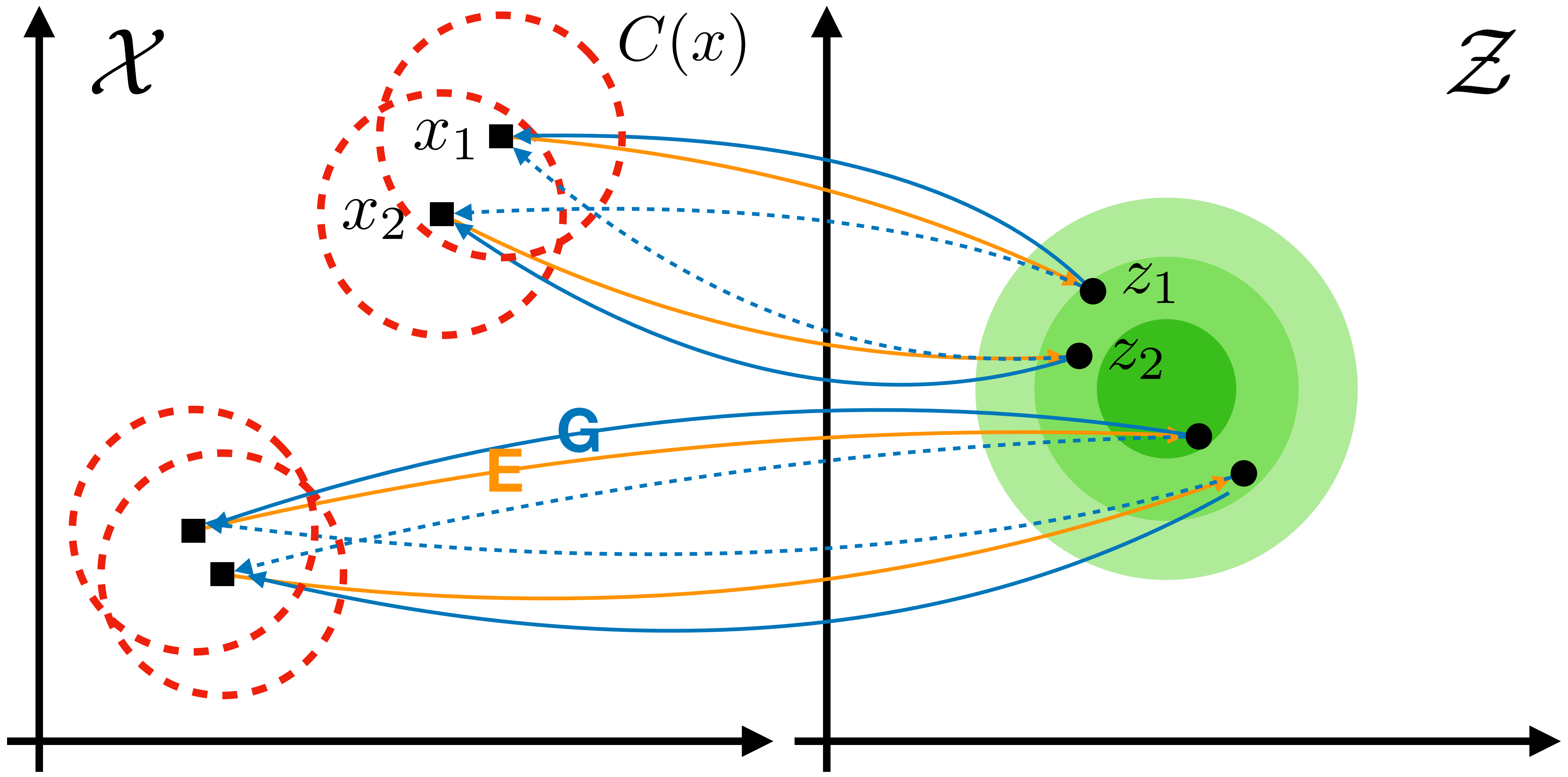
AAE Can Learn a Random Mapping Between X and Z

Theorem 1. *With high-capacity encoder/decoder networks, any assignment between* $\{x_1, \cdots, x_n\}$ *and* $\{z_1, \cdots, z_n\}$ *can achieve the same optimal value under the AAE objective*



DAAE Learns to Map Similar X to Close Z

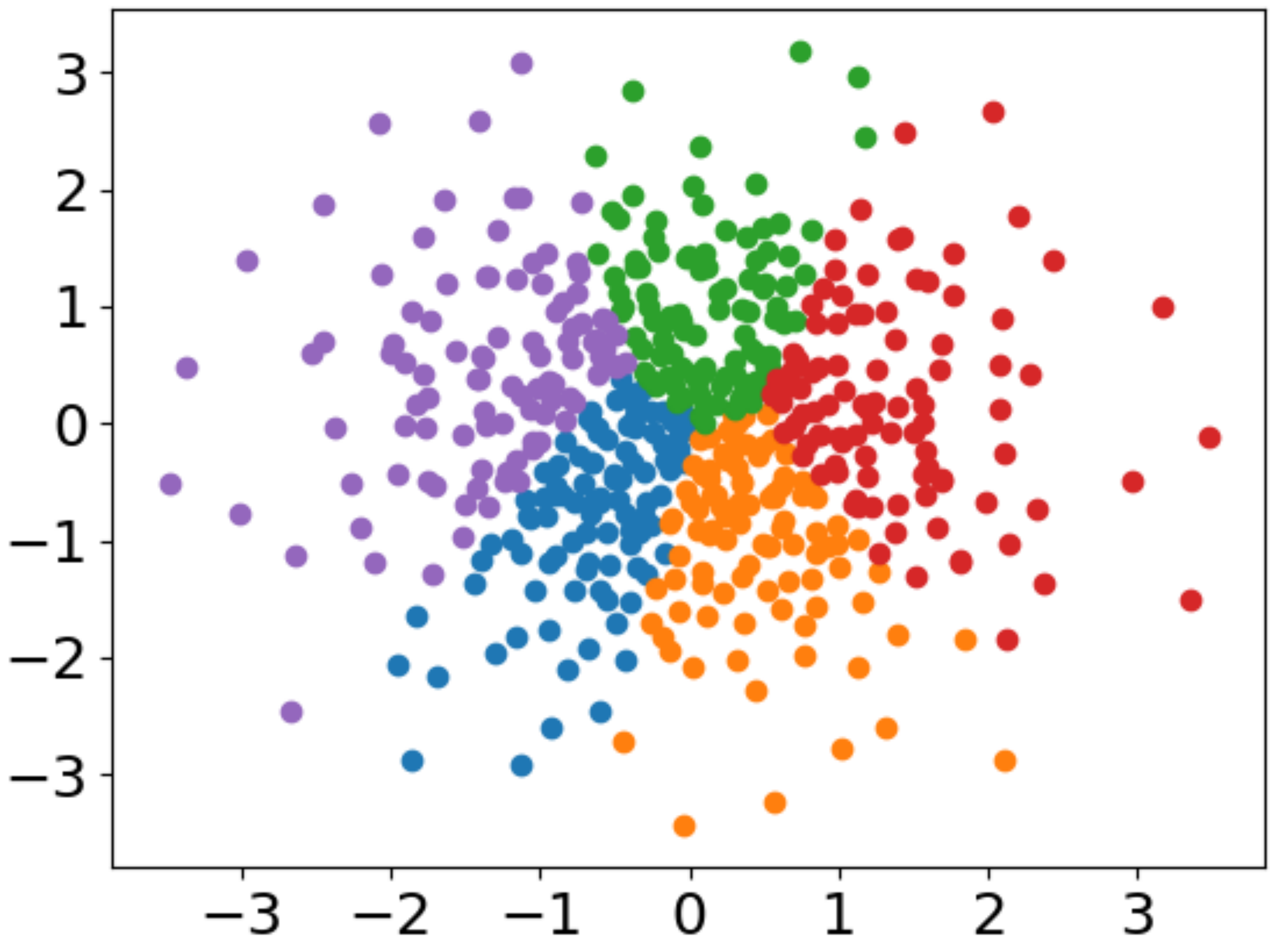
Theorem 2. *In a simple scenario with only four examples, the optimal value under the DAAE objective is achieved when close pairs of* x *are mapped to close pairs of* z



MIT CSAIL
aws

DAAE Learns to Map Similar X to Close Z

Theorem 3 (sketch). *Suppose $x_1, \cdots, x_n$ are divided into n/K clusters of equal size K. Let the perturbation process C be uniform within clusters. The DAAE objective is "best achieved"* when examples in the same cluster are mapped to the latent space in a manner that is well-separated from encodings of other clusters



Experiments

Compare DAAE with:

- AAE [Makhzani et al., 2015]
- Latent-noising AAE (LAAE) [Rubenstein et al., 2018]
- β-VAE [Higgins et al., 2017]
- ARAE [Zhao et al., 2018]

Evaluate:

- Neighborhood Preservation
- Generation-Reconstruction Trade-Off
- Style Transfer
- Sentence Interpolation

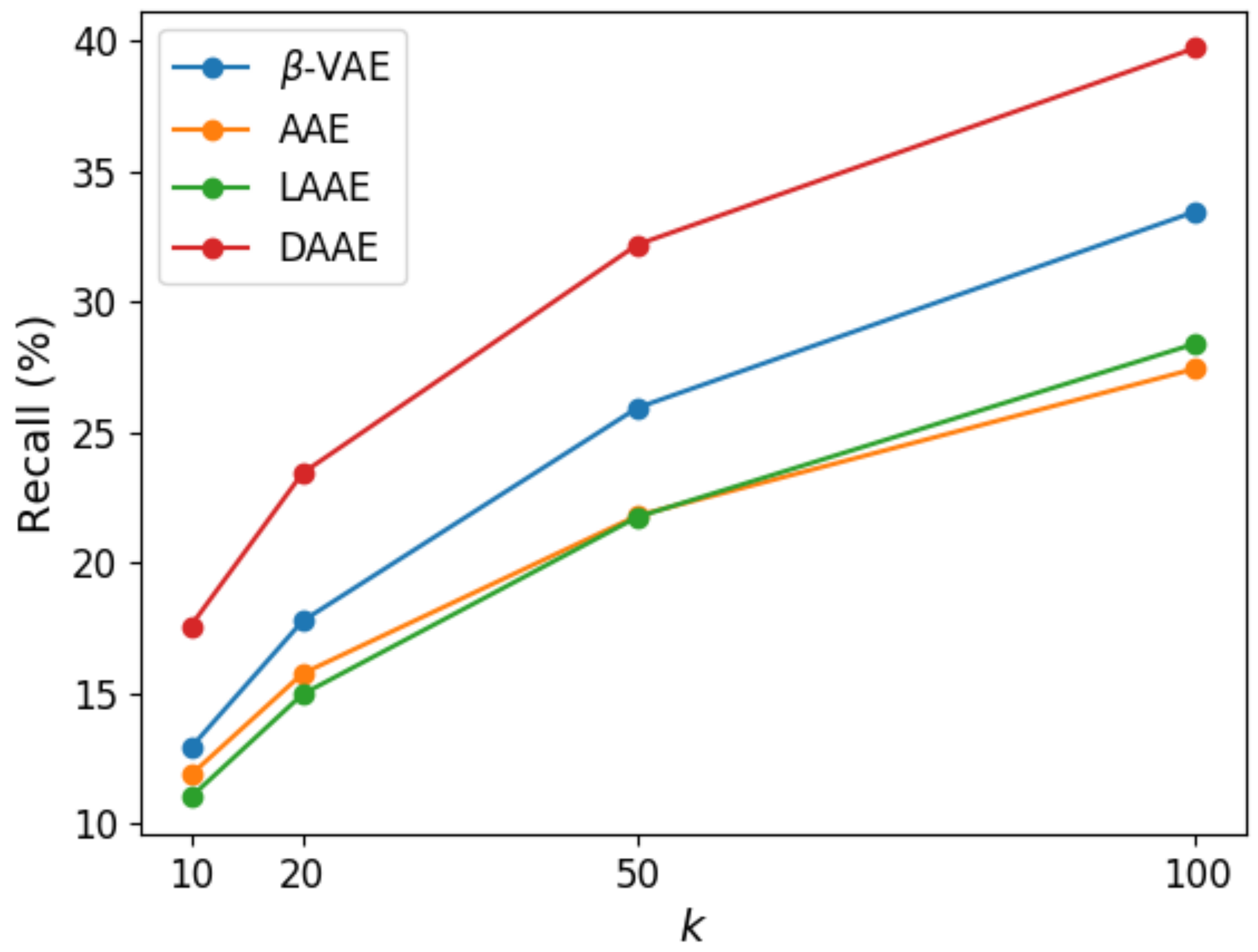
Datasets:

- Yelp reviews
- Yahoo answers

Neighborhood Preservation

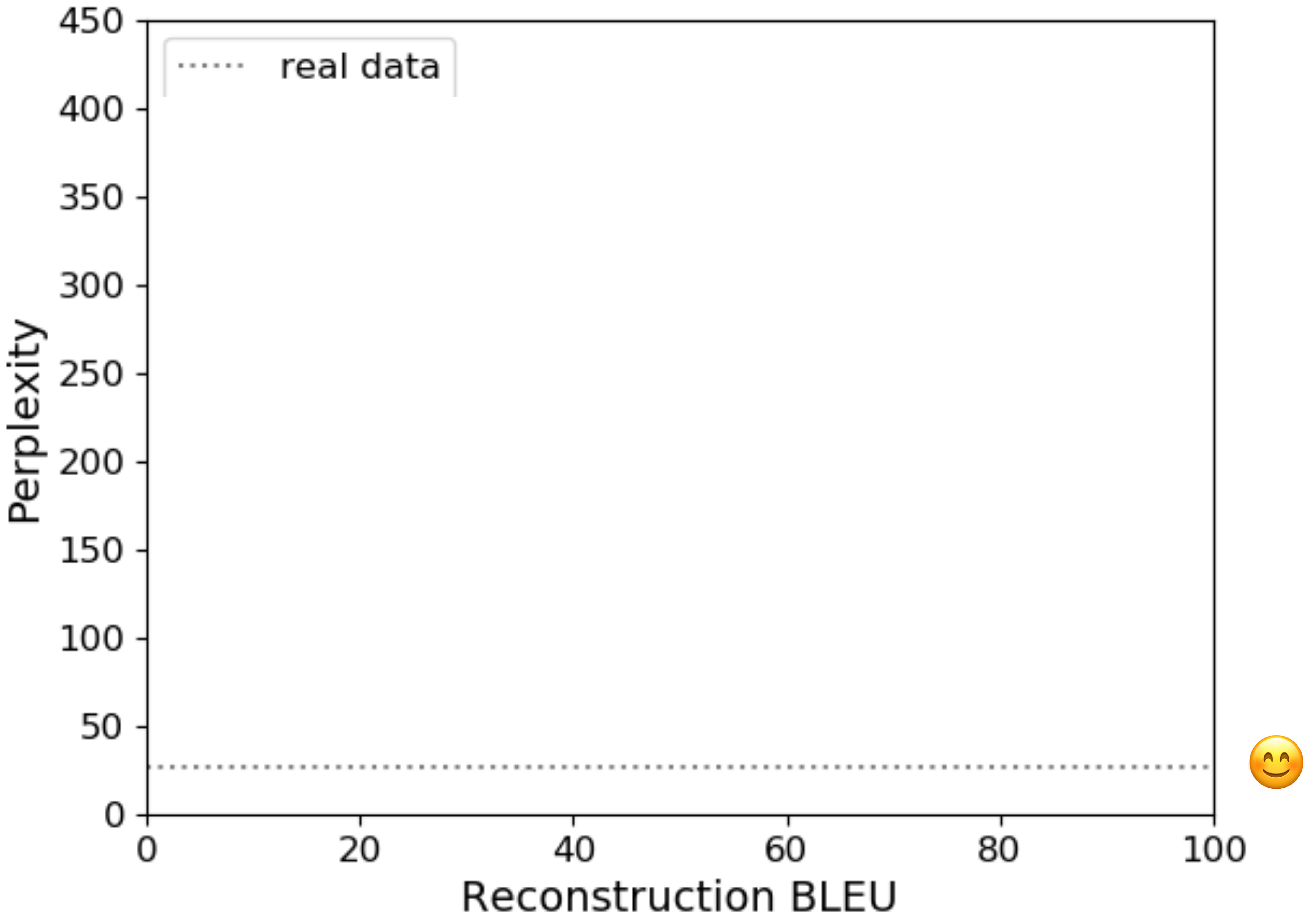
Source	my waitress katie was fantastic , attentive and personable .
AAE	my cashier did not smile , barely said hello . the service is fantastic , the food is great . the employees are extremely nice and helpful .
DAAE	the manager , linda , was very very attentive and personable . stylist brenda was very friendly , attentive and professional . the manager was also super nice and personable .

Nearest neighbors (NN) in the latent Euclidean space

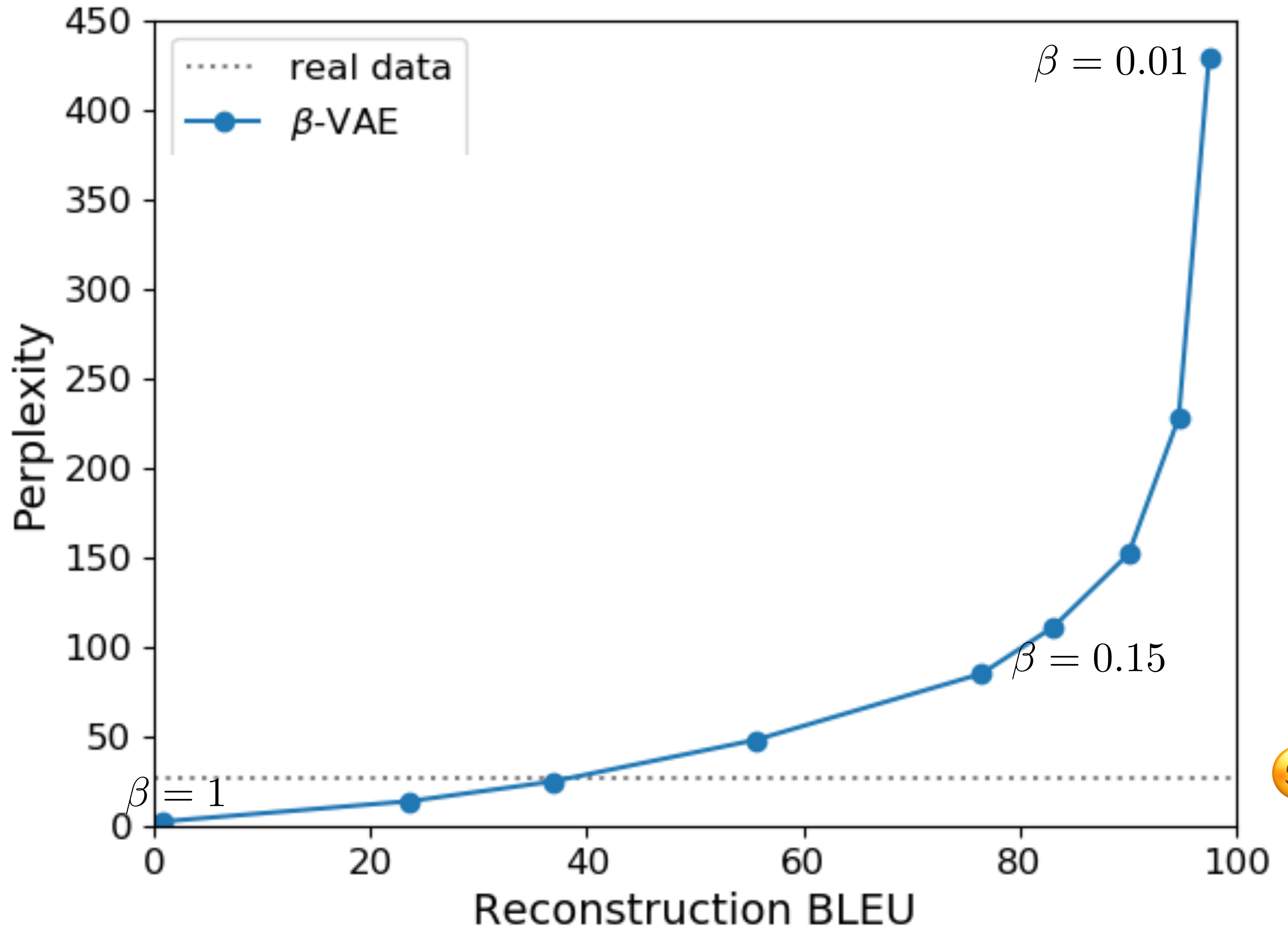


For each sentence's 10-NN in terms of normalized edit distance, count how many of them lie among the k-NN in the latent space

Generation-Reconstruction Trade-Off

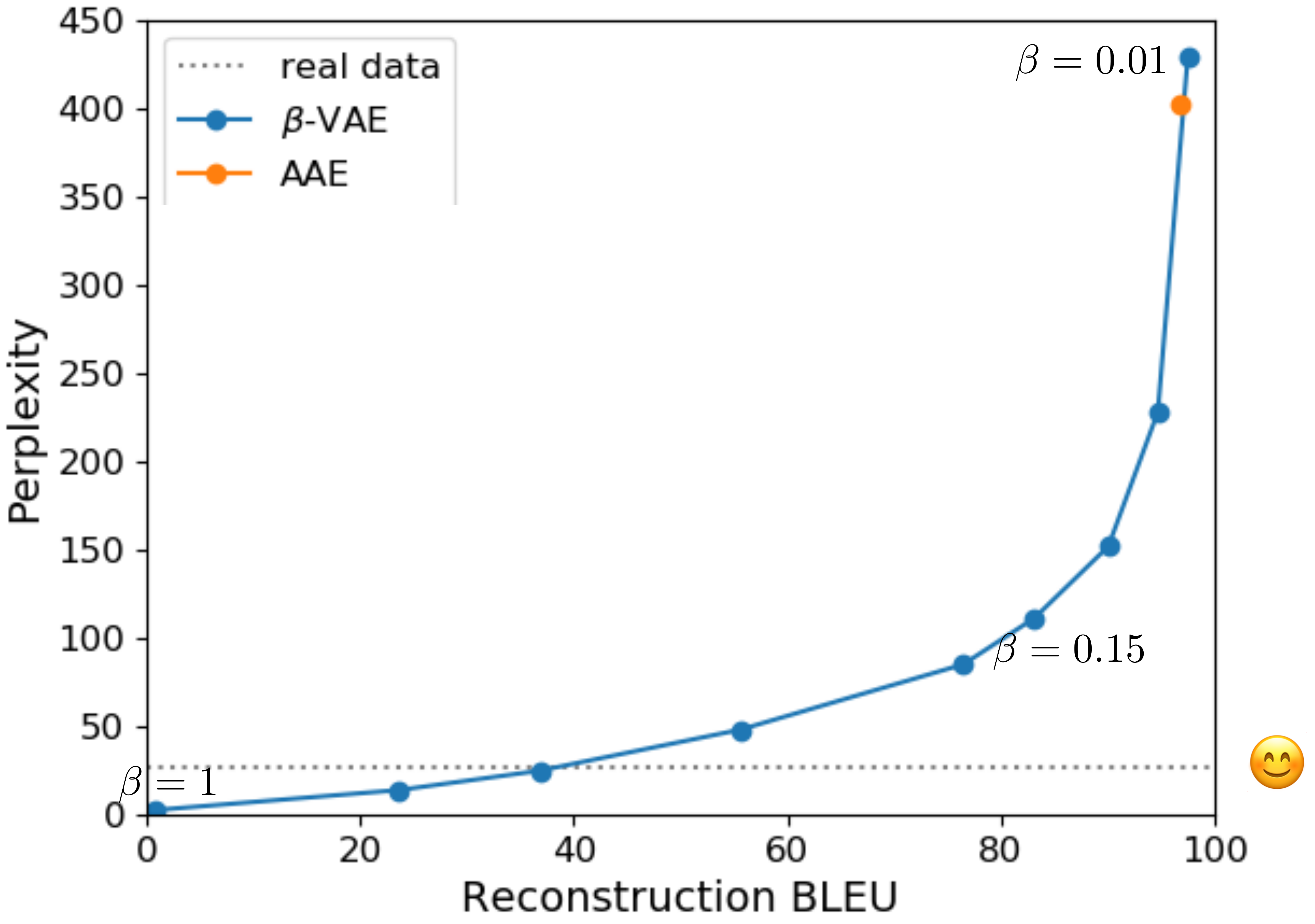


Generation-Reconstruction Trade-Off



😊

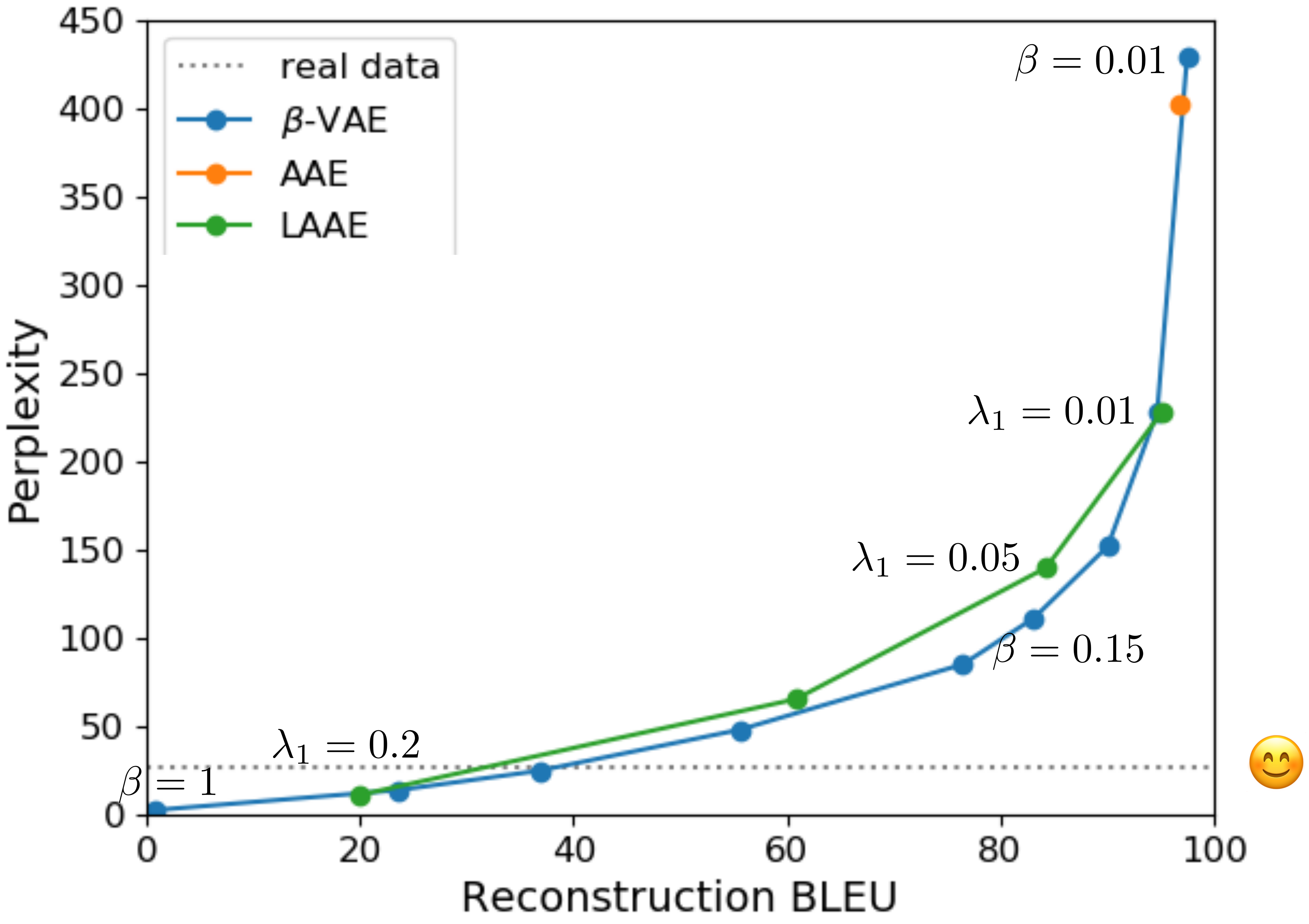
Generation-Reconstruction Trade-Off



MIT CSAIL

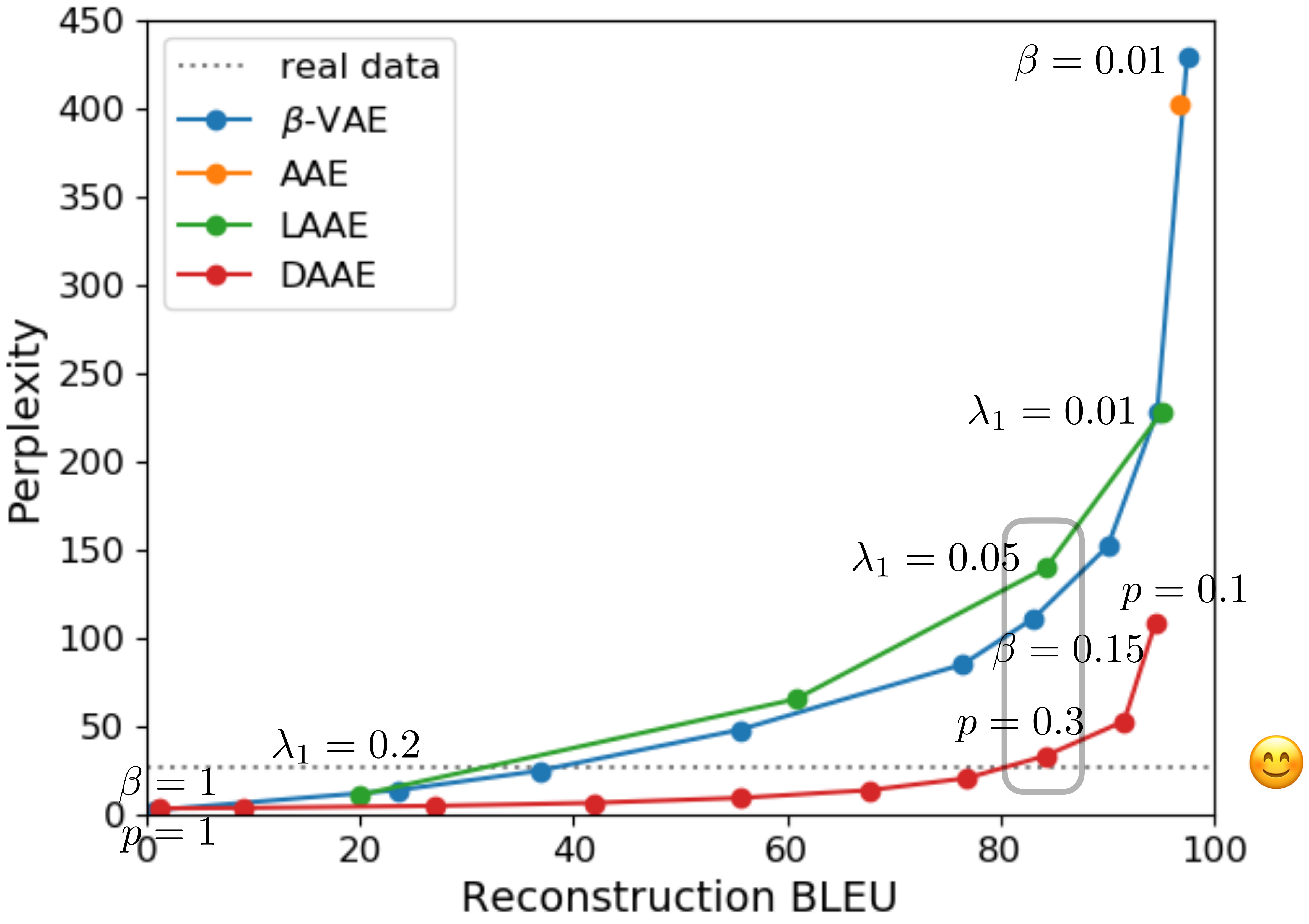
aws

Generation-Reconstruction Trade-Off

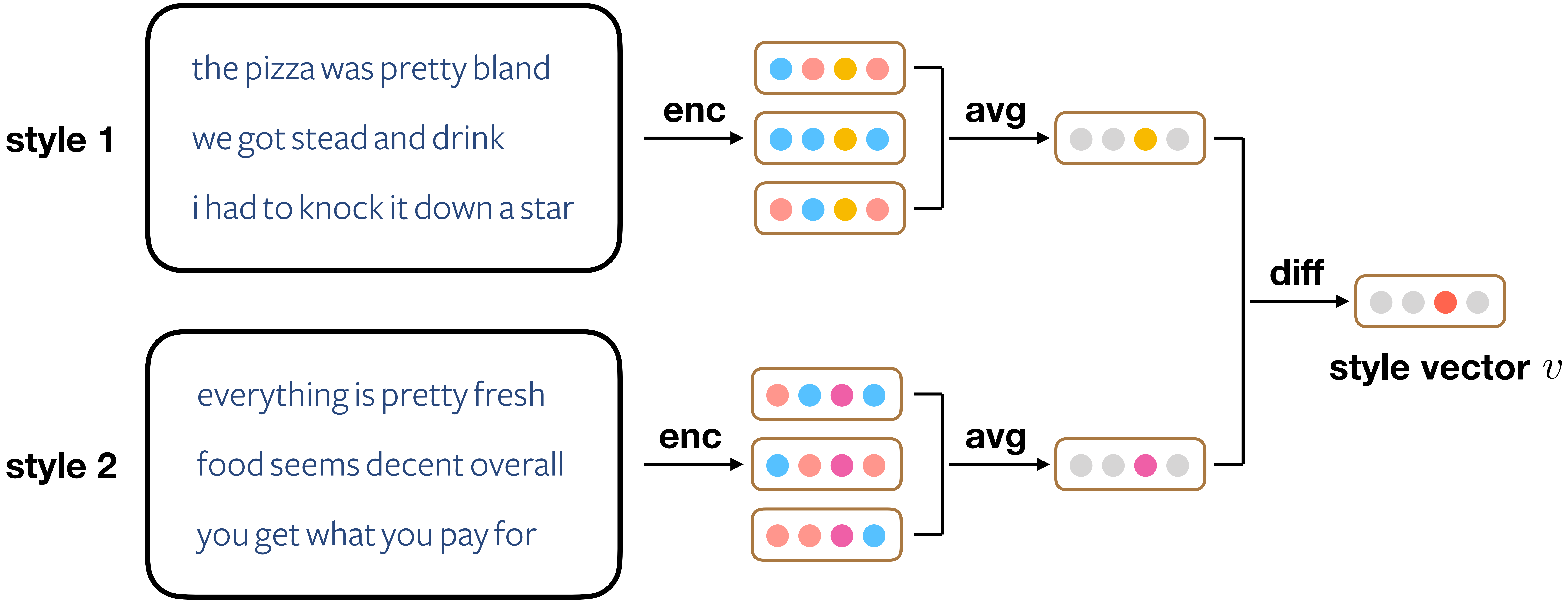


MIT CSAIL aws

Generation-Reconstruction Trade-Off



Unsupervised Text Style Transfer

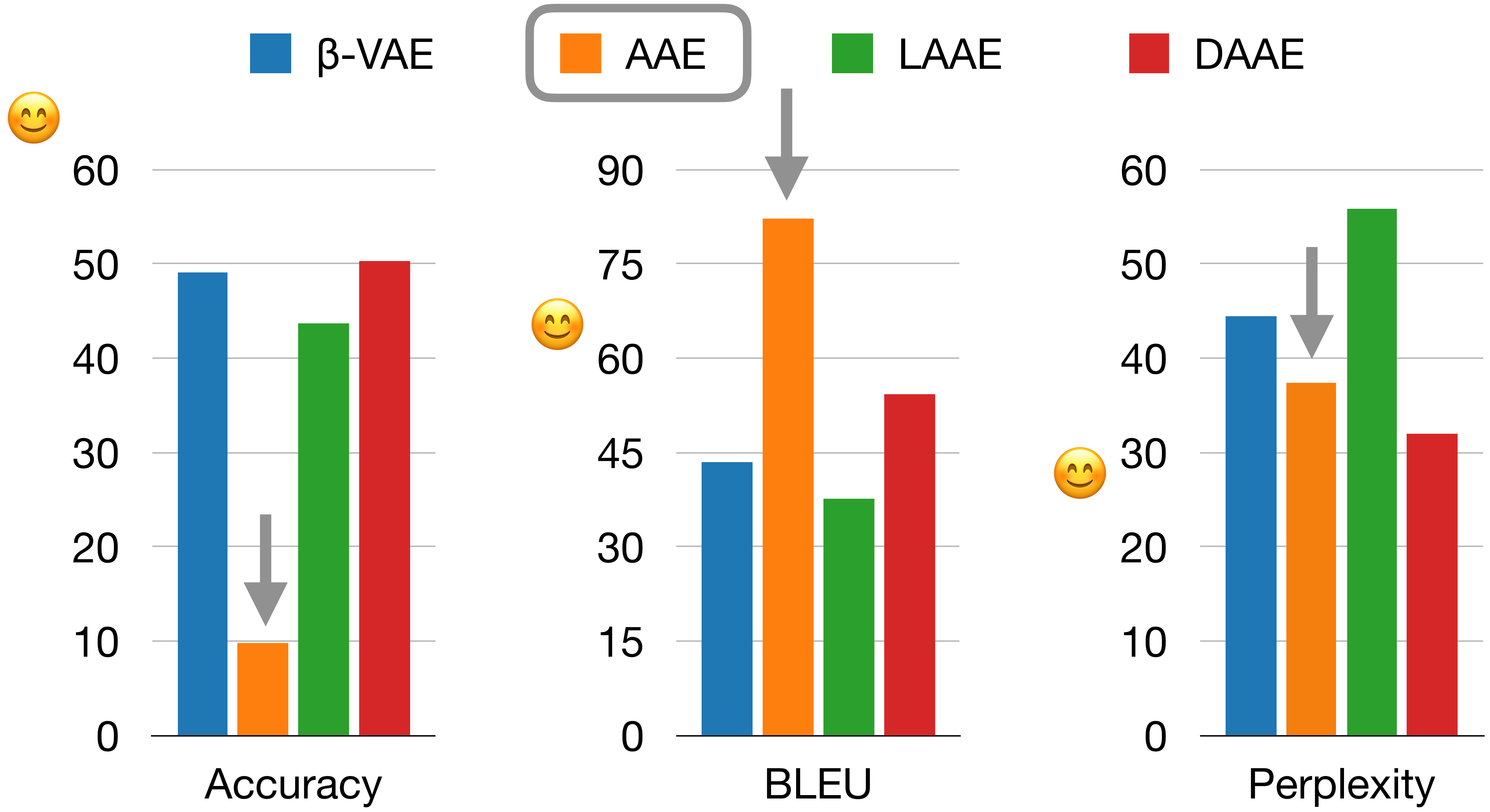


dec(enc(input **) ± v) = ?**

No style labels required during training!

Tense Transfer

- AAE has the highest BLEU but the lowest ACC ⟶ not change the source sentence

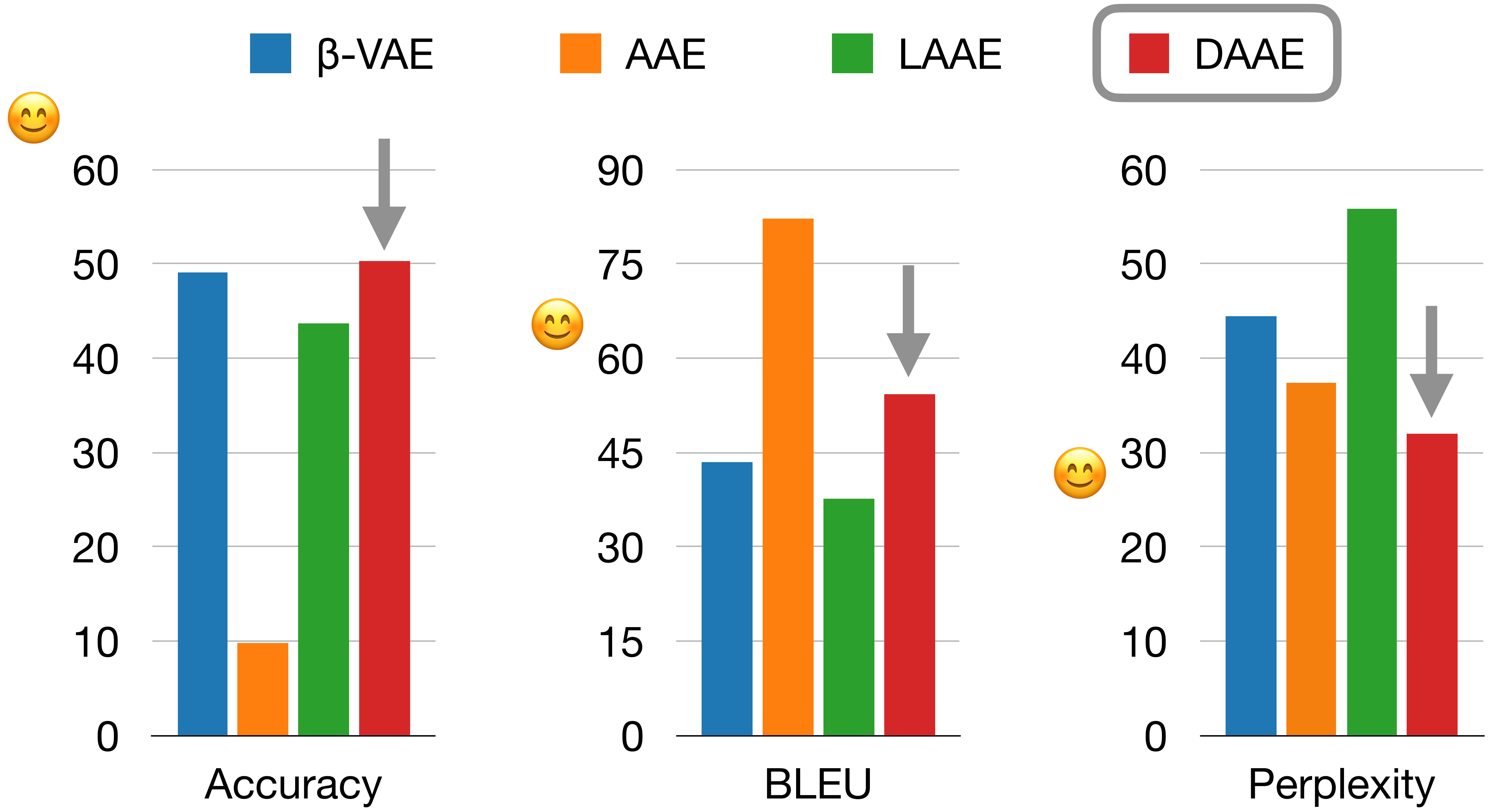


MIT CSAIL aws

Tense Transfer

- DAAE achieves the highest ACC, the lowest PPL, relatively high BLEU

✓ *proper tense* ✓ *high quality* ✓ *faithful to source*



Tense Transfer

- DAAE achieves the highest ACC, the lowest PPL, relatively high BLEU

✓ *proper tense* ✓ *high quality* ✓ *faithful to source*

Input	the staff is rude and the dr. does not spend time with you .
β-VAE	the staff was rude and the dr. did not spend time with your attitude .
AAE	the staff was rude and the dr. does not spend time with you .
LAAE	the staff was rude and the dr. is even for another of her entertained .
DAAE	the staff was rude and the dr. did not make time with you .

MIT CSAIL aws

Sentiment Transfer

- As the scaling factor increases, the resulting sentences generated by DAAE get more and more positive/negative

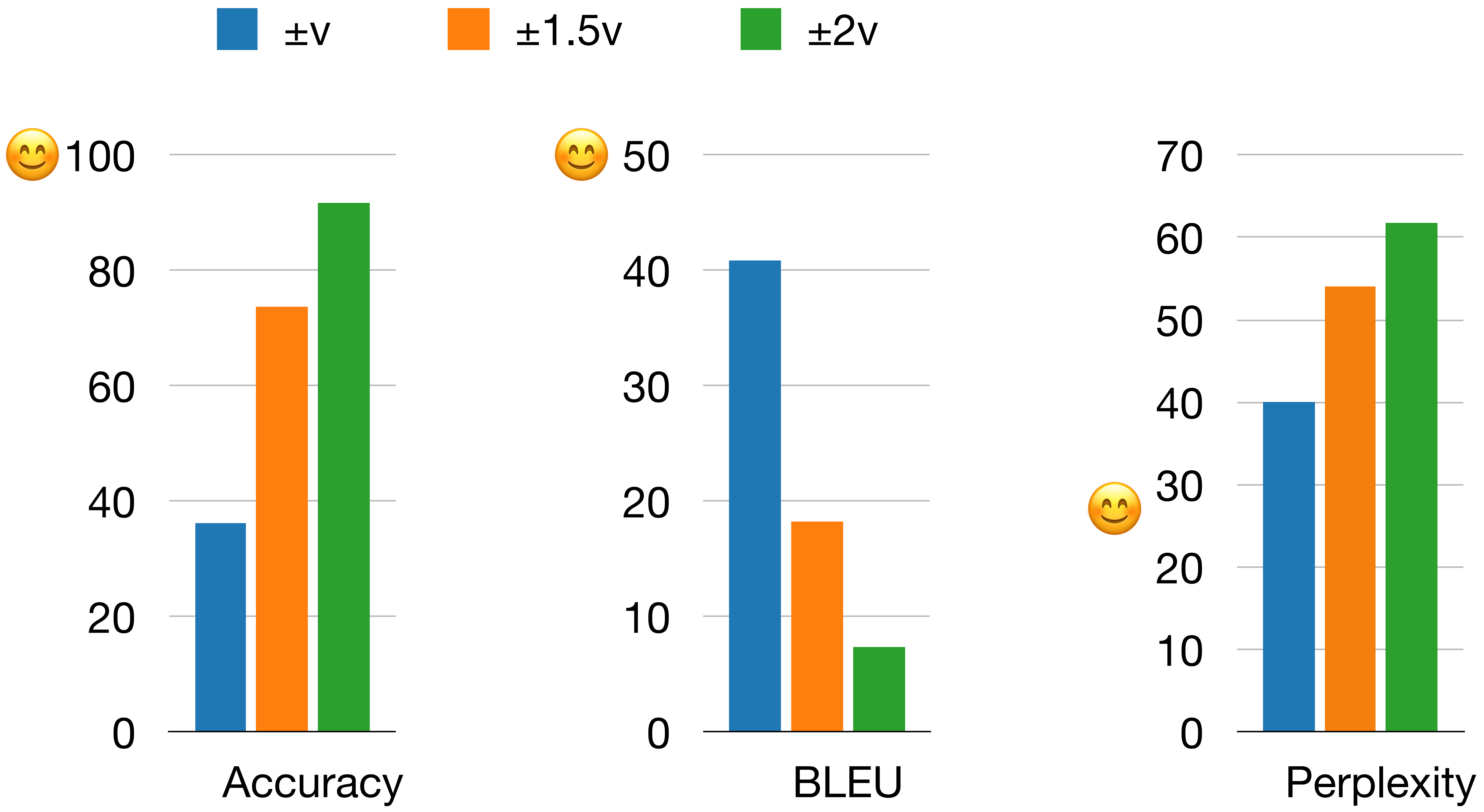
Input	the food is entirely tasteless and slimy .
+v	the food is tremendous and fresh .
+1.5v	the food is sensational and fresh .
+2v	the food is gigantic .

Input	the patrons all looked happy and relaxed .
-v	the patrons all helped us were happy and relaxed .
-1.5v	the patrons that all seemed around and left very stressed .
-2v	the patrons actually kept us all looked long and was annoyed .

MIT CSAIL aws

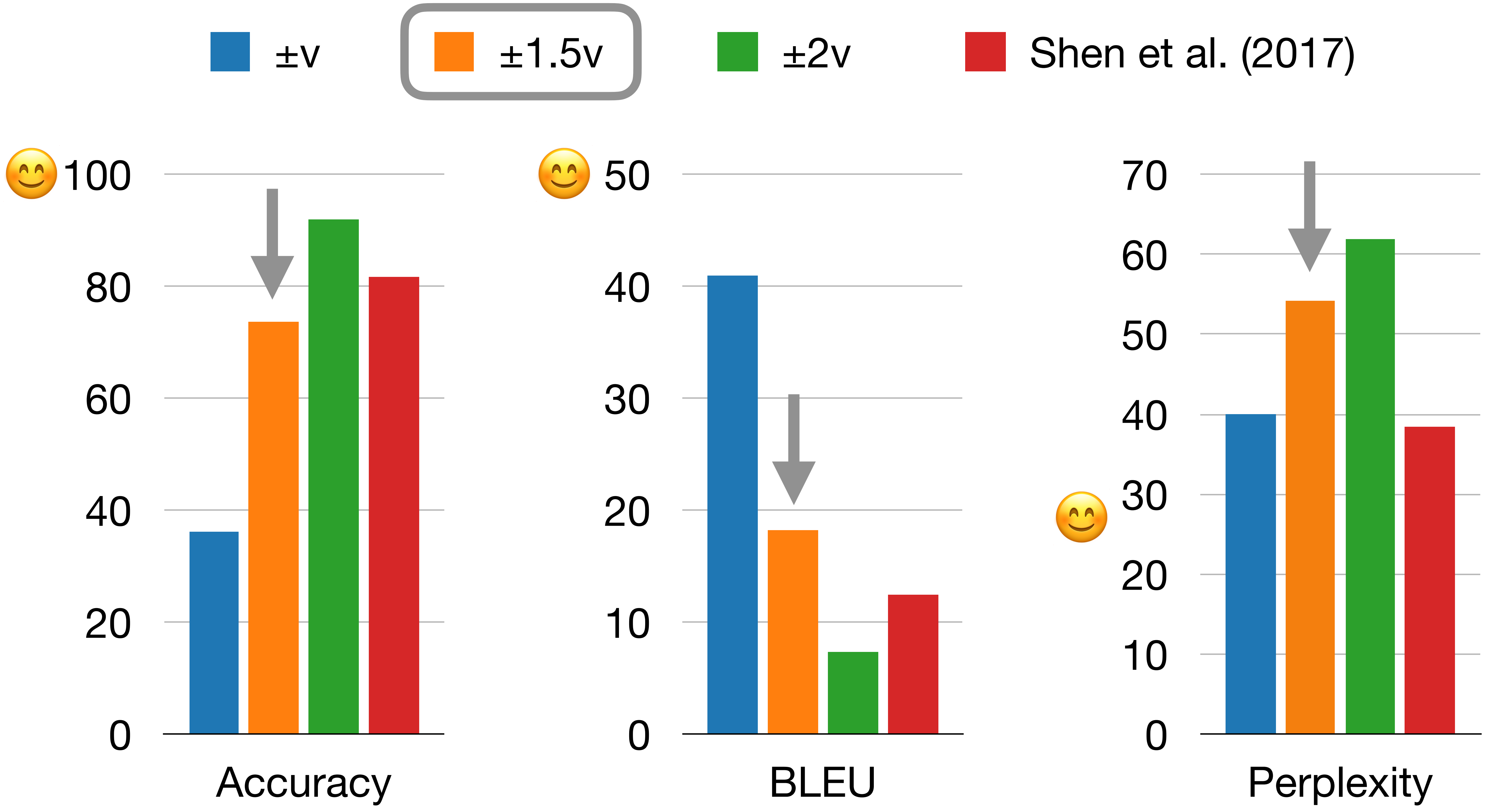
Sentiment Transfer

- As the scaling factor increases, the resulting sentences generated by DAAE get more and more positive/negative

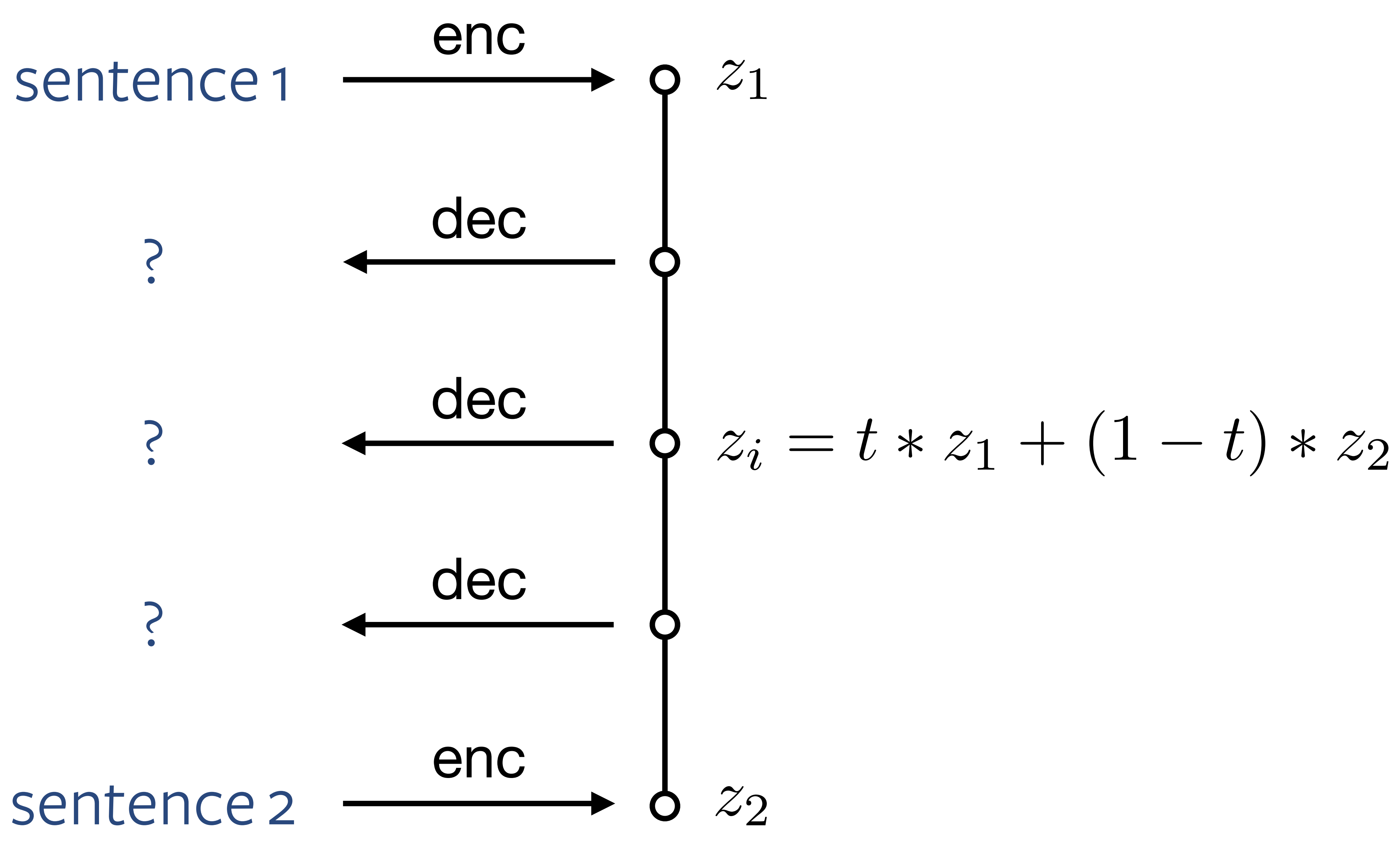


Sentiment Transfer

- DAAE with $\pm 1.5v$ is comparable to previous models trained with sentiment labels [Shen et al., 2017]



Sentence Interpolation via Latent Space Traversal



MIT CSAIL aws

Sentence Interpolation via Latent Space Traversal

AAE	it 's so much better than the other chinese food places in this area . it 's so much better than the other food places in this area . better , much better . better than other places . better than other places .
DAAE	it 's so much better than the other chinese food places in this area . it 's much better than the other chinese places in this area . better than the other chinese places in this area . better than the other places in charlotte . better than other places .

MIT CSAIL aws

Takeaways

- Minimizing $D(p_{\text{data}}(x) \| p_{\text{model}}(x))$ does NOT ensure X-structure is preserved in Z-space
- Denoising helps induce latent space organization
- DAAE best preserves sequence neighborhood, provides superior generation-reconstruction trade-off, and enables *zero-shot* style transfer

Moving Forward

- Better/task-specific text perturbations
- Additional properties of latent space geometry
- Finer control over text generation

https://github.com/shentianxiao/text-autoencoders

Thank you!

MIT CSAIL aws