Can One Modality Model Synergize Training of Other Modality Models?

Jae-Jun Lee, Sung Whan Yoon

Ulsan National Institute of Science and Technology
johnjaejunlee95@unist.ac.kr , shyoon8@unist.ac.kr

25 April, 2025, @ICLR 2025, Singapore

ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

1. Introduction

Huge Success of recent Multimodal Learning

1. Introduction

Huge Success of recent Multimodal Learning

- **Rise of *Foundation Models* (Large Models)**

Foundation Models

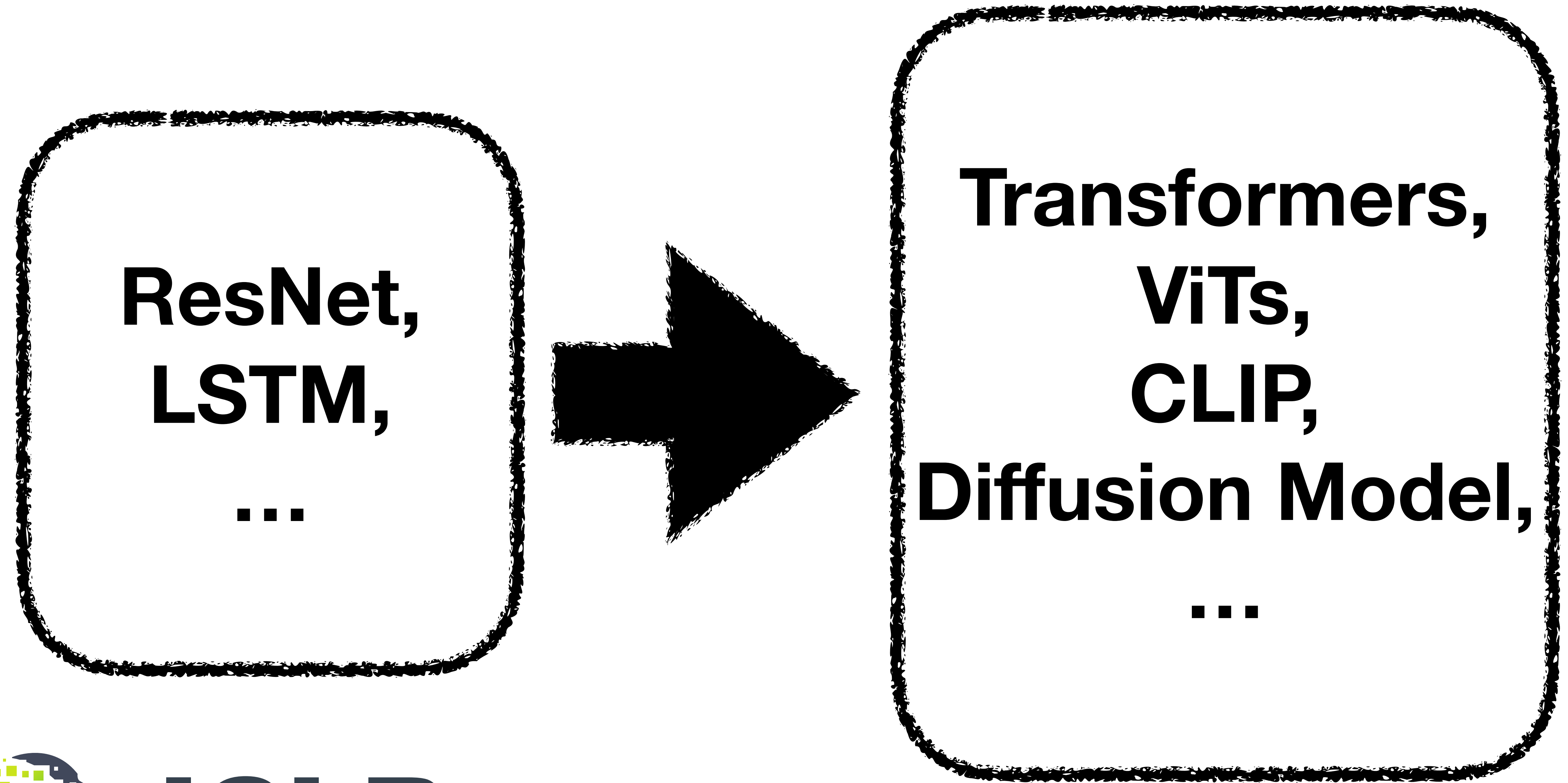
ResNet, LSTM, … → **Transformers, ViTs, CLIP, Diffusion Model, …**

1. Introduction

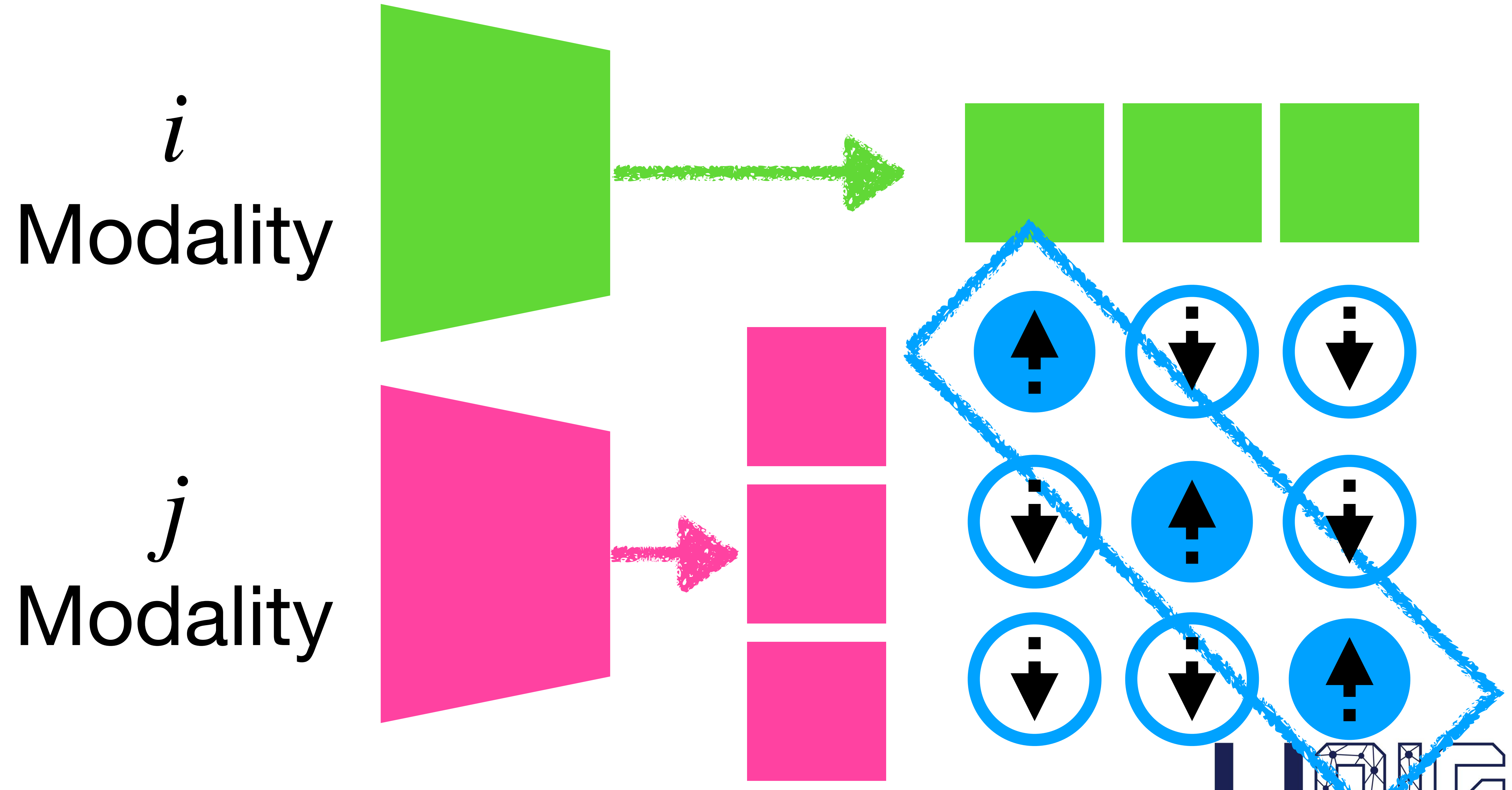
Huge Success of recent Multimodal Learning

- **Rise of *Foundation Models* (Large Models)**
- ***Contrastively leverage information* across different modalities.**

Foundation Models



Train contrastively across modalities



ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

1. Introduction

Problem Formulation - Limitations of recent Multimodal Learning

- However, ***significant limitations*** remain:

1. Introduction

Problem Formulation - Limitations of recent Multimodal Learning

- However, ***significant limitations*** remain:
  - ***Require high-quality data*** **describing each modality sufficiently.**

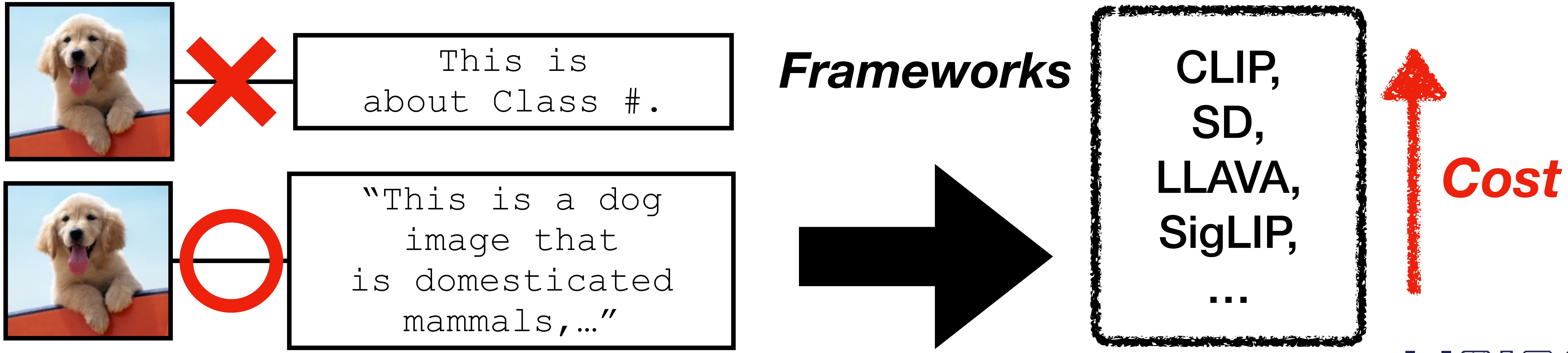
ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

1. Introduction

Problem Formulation - Limitations of recent Multimodal Learning

- However, ***significant limitations*** remain:
  - ***Require high-quality data*** **describing each modality sufficiently.**
  - Multimodal theoratical aspects focus ***when paired-datasets are available, where it requires high computational cost.***



1. Introduction

Our Approach: Synergistic Multimodal Learning in 2 Perspectives

1. Introduction

Our Approach: Synergistic Multimodal Learning in 2 Perspectives

- ***Theoratical Perspective*: Derive how one modality can promote the training of other modality mathematically based on *2-Wasserstein distance between distribution of latent features of each modality*, where it reveals that it doesn't requires high quality of paired-datasets.**

ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

1. Introduction

Our Approach: Synergistic Multimodal Learning in 2 Perspectives

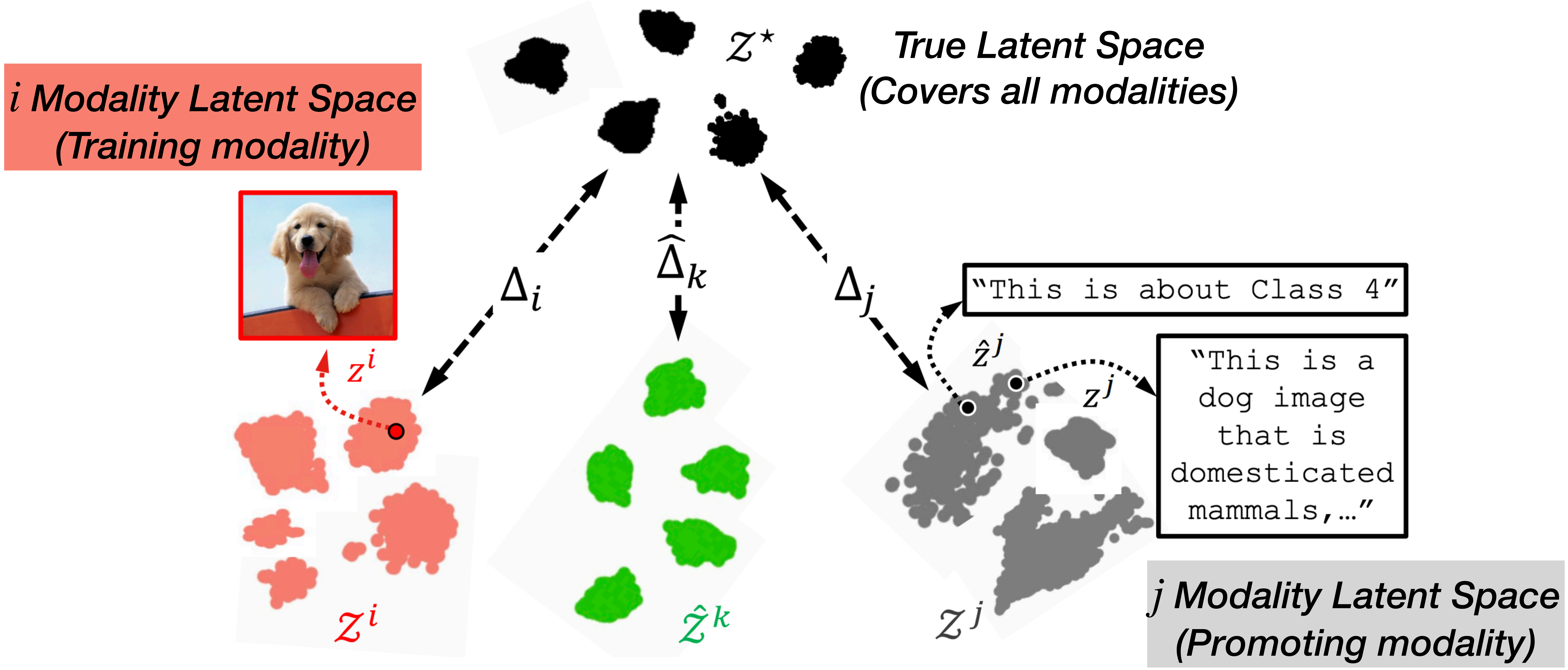
- ***Theoratical Perspective*****:** Derive how one modality can promote the training of other modality mathematically based on ***2-Wasserstein distance between distribution of latent features of each modality***, where it reveals that it doesn't requires high quality of paired-datasets.
- ***Empirical Perspective*****:** Demonstrates how a pretrained modality model can aid in training another modality, even with ***imperfect supervision between paired datasets***.

ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

2. Motivation (Theoratical Perspective)

Skeptual Concept *based on Our Hypotheses*

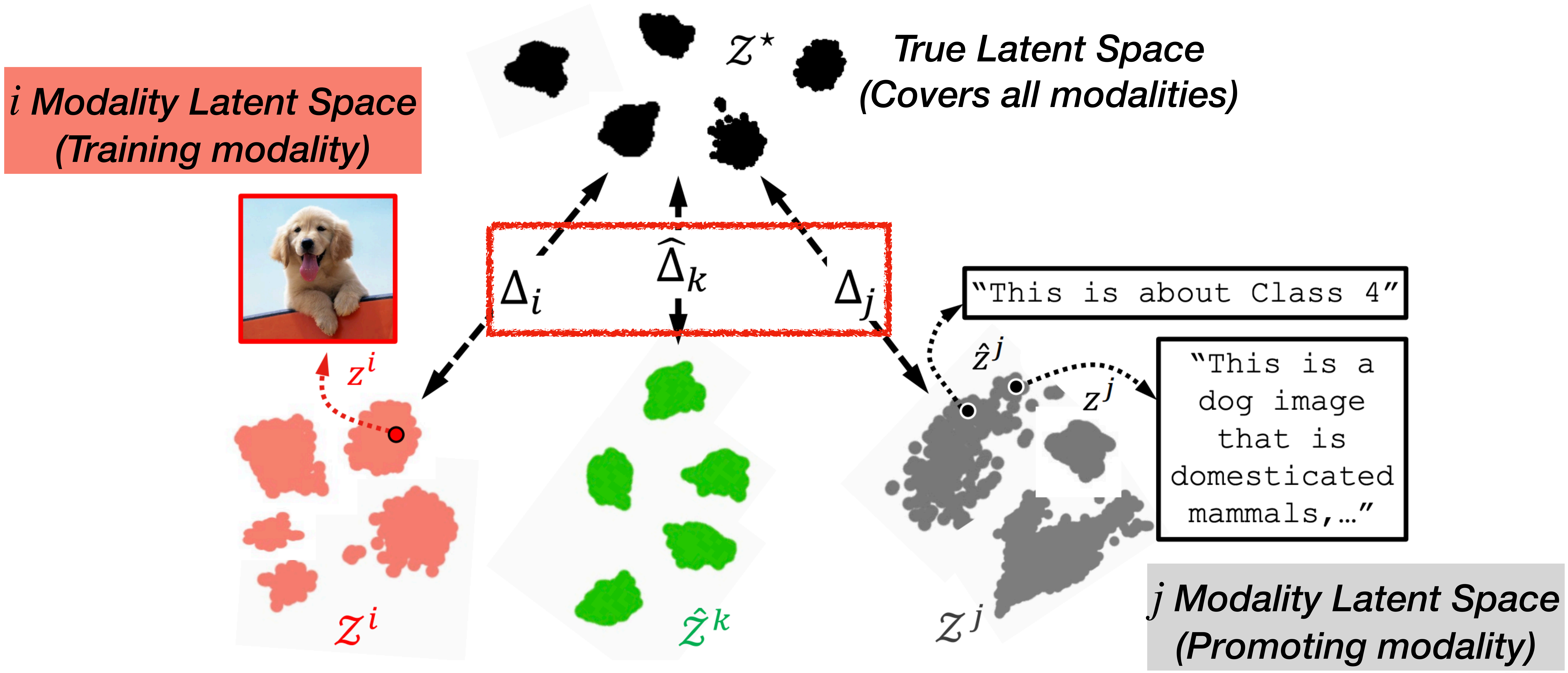


ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

2. Motivation (Theoratical Perspective)

***Skeptual Concept** based on Our Hypotheses*

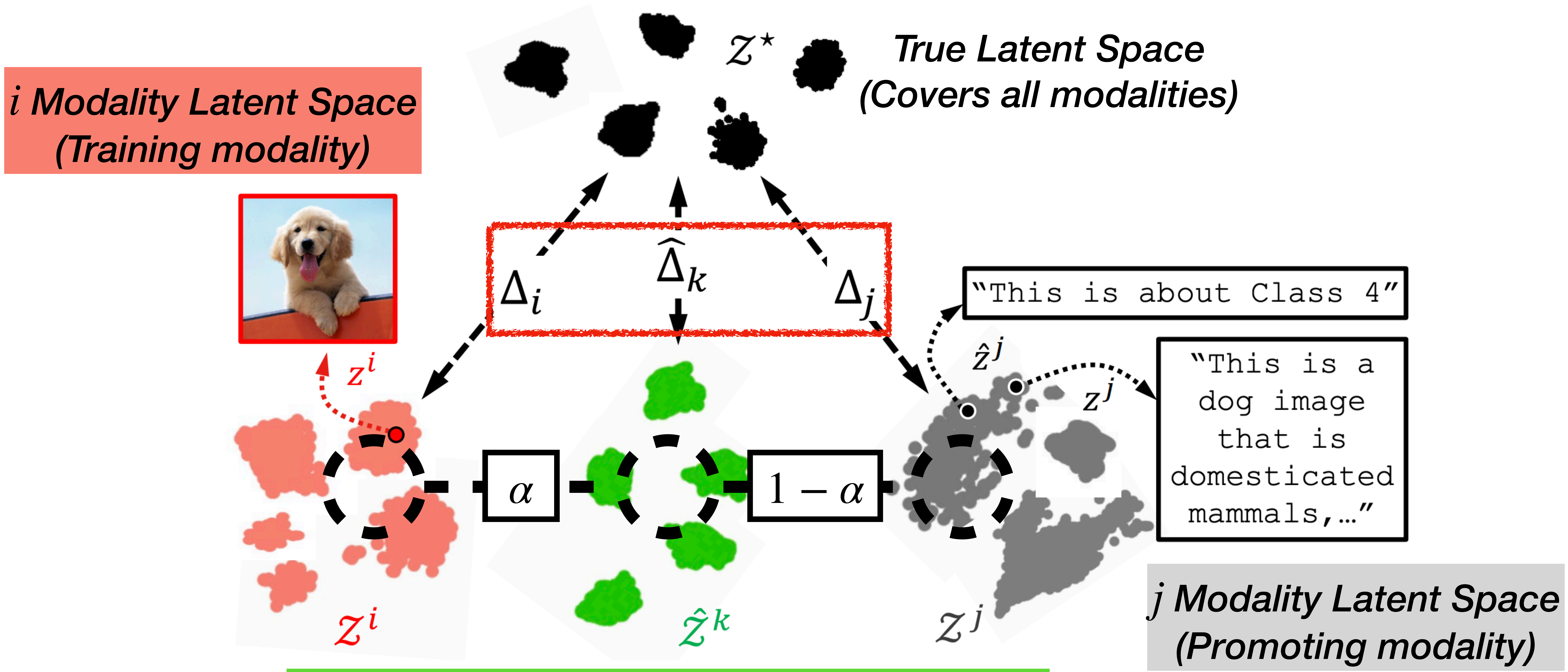


ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

2. Motivation (Theoratical Perspective)

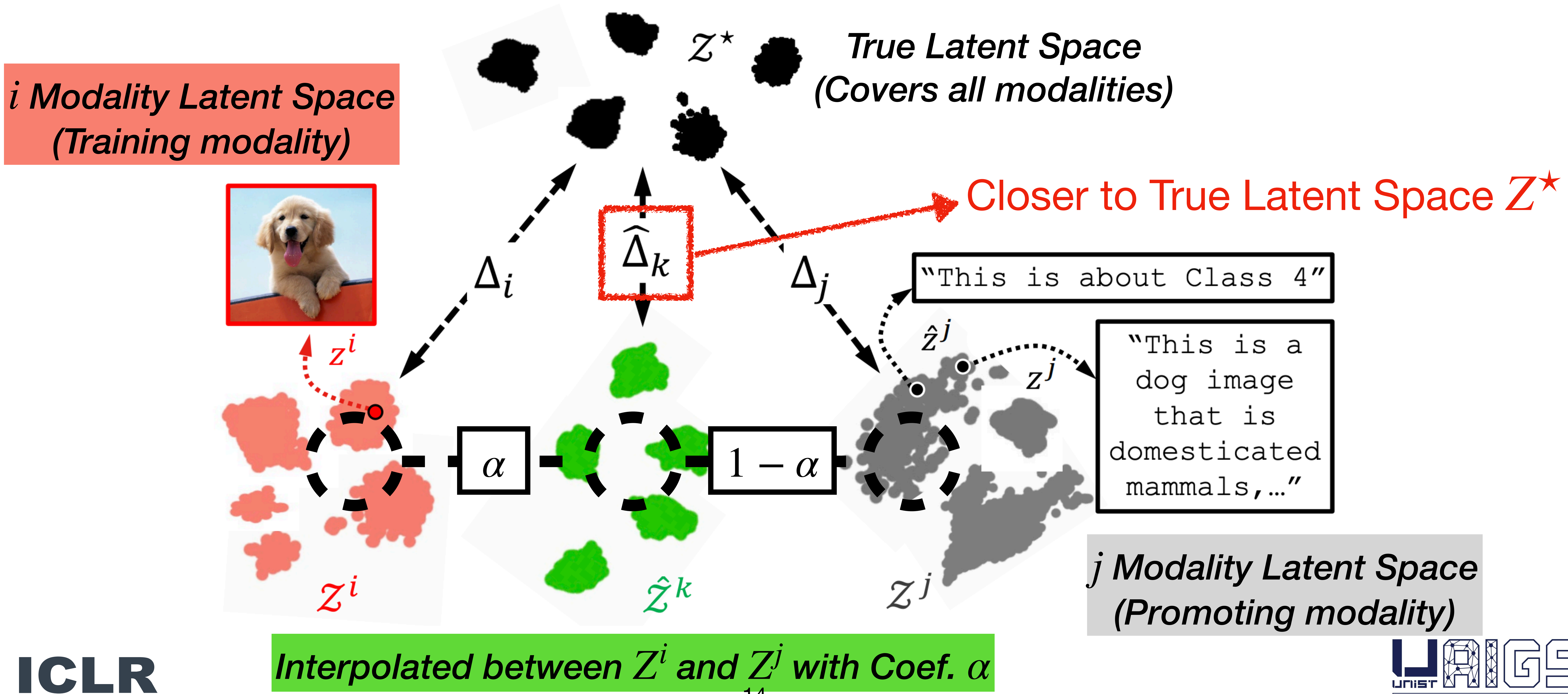
Skeptual Concept *based on Our Hypotheses*



Interpolated between Z^i and Z^j with Coef. α

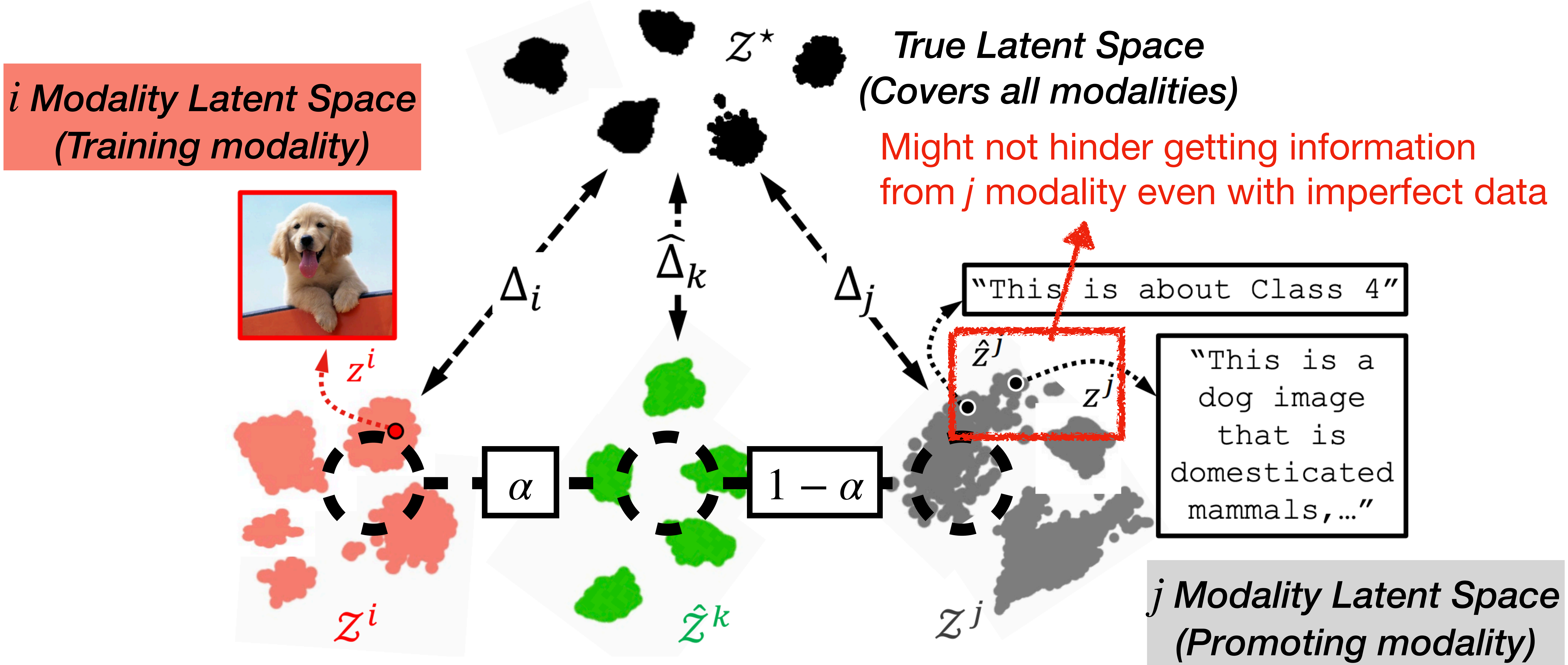
2. Motivation (Theoratical Perspective)

Skeptual Concept *based on Our Hypotheses*



2. Motivation (Theoratical Perspective)

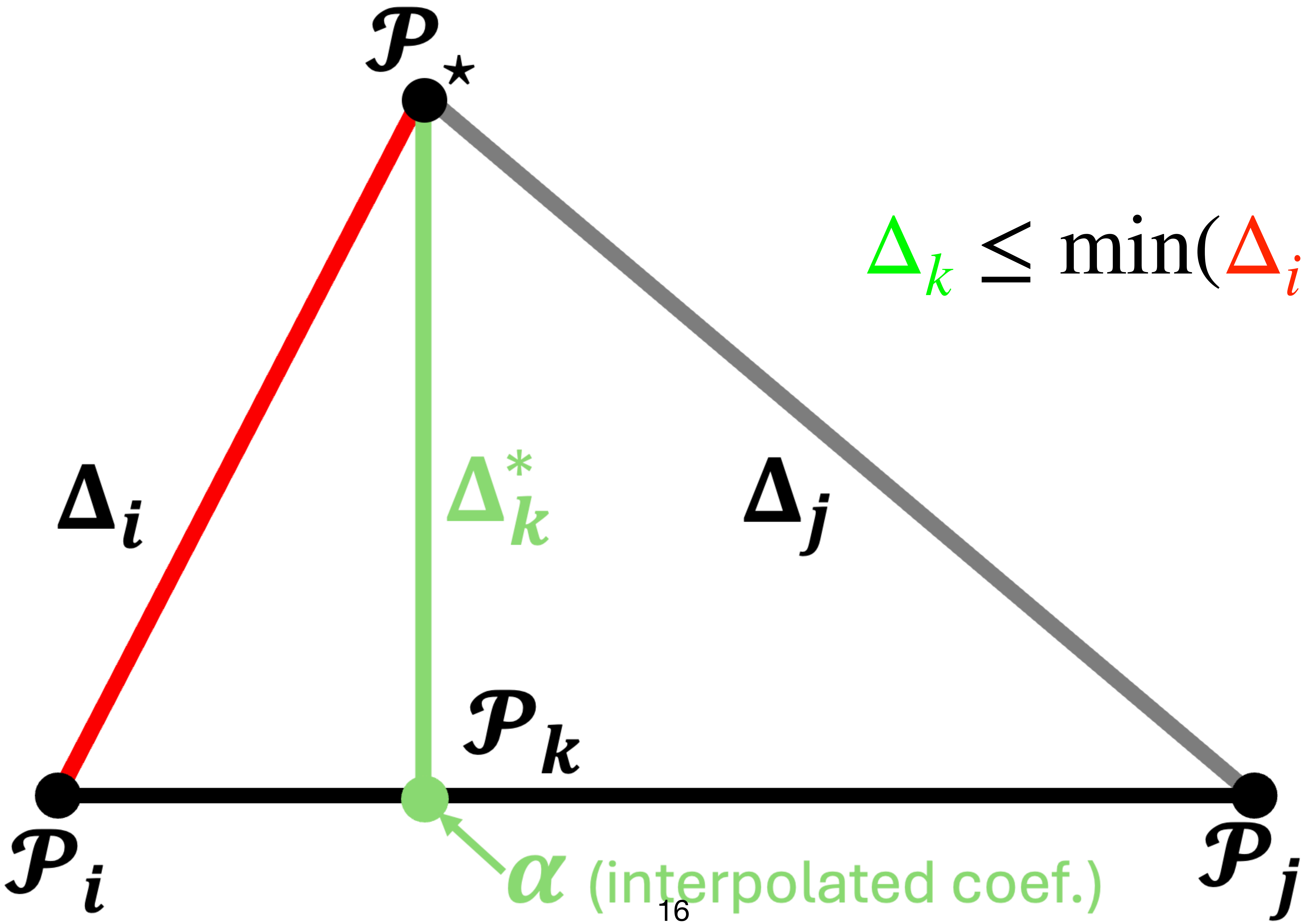
Skeptual Concept *based on Our Hypotheses*



Interpolated between Z^i and Z^j with Coef. α

2. Motivation (Theoratical Perspective)

Skeptual Concept *based on Our Hypotheses* ***(Easier verison)***



$$\Delta_k \leq \min(\Delta_i, \Delta_j)$$

3. Experiment (Empirical Perspective)

Experimental Settings for Synergistic Multimodal Learning

- **Imperfect Supervision:**
  - Conducting ***imperfectly paired datasets***, where paired data provide only partial or insufficient descriptions of each other.
  - Ex). *[*  *, "This is about class #."]* ⇒ ***Imperfect (Vision, Text)***

3. Experiment (Empirical Perspective)

Experimental Settings for Synergistic Multimodal Learning

- **Imperfect Supervision:**
  - Conducting ***imperfectly paired datasets***, where paired data provide only partial or insufficient descriptions of each other.
  - Ex). [🐕‍🦺 , *"This is about class #."*] ⇒ ***Imperfect (Vision, Text)***
- **Matching modalities at Latent Feature Space (or Subspace):**
  - Each modality is analyzed and compared within the latent feature space, extracted from the each modality model. ⇒ ***Need new loss functions***

ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

3. Experiment (Empirical Perspective)

Loss Functions

Classification Loss: $\mathcal{L}_{cls} = \underset{(\mathbf{x}_m^i, y_m^i) \sim \mathcal{S}^i}{\mathbb{E}} \left[\mathcal{L}_{CE} \left(h \circ g(\mathbf{x}_m^i), y_m^i \right) \right]$

Latent Loss: $\mathcal{L}_z = \underset{(\mathbf{x}_m^i, y_m^i, \hat{z}_m^i) \sim \mathcal{S}^i \times \hat{Z}^j}{\mathbb{E}} \left[|| g(\mathbf{x}_m^i) - \hat{z}_m^i ||_2^2 \right]$

$\Rightarrow$ **Total Loss:** $\mathcal{L}_{\textbf{total}} = (1 - \alpha)\mathcal{L}_{\textbf{cls}} + \alpha \mathcal{L}_{\mathbf{z}}$

ICLR

UNIST AIGS

ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

3. Experiment (Empirical Perspective)

Experimental Settings of $\hat{z}^j$ (imperfect supervision)

Datasets & Cases	**Implementation of $\hat{z}^j_m$**
ImageNet-1k [**L**→**V**]	**[L]** $\Rightarrow$ `This is about Class #.`†
IEMOCAP [**L**→**A**]	**[L]** $\Rightarrow$ `This is about Emotion #.`†
IEMOCAP [**A**→**L**]	**[A]** $\Rightarrow$ Add Gaussian Noise: $\xi \sim \mathcal{N}(0, \lambda I)^{\dagger\dagger}$ & Random Shuffling
AVMNIST [**V**→**A**]	**[V]** $\Rightarrow$ Random Shuffled Image (mismatch paired sets)
AVMNIST [**A**→V]	**[A]** $\Rightarrow$ Add Gaussian Noise: $\xi \sim \mathcal{N}(0, \lambda I)^{\dagger\dagger}$ & Random Shuffling

†: # is a random number that does not directly correspond to the actual label.
††: λ is a parameter that controls the variance of the Gaussian noise. We applied $\lambda = 10^{-3}$

ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

3. Experiment (Empirical Perspective)

Empirical Results: Vision-Langauge

Table 1: Classification results on ImageNet-1K and evaluation benchmarks (OOD and robustness)

Model [L→V]	IN	V2	Rend.	Sketch	A	Style.	C (↓)
ResNet-50 (reproduced)	77.83	66.20	39.28	27.35	6.44	8.59	66.01
+ **BERT** (Devlin et al., 2018)	78.41	67.10	40.38	28.19	**8.47**	**9.64**	**64.96**
+ **RoBERTa** (Liu et al., 2019)	**78.54**	**67.30**	**40.92**	**28.78**	8.25	9.19	65.32
ViT-B/32 (reproduced)	75.04	62.02	40.31	27.34	9.23	16.56	55.45
+ BERT (Devlin et al., 2018)	76.59	63.37	41.28	28.53	11.31	18.11	53.28
+ **RoBERTa** (Liu et al., 2019)	**76.75**	**64.00**	**41.81**	**29.50**	**11.55**	**18.75**	**52.95**
ViT-B/16 (reproduced)	80.07	68.60	44.72	31.22	24.20	18.81	51.21
+ **BERT** (Devlin et al., 2018)	81.62	70.07	**45.72**	33.13	25.12	**20.31**	49.27
+ **RoBERTa** (Liu et al., 2019)	**81.90**	**70.55**	45.41	**33.19**	**26.89**	19.93	**48.51**

3. Experiment (Empirical Perspective)

Empirical Results: Language-Audio, Vision-Audio

Table 2: Classification results on IEMOCAP and AVMNIST datasets on each cases of $[M_j \rightarrow M_i]$.

Datasets	Model [L→A]	Accuracy	Model [A→L]	Accuracy
IEMOCAP††	Wav2Vec2† (Ravanelli et al., 2021)	59.46	BERT (Devlin et al., 2018)	55.81
	+ BERT-B (Devlin et al., 2018)	60.44	+ **Wav2Vec2-B** (Baevski et al., 2020)	**56.49**
	+ **BERT-L** (Devlin et al., 2018)	**61.20**	+ Wav2Vec2-L (Baevski et al., 2020)	56.05
Datasets	**Model [V→A]**	**Accuracy**	**Model [A→V]***	**Accuracy**
AVMNIST	Audio Model (Li et al., 2023)	41.28	Vision Model (Li et al., 2023)	65.18
	+ ResNet-18 (He et al., 2016)	42.08	+ Wav2Vec2-B (Baevski et al., 2020)	66.37
	+ **ResNet-34** (He et al., 2016)	**42.44**	+ **Wav2Vec2-L** (Baevski et al., 2020)	**66.69**

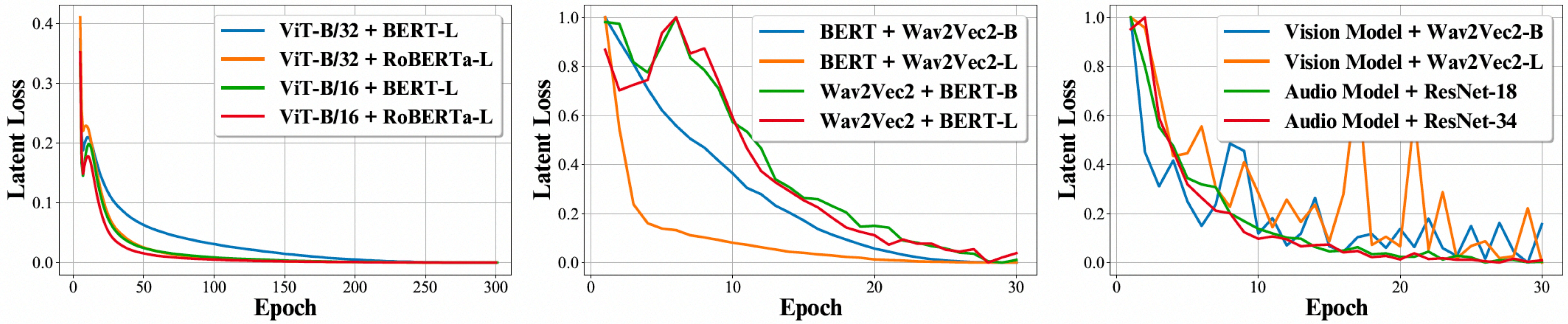
†: SpeechBrain (Ravanelli et al., 2021) experimented with 4 out of 6 labels; we used the all labels.
††: Owing to transformer-type model requires numerous data, we fine-tuned the pretrained model.
*: Since the audio data in AVMNIST is based on spectrograms, we use the original raw audio data prior to its conversion into spectrogram.

ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

4. Analysis

Convergence of Latent Loss (Magnitude of Losses)



(a) ImageNet-1K $\mathcal{L}_z$ (b) IEMOCAP $\mathcal{L}_z$ (c) AVMNIST $\mathcal{L}_z$

⇒ All cases converges almost to, but not exactly to, zero due to interpolation.

4. Analysis

Wasserstein Distance between Paired Modalities

IEMO. [L→A]	WD
$W_2(\mathcal{P}_A, \hat{\mathcal{P}}_k)$	0.494
$W_2(\hat{\mathcal{P}}_L, \hat{\mathcal{P}}_k)$	0.141
$W_2(\mathcal{P}_A, \hat{\mathcal{P}}_L)$	**0.977**

IEMO. [A→L]	WD
$W_2(\mathcal{P}_L, \hat{\mathcal{P}}_k)$	0.965
$W_2(\hat{\mathcal{P}}_A, \hat{\mathcal{P}}_k)$	0.460
$W_2(\mathcal{P}_L, \hat{\mathcal{P}}_A)$	**1.005**

AVMN. [V→A]	WD
$W_2(\mathcal{P}_A, \hat{\mathcal{P}}_k)$	0.025
$W_2(\hat{\mathcal{P}}_V, \hat{\mathcal{P}}_k)$	0.754
$W_2(\mathcal{P}_A, \hat{\mathcal{P}}_V)$	**0.790**

AVMN. [A→V]	WD
$W_2(\mathcal{P}_V, \hat{\mathcal{P}}_k)$	0.908
$W_2(\hat{\mathcal{P}}_A, \hat{\mathcal{P}}_k)$	0.502
$W_2(\mathcal{P}_V, \hat{\mathcal{P}}_A)$	**0.954**

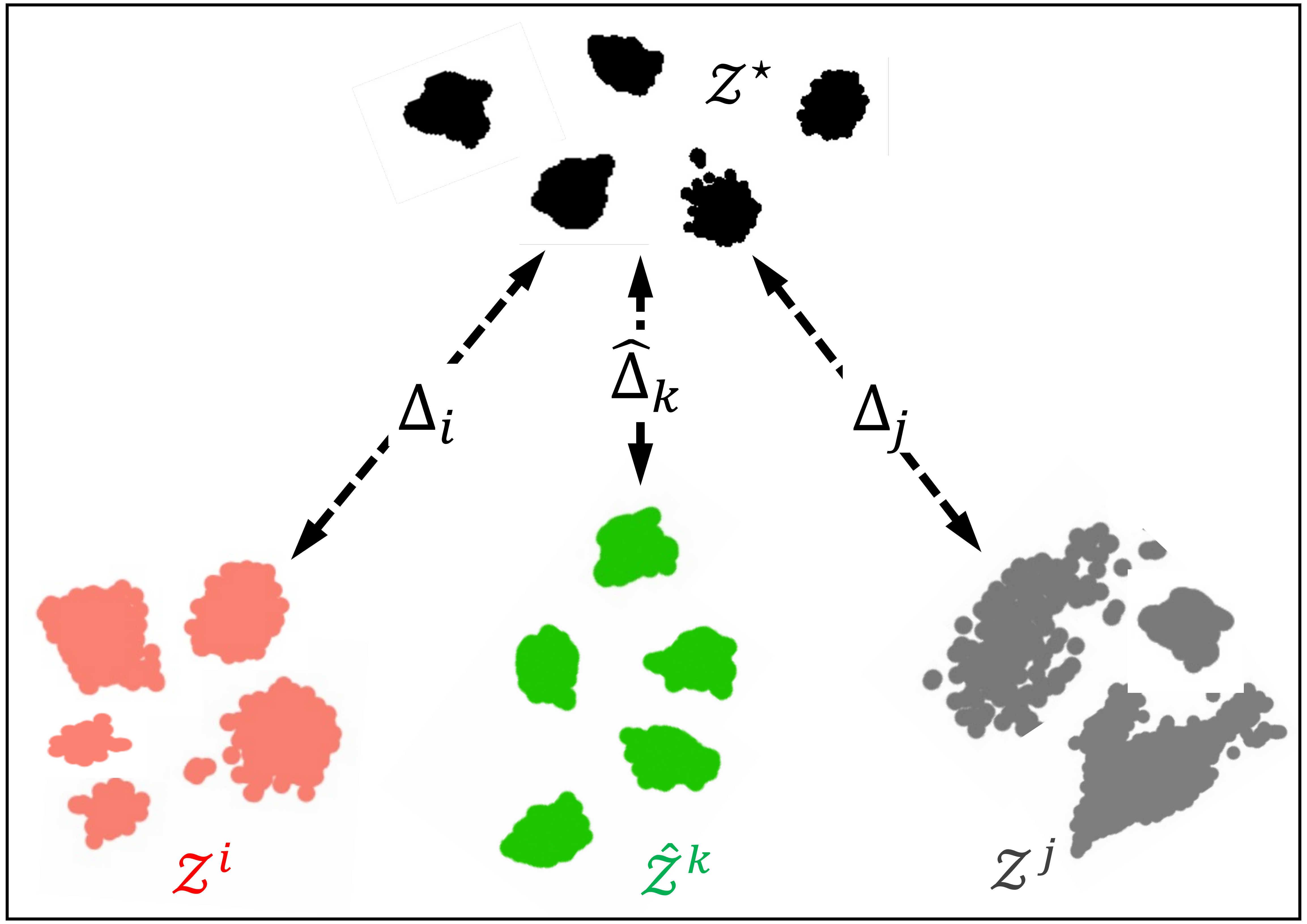
Interpolated representation are both smaller than WD between modalities

ICLR

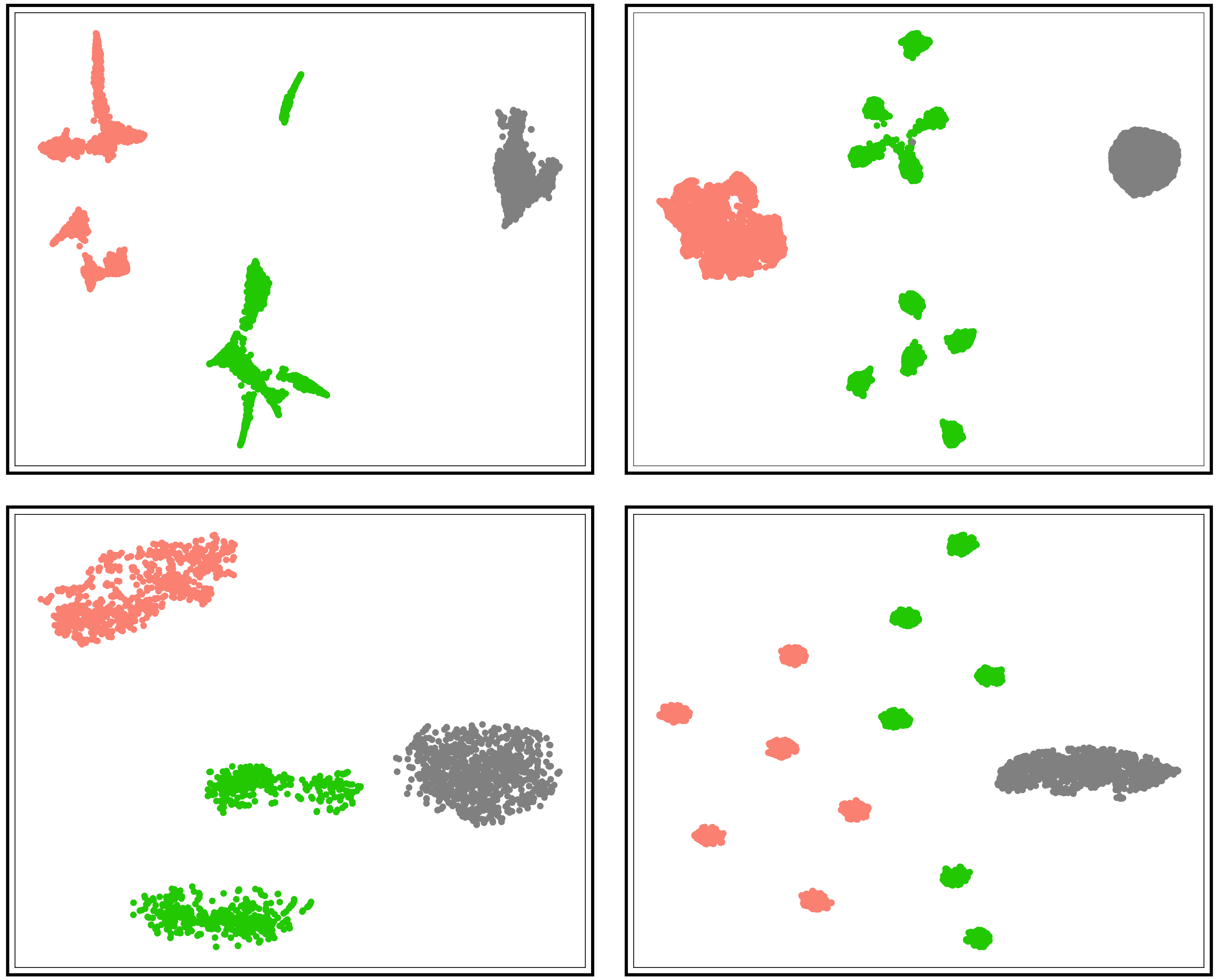
UNIST UAIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

4. Analysis

Wasserstein Distance between Paired-Modalities



Our Hypothesis



Our Results (t-SNE Visualizations)

4. Analysis

Ablation Studies: Usage of Paired Supervision z^j vs. $\hat{z}^j$

Model [L→V]	$\hat{z}^j$	z^j
ResNet-50 + RoBERTa	78.54	78.61 (+0.07)
ViT-B/32 + RoBERTa	76.75	76.99 (+0.24)
ViT-B/16 + RoBERTa	81.90	82.54 (+0.64)

Slightly improved but minimal gains

ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

5. Conclusion

- Our paper demonstrate that **a modality can enhance learning in another**, even with **weakly related or mismatched supervision,**
- Both **theoretical and empirical frameworks** support this finding, reinforcing its validity.
- Exploring more complex multimodal settings, incorporating additional modalities, and scaling to larger models for further advancements.

ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

Thank you!