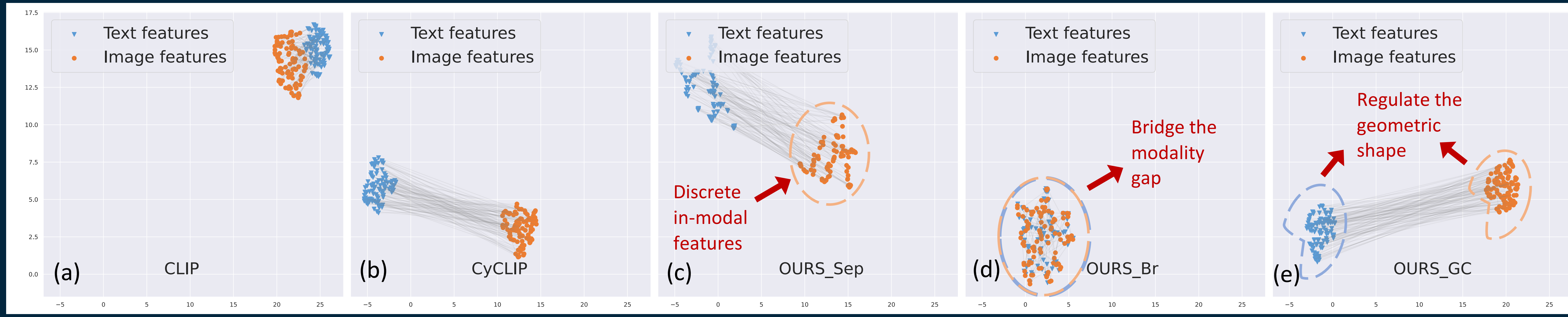
Search

Understanding and Constructing Latent Modality Structures in Multi-Modal Representation Learning

Qian Jiang[1,2], Changyou Chen[1], Han Zhao[1,2], Liqun Chen, Qing Ping[1], Son Dinh Tran[1], Yi Xu[1], Belinda Zeng[1], Trishul Chilimbi[1]

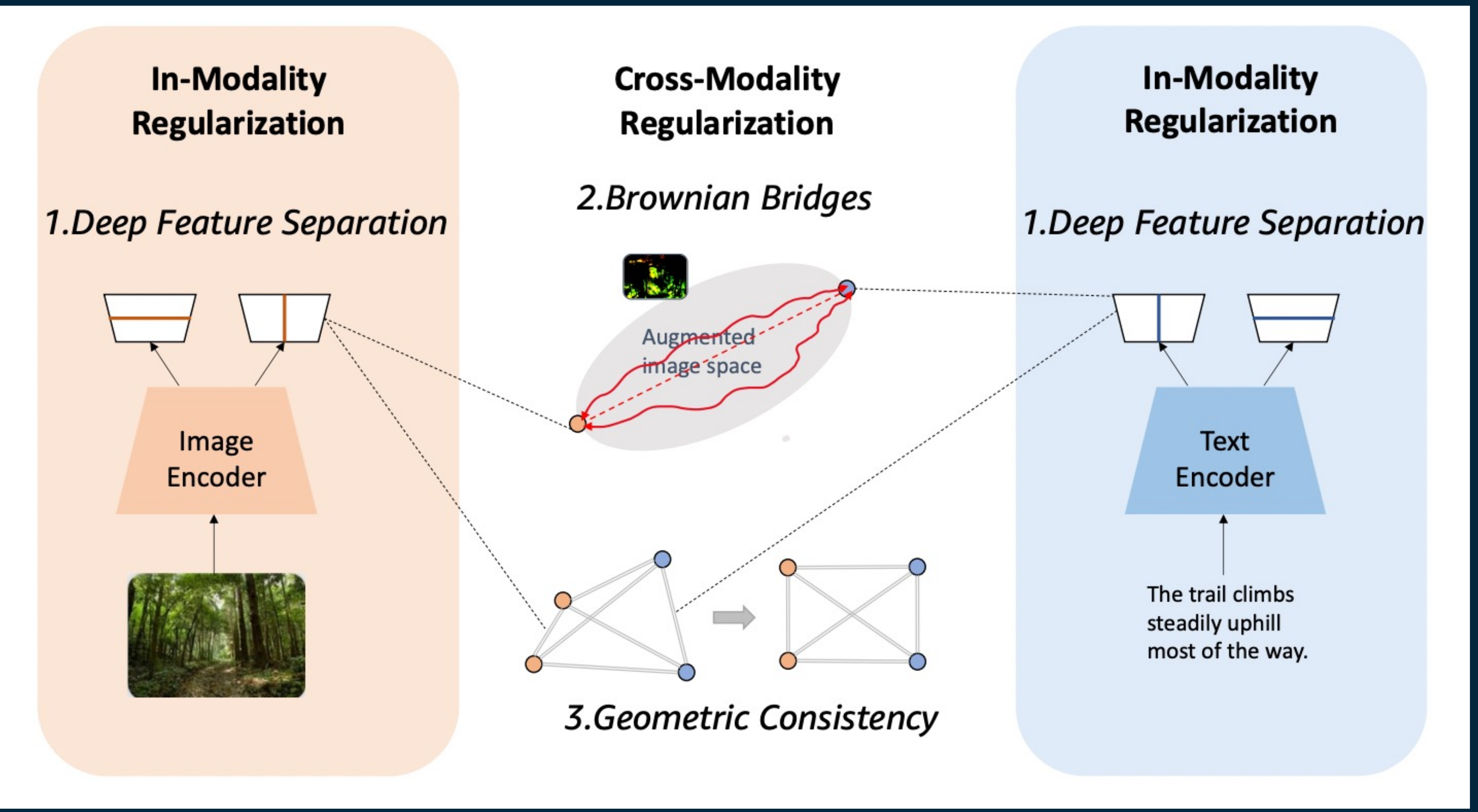
Amazon[1]
University of Illinois at Urbana-Champaign[2]



One-page overview

- We study the impact of modality alignment with empirical and theoretic analysis

- We propose three regularizations to construct latent feature structures
  - intra-modality regularization via deep feature separation
  - inter-modality regularization via Brownian bridge
  - intra-inter-modality regularization via geometric consistency



- We demonstrate improved performance on both two-tower-based models (e.g. CLIP) and fusion-based models (e.g. ALBEF) on a variety of vision-language tasks.

Outline

- Background
- Analysis
- Method
- Experiments
- Summary

Vision-language Pre-training (VLP)

- VLP aims to learn multimodal representations from large-scale image-text pairs
- Many downstream tasks (Image Retrieval/Text Retrieval, Visual Question Answering, etc.) benefit from multi-modal training
- Aligning different modalities plays the crucial rule for obtaining meaningful features

Image modality

Text modality

The trail climbs steadily uphill most of the way

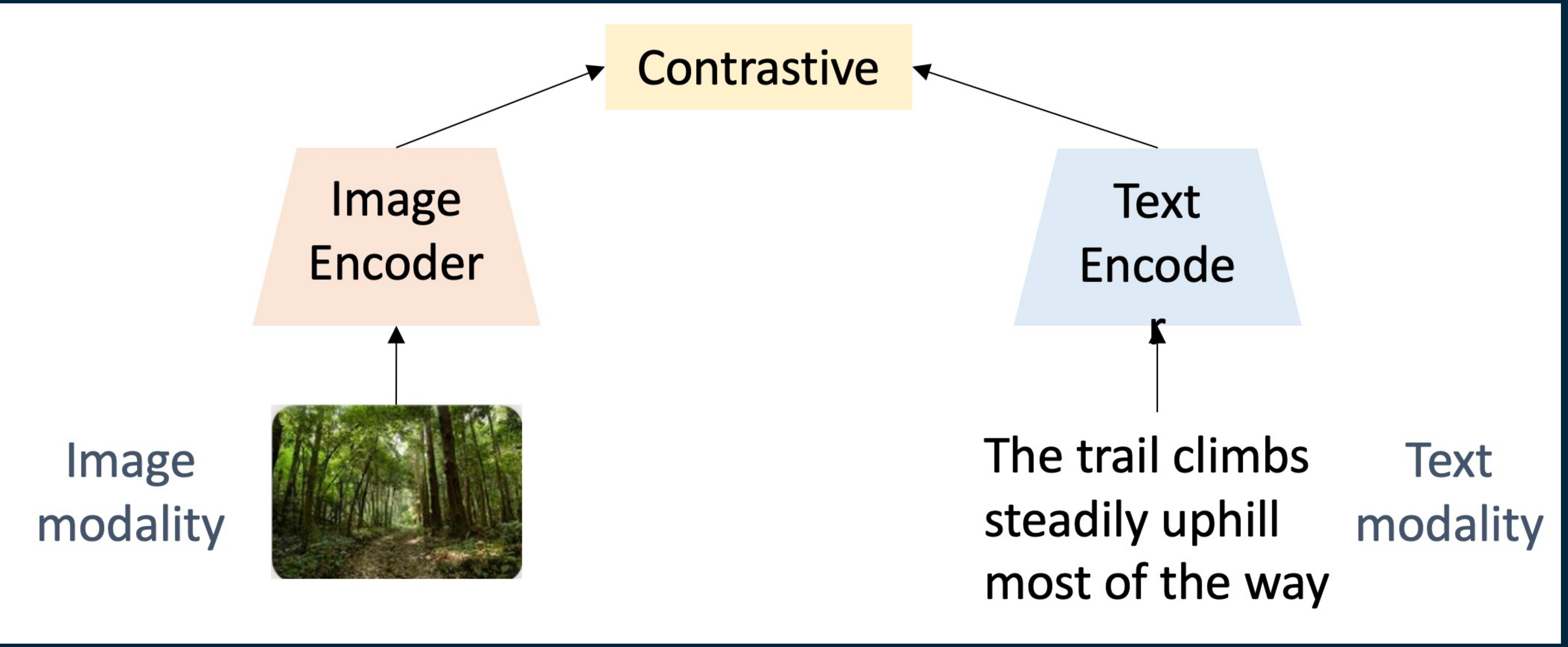
Image-Text Pairs

Vision-language Contrastive Learning

- Typically Joint image-text models are trained with contrastive learning e.g. CLIP

- Model has two separate encoders
- Model receives training samples in pairs

- Learn to align images with their corresponding texts by pulling the positive pairs together and pushing negative pairs apart.

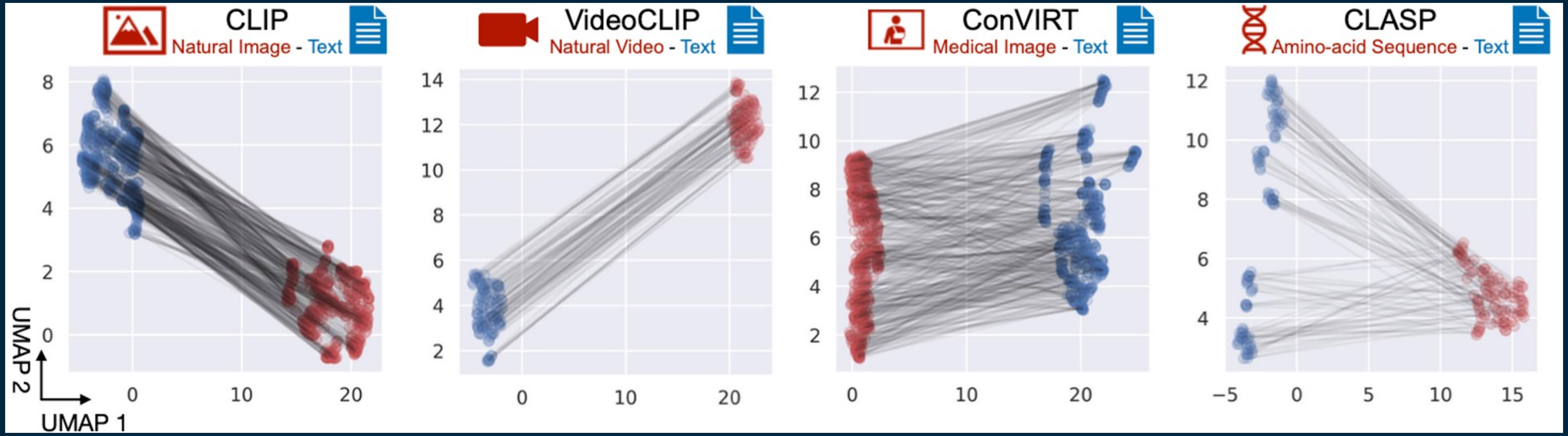


CLIP Framework

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

Modality Gap

In contrastive learning, image and text features still reside in different regions of feature space. Such a phenomenon is called modality gap.



Modality gap in different models

Liang, Weixin, et al. "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning." *arXiv preprint arXiv:2203.02053* (2022).

Understanding Modality Gap

Key question : With the existence of modality gap, how to better align modalities?

- How about perfect alignment? With zero modality gap, we can achieve perfect alignment. Is this the ideal way to go?

  - ✓ Empirical Analysis
  - ✓ Theoretic Analysis

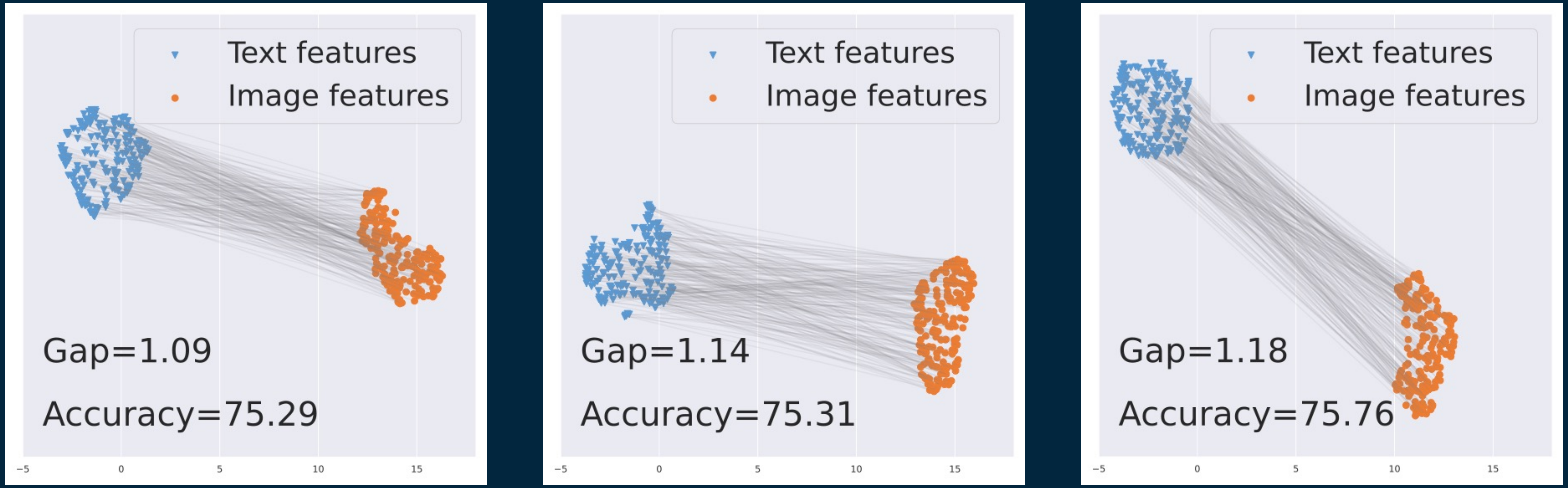
Notations

- X_T and X_V denote the inputs from two modalities
- Y denote the task label
- g_T and g_V denote the modality specific encoders
- $Z_T = g_T(X_T)$ and $Z_V = g_V(X_V)$ denote the extracted features

Empirical Analysis

X_T and X_V : inputs from two modalities
Y : the task label
g_T and g_V: the modality specific encoders
$Z_T = g_T(X_T)$ and $Z_V = g_V(X_V)$: the features

- Adjusting modality gap by optimizing $\mathcal{L}_{\mathrm{Align}} = 1/\langle \mathrm{Z_T}, \mathrm{Z_V} \rangle^2$ with different loss scale
- Train on COCO and evaluate zero-shot image-text retrieval performance on Flick30K



- There is no clear-cut relationship between the gap of these two modalities and the downstream retrieval performance.

Notations

- X_T and X_V denote the inputs from two modalities
- Y denote the task label
- g_T and g_V denote the modality specific encoders
- $Z_T = g_T(X_T)$ and $Z_V = g_V(X_V)$ denote the extracted features
- $I(X_T; X_V)$ denotes the Shannon mutual information between $X_T$and X_V
- $I(X_T; Y)$ denotes the information provided by X_T towards predicting Y
- p denotes the joint distribution of (X_T, X_V, Y)

Theoretic Analysis

X_T and X_V : inputs from two modalities
Y : the task label
Z_T and Z_V : the features
$I(X_T;Y)$: the information provided by X_T towards predicting Y

- Define information gap $\Delta p := |I(X_T;Y) - I(X_V;Y)|$ to characterize the gap of information provided by two modalities towards predicting the target variable Y.

- We prove *Theorem 1* For a pair of modality encoders $g_T(\cdot)$ and $g_V(\cdot)$, if the multi-modal features $Z_T = g_T(X_T)$ and $Z_V = g_V(X_V)$ are perfectly aligned in the feature space, i.e., $Z_T = Z_V$, then $\inf_h \mathbb{E}_p[\ell_{CE}(h(Z_T, Z_V), Y)] - \inf_{h'} \mathbb{E}_p[\ell_{CE}(h'(X_T, X_V), Y)] \geq \Delta p$

- The optimal prediction error we can hope to achieve by using aligned features is at least Δp larger than that we can achieve using the input modalities directly.

- In other words, using perfectly aligned features leads to an information loss of Δp

Implications

- **Recall *Theorem 1*** With perfectly alignment:

$$\inf_{h} \mathbb{E}_p[\ell_{CE}(h(Z_T, Z_V), Y)] - \inf_{h'} \mathbb{E}_p[\ell_{CE}(h'(X_T, X_V), Y)] \geq \Delta p$$

- When Δp is large, i.e. when one modality is much more informative, perfect modality alignment could render the learned aligned features Z_T and Z_V uninformative of Y, leading to a large downstream prediction error

- Features with zero modality gap can only preserve predictive information present in both of the modalities at the cost of losing the modality-specific information

Methods

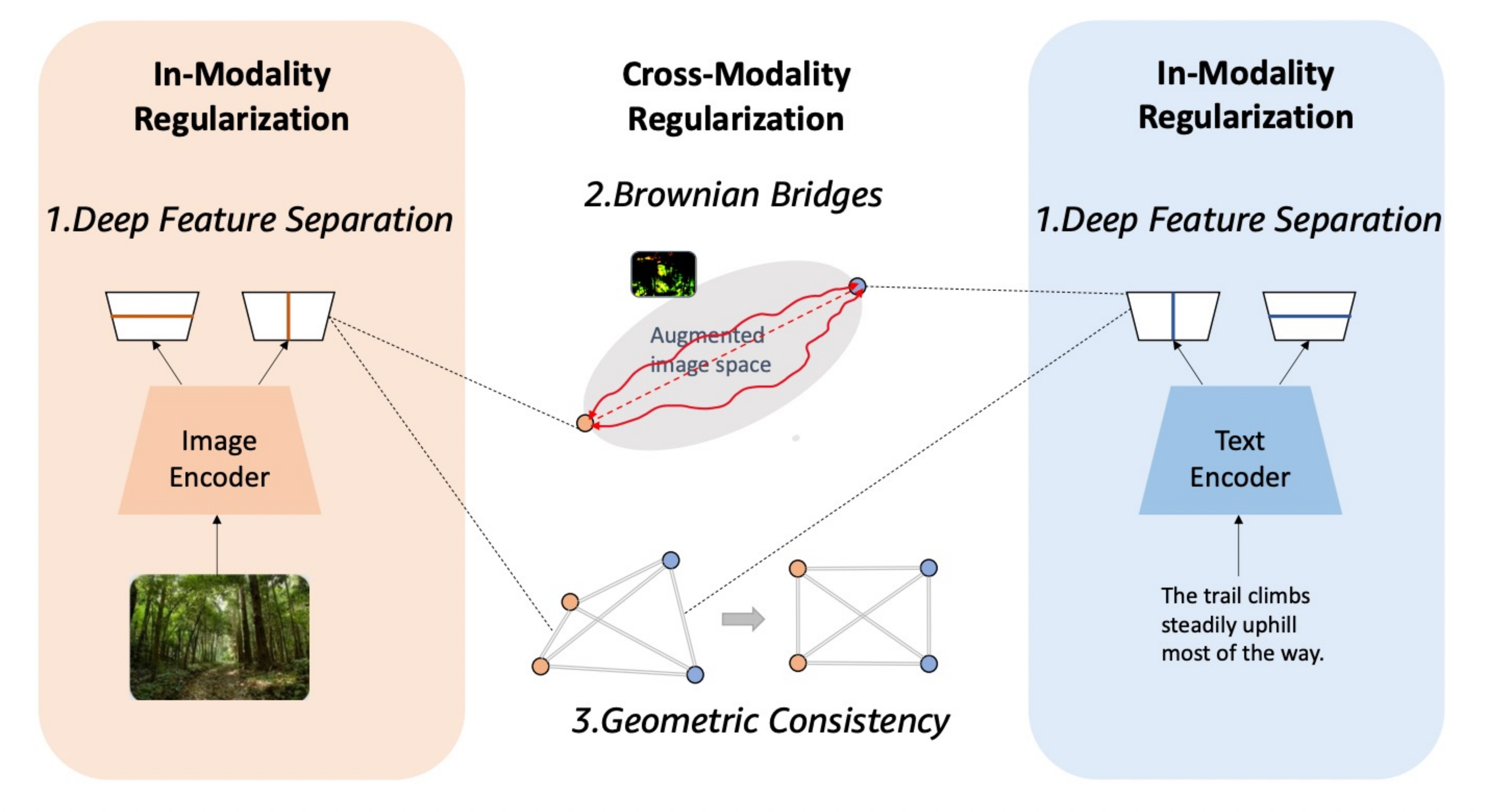
Key question : how to better align modalities?

✗ Perfect alignment ?

✓ More meaningful alignment by constructing latent modality structures:

- Intra-modality regularization
- Inter-modality regularization
- Intra-Inter-modality regularization

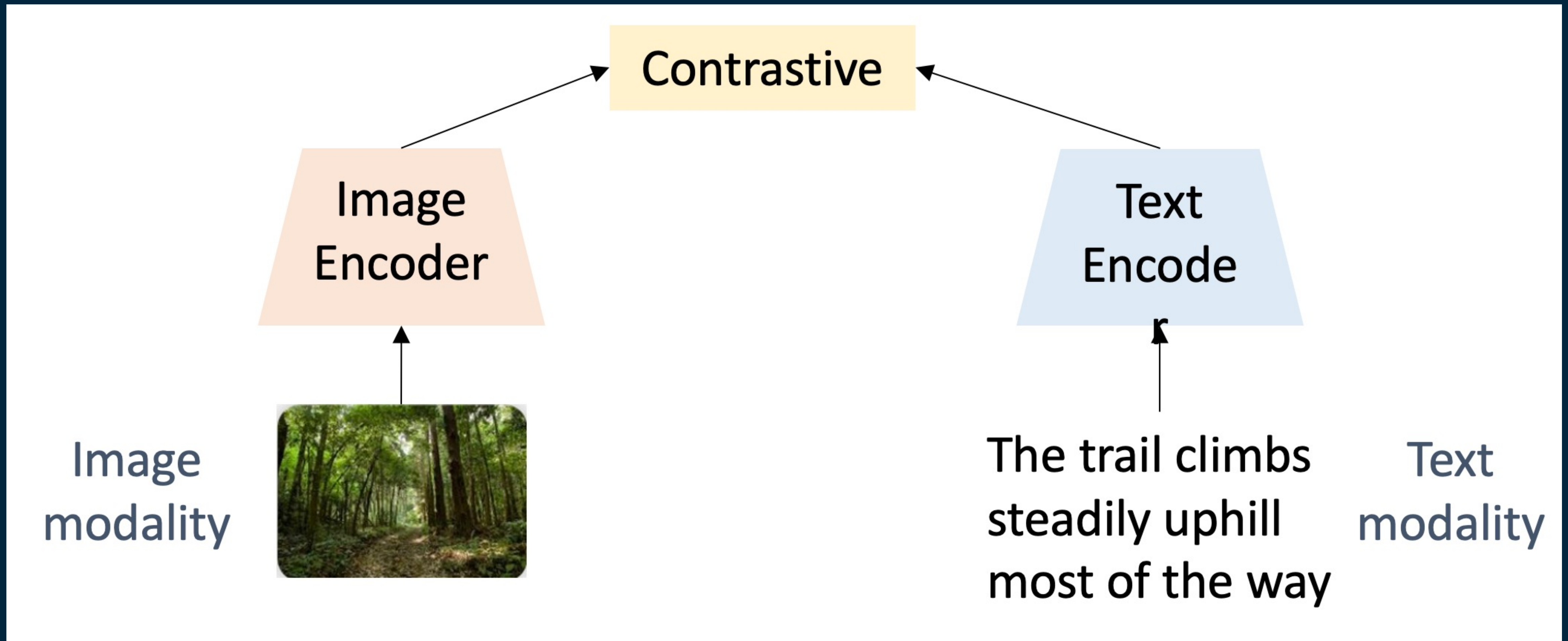
Methods



Overview of methods

Basic contrastive learning framework

- We incorporate our methods over the contrastive learning framework:
- $\mathcal{L}_{Con} = \frac{1}{4}(\mathcal{L}_{V2T} + \mathcal{L}_{T2V} + \mathcal{L}_{V2V} + \mathcal{L}_{T2T})$
- $\mathcal{L}_{V2T} = -\frac{1}{N}\sum_{j=1}^{N} log \frac{e^{\langle z_{V\,j}, z_{T\,j}\rangle/\tau}}{\sum_{k=1}^{N} e^{\langle z_{V\,j}, z_{T\,k}\rangle/\tau}}$
- $\mathcal{L}_{V2V} = -\frac{1}{N}\sum_{j=1}^{N} log \frac{e^{\langle z_{V\,j}, z_{V\,j}^{a}\rangle/\tau}}{\sum_{k=1}^{N} e^{\langle z_{V\,j}, z_{V\,k}\rangle/\tau}}$



Basic contrastive framework

Intra-modality regularization via deep feature separation

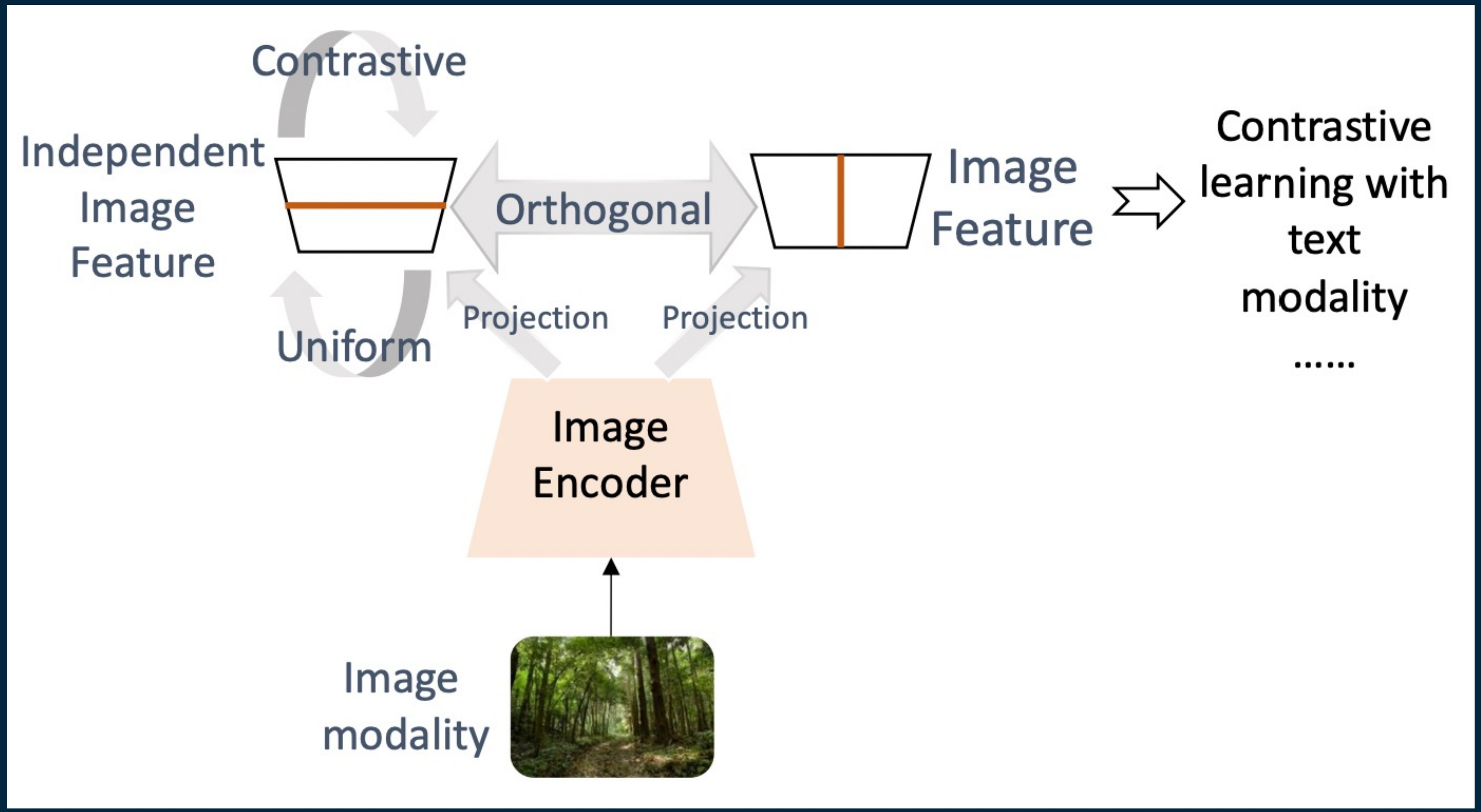
Recall the implication from ***Theorem 1***

- Features with zero modality gap can only preserve predictive information present in both of the modalities at the cost of losing the modality-specific information

- Can we preserve the modality-specific information?
  - use a new feature to store the modality-specific information
  - optimize the new feature to be :
    - ✓ complementary to the original feature
    - ✓ meaningful

Intra-modality regularization via deep feature separation

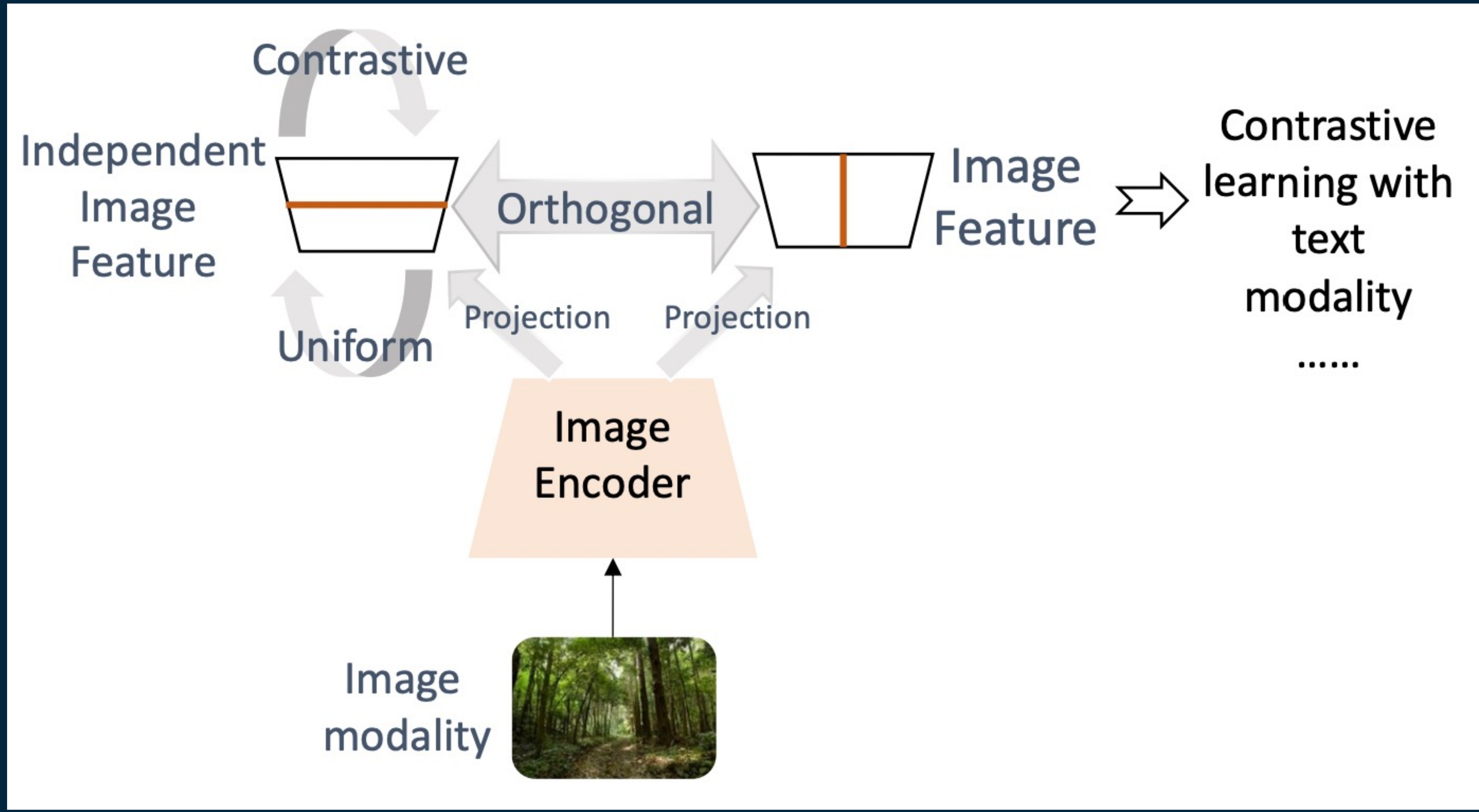
- Use one projection layer to obtain the independent feature
- Optimize the independent feature z_V^i to contain complementary information to the original feature
- Use orthogonal loss to encourage the independent feature to be orthogonal to the original feature:
- $\mathcal{L}_{\mathrm{Ortho}} = \frac{1}{N}\sum_{j=1}^{N}\left\langle z_{V_j}, z_{V_j}^{i}\right\rangle^2$



Intra-modality regularization via deep feature separation

Intra-modality regularization via deep feature separation

- Optimze z_V^i to be meaningful
- Use contrastive loss:
- $\mathcal{L}_{V2V}^i = -\frac{1}{N}\sum_{j=1}^{N} log \frac{e^{\left\langle z_{V\,j}^i, z_{V\,j}^i \right\rangle/\tau}}{\sum_{k=1}^{N} e^{\left\langle z_{V\,j}^i, z_{V\,k}^i \right\rangle/\tau}}$
- Use Uniform loss with Gaussian potential kernel to encourage pairwise difference:
- $\mathcal{L}_{Uni} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N} G_t(z_{V\,j}^i, z_{V\,k}^i)$
- $G_t = e^{-t\|u-v\|^2}, t = 2$



Intra-modality regularization via deep feature separation

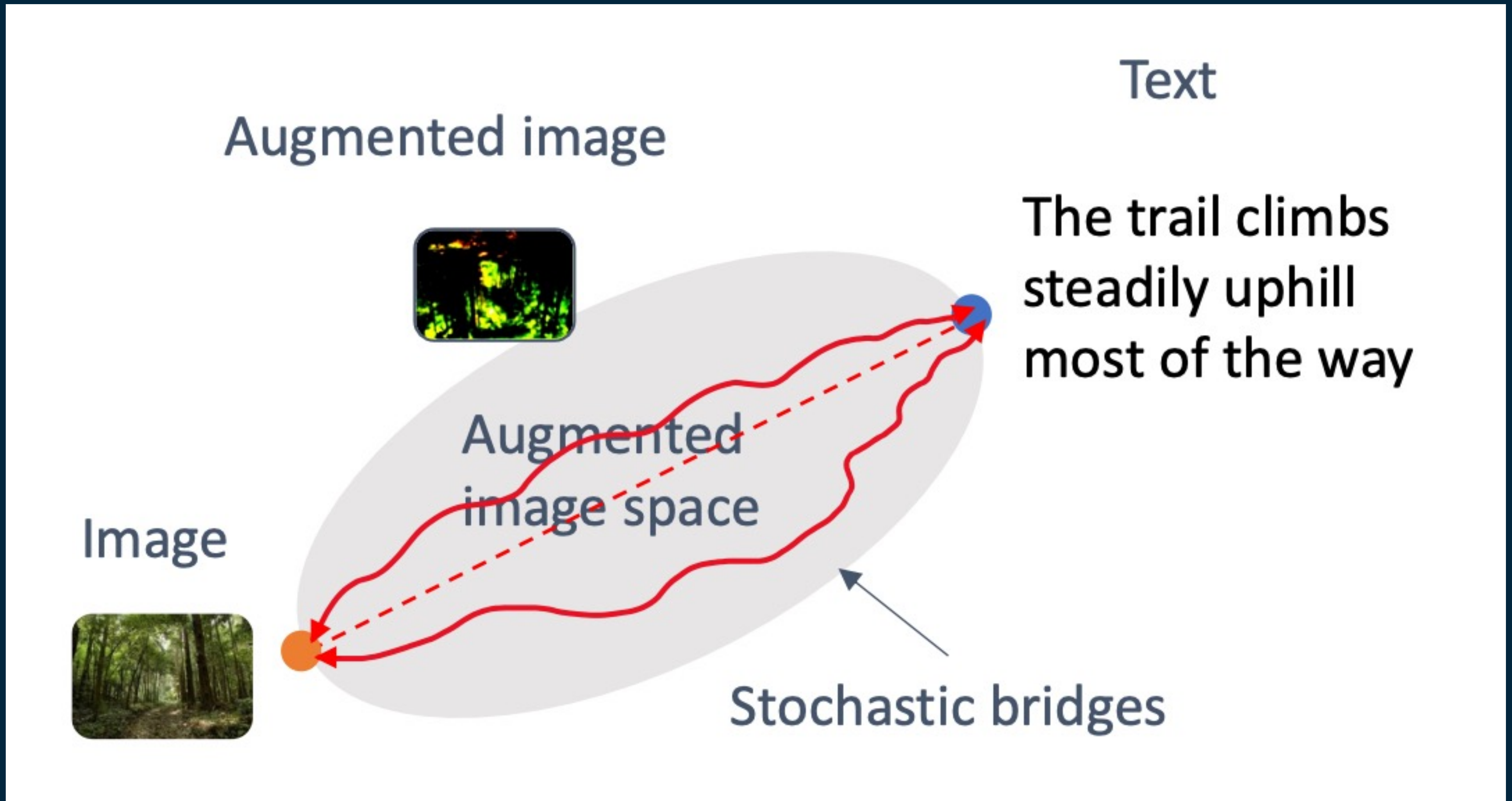
Inter-modality regularization via Brownian Bridge

With the modality gap

- How can we connect two modalities ?
- Use a latent structure to explicitly guide the transition from the image modality to the associated text modality
- Apply Brownian bridge that define stochastic paths (called bridges) between a pair of fixed starting and ending points (corresponding to the two modalities in our setting)

Inter-modality regularization via Brownian Bridge

- Use augmented image feature $\mathrm{Z}_{\mathrm{V}}^{\mathrm{a}}$ to guide the transition
- We define a stochastic path such that $\mathrm{Z}_{\mathrm{V}}^{\mathrm{a}}$ is constrained to stay on the path between Z_{V} and Z_{T}:
- $p(Z_V^a \mid Z_V, Z_T) = N(Z_V^a;\ \mu(Z_V, Z_T, t), t(1-t)\boldsymbol{I})$
- $\mu(Z_V, Z_T, t) = \frac{tZ_V + (1-t)Z_T}{\|tZ_V + (1-t)Z_T\|}$



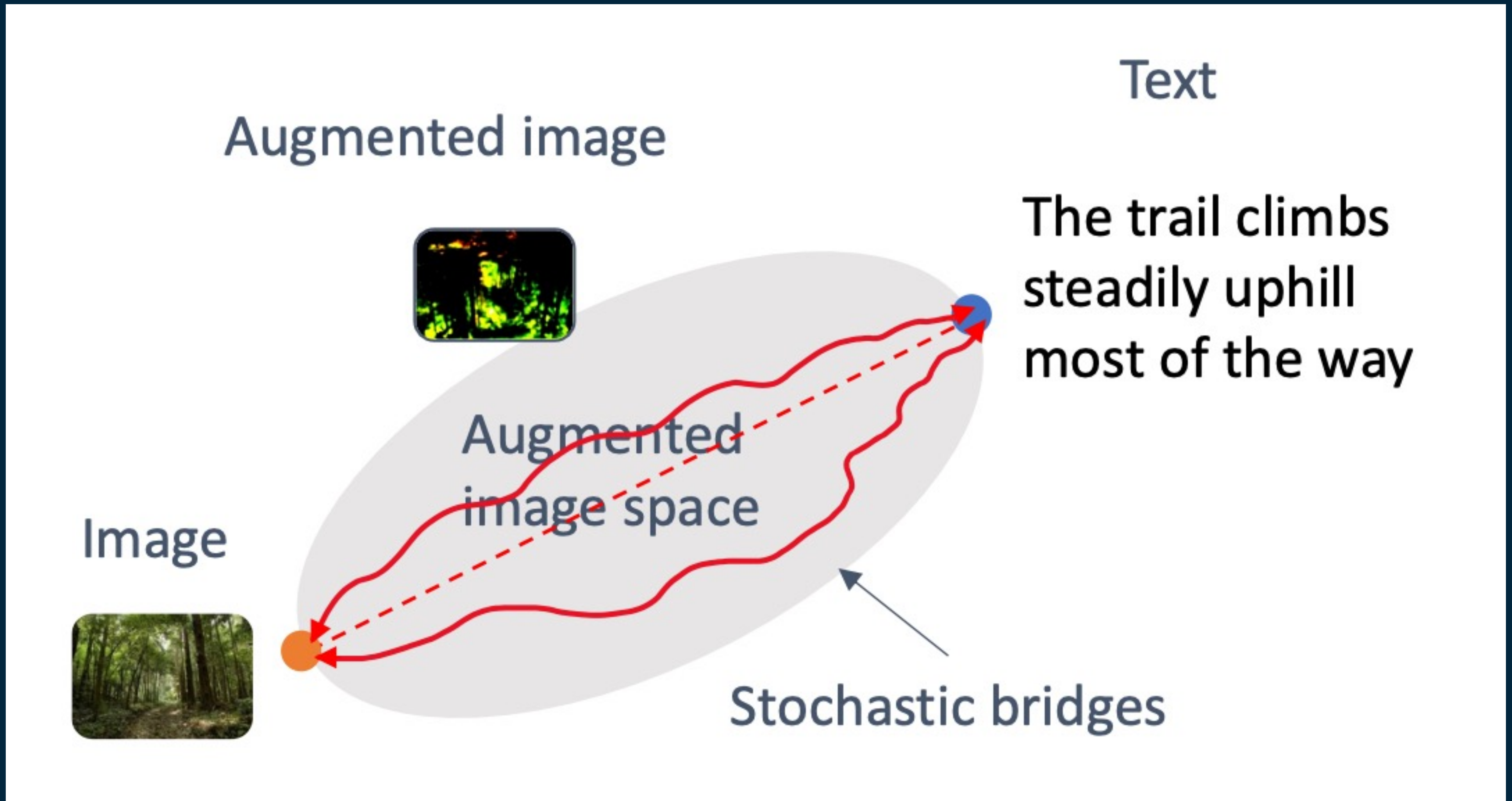
Inter-modality regularization via Brownian bridge

Inter-modality regularization via Brownian Bridge

- To optimize, we simply align Z_V^a with the mean of the Brownian bridge $\mu(\mathrm{Z_V}, \mathrm{Z_T}, \mathrm{t})$:

- $\mathcal{L}_{Br} = \frac{1}{N}\sum_{j=1}^{N}\|Z_V^a - \mu(Z_V, Z_T, t)\|^2$

$$= \frac{1}{N}\sum_{j=1}^{N}\frac{t\langle z_{V\,j}, z_{V\,j}^a\rangle + (1-t)\langle z_{V\,j}^a, z_{T\,j}\rangle}{t^2 + (1-t)^2 + 2t(1-t)\langle z_{V\,j}, z_{T\,j}\rangle}$$



Inter-modality regularization via Brownian bridge

Inter-intra modality regularization via geometric consistency

- Is there a way to combine both inter-modality and intra-modality regularization?
- Consider the distances between inter-modality feature pairs and intra-modality feature pairs
- To construct more meaningful latent structure:
  - ✓ Encourage the geometry symmetry of the feature pair distances

Inter-intra modality regularization via geometric consistency

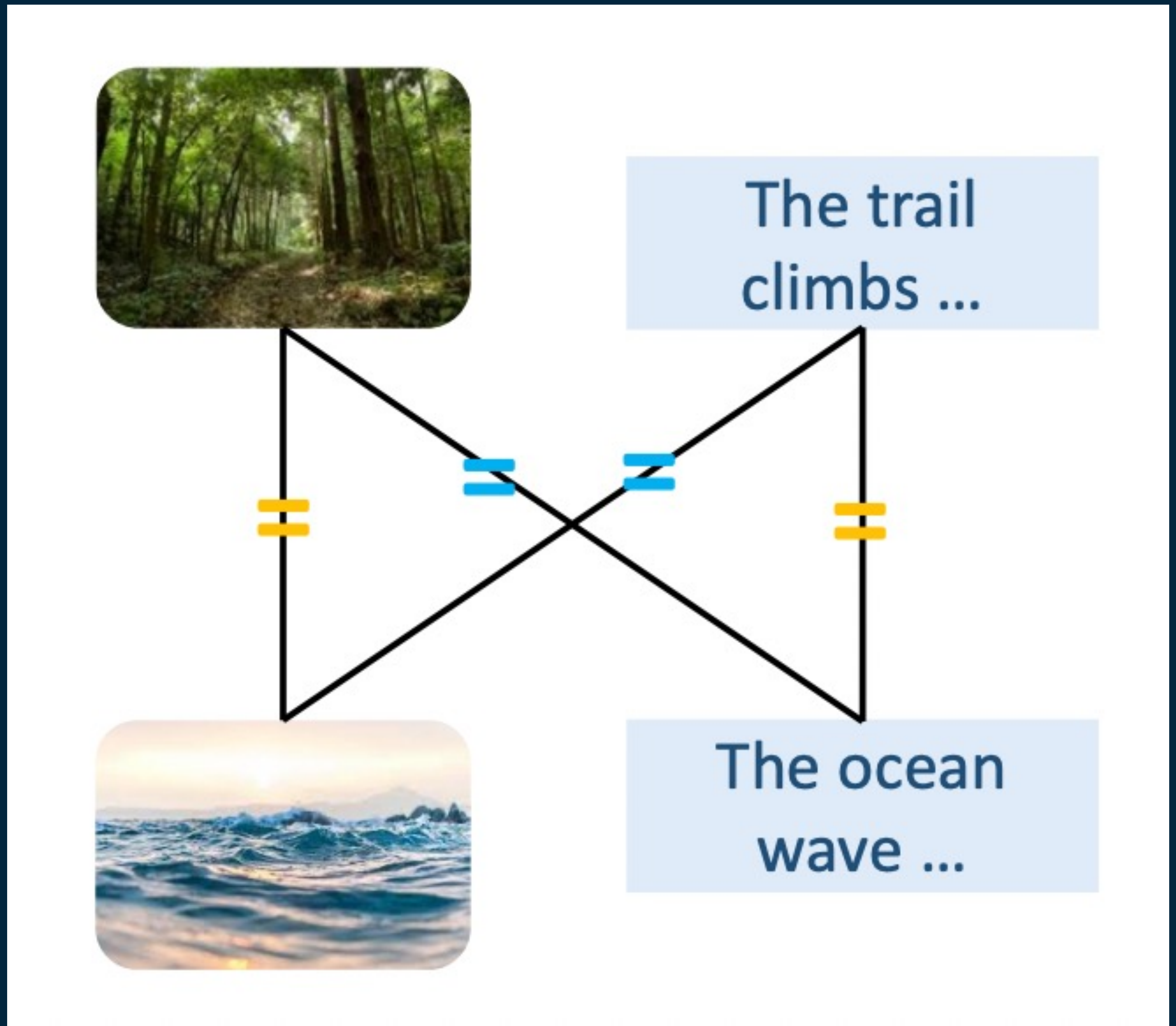
- Enforce geometric consistency on the original features

- Inter-modality consistency:
- $\langle z_{V_1}, z_{T_2} \rangle \sim \langle z_{V_2}, z_{T_1} \rangle$

 The ocean wave …  The trail climbs …

- Intra-modality consistency:
- $\langle z_{V_1}, z_{V_2} \rangle \sim \langle z_{T_1}, z_{T_2} \rangle$

  The trail climbs … The ocean wave …



Inter-intra-modality regularization via geometric consistency

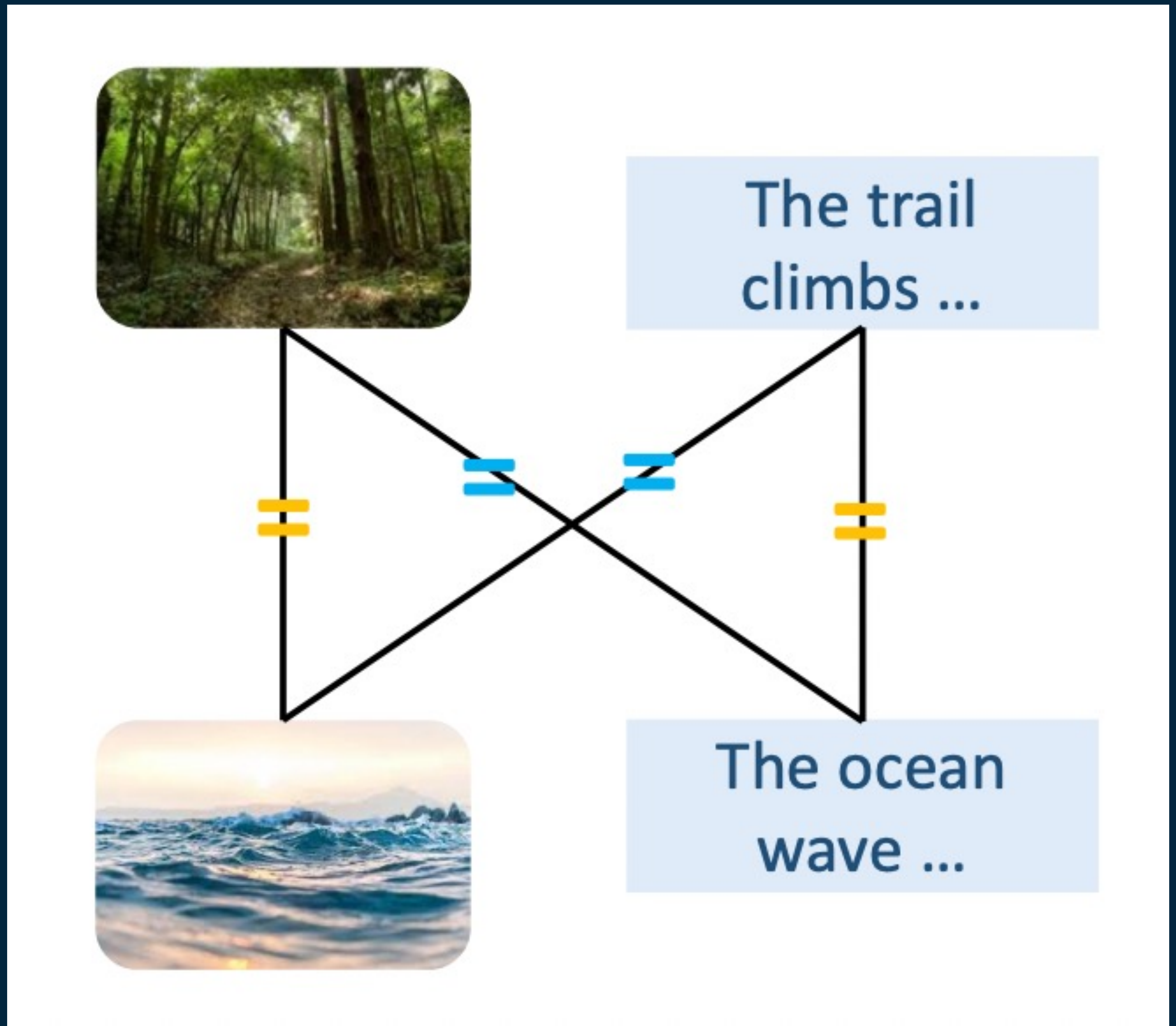
Inter-intra modality regularization via geometric consistency

- To optimize the original features:

- $\mathcal{L}_{GC} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}[$

$$\left(\left\langle z_{V_j}, z_{T_k}\right\rangle - \left\langle z_{V_k}, z_{T_j}\right\rangle\right)^2$$

$$+\left(\left\langle z_{V_j}, z_k\right\rangle - \left\langle z_{T_j}, z_{T_k}\right\rangle\right)^2]$$



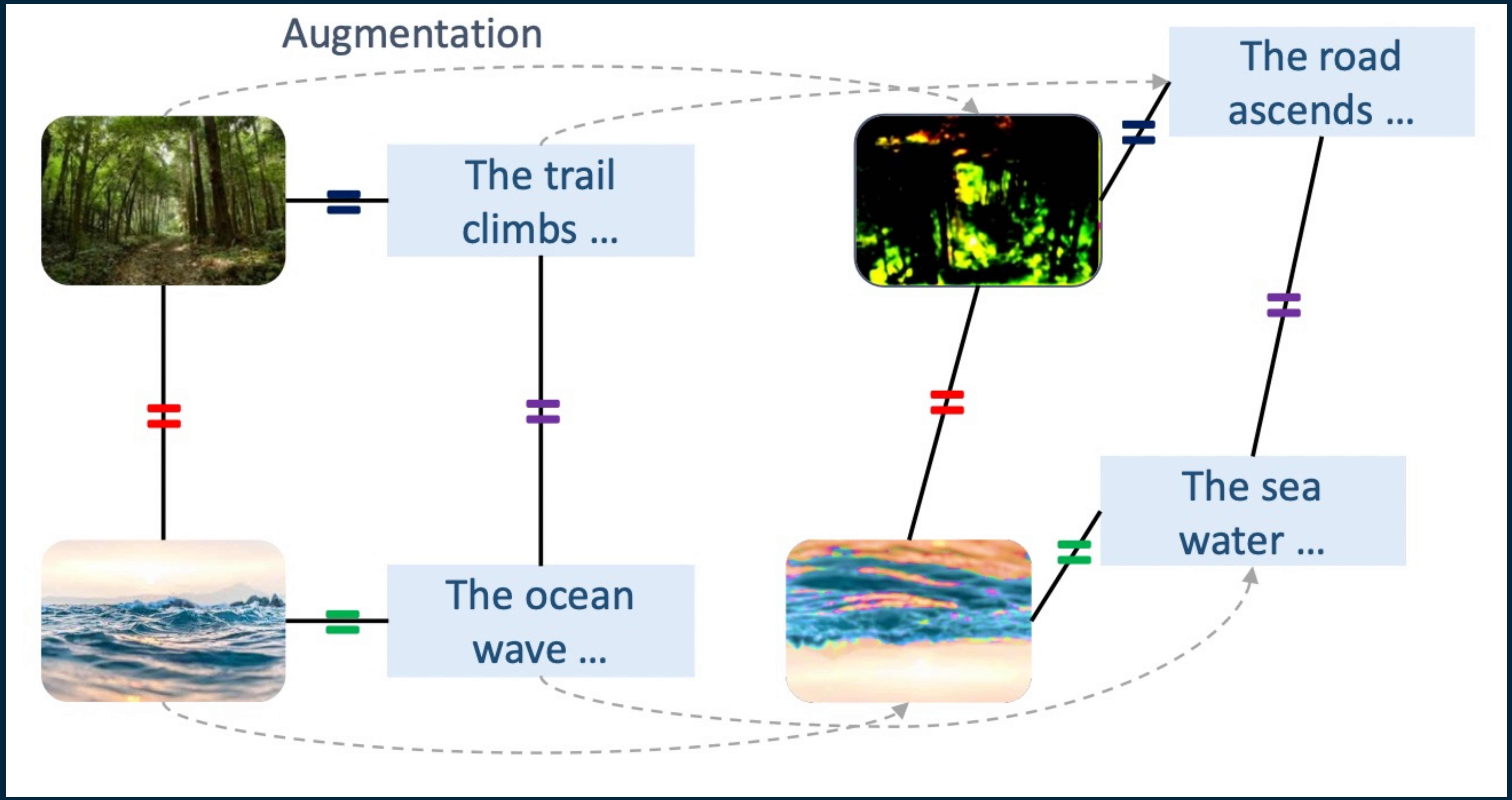
Inter-intra-modality regularization via geometric consistency

Inter-intra-modality regularization

- Enforce geometric consistency on the augmented features

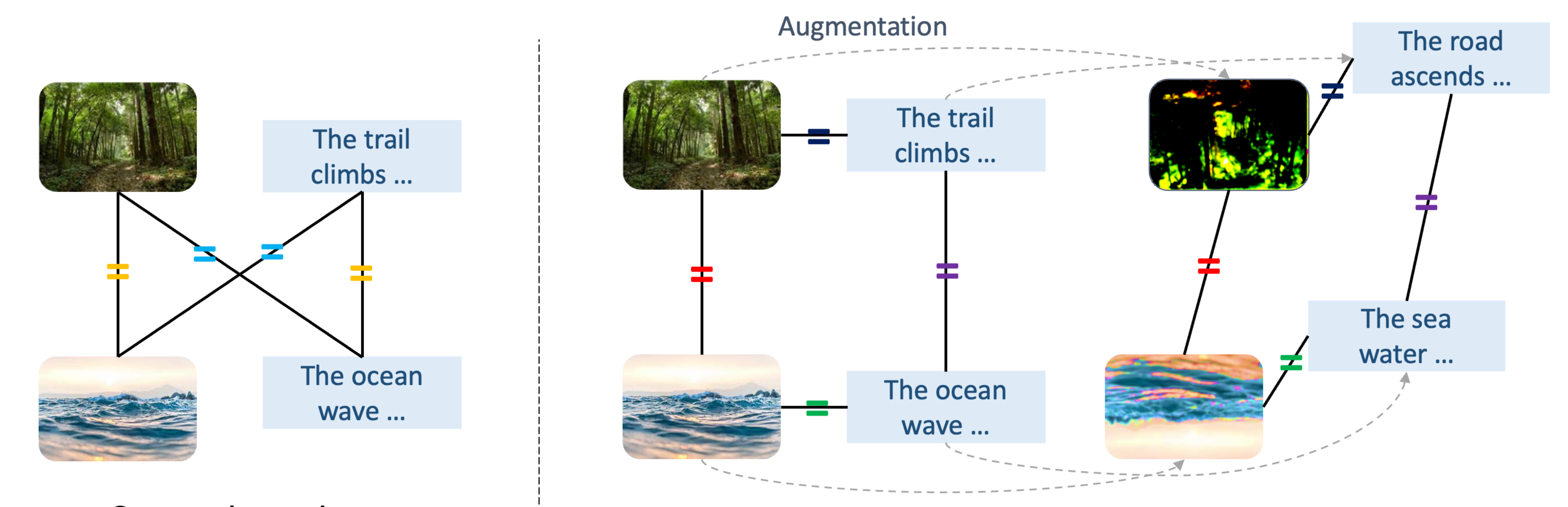
- $$\mathcal{L}_{GC}^{a} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\Big[\left(\left\langle z_{V_j}, z_{V_k}\right\rangle - \left\langle z_{\boldsymbol{V}_j}^{\boldsymbol{a}}, z_{\boldsymbol{V}_k}^{\boldsymbol{a}}\right\rangle\right)^2 + \left(\left\langle z_{T_j}, z_{T_k}\right\rangle - \left\langle z_{T_j}^{a}, z_{T_k}^{a}\right\rangle\right)^2 \Big]$$

$$+ \frac{1}{N}\sum_{j=1}^{N}\left(\left\langle z_{T_j}, z_{T_j}\right\rangle - \left\langle z_{V_j}^{a}, z_{T_j}^{a}\right\rangle\right)^2$$



Inter-intra-modality regularization via geometric consistency

Inter-intra modality regularization via geometric consistency

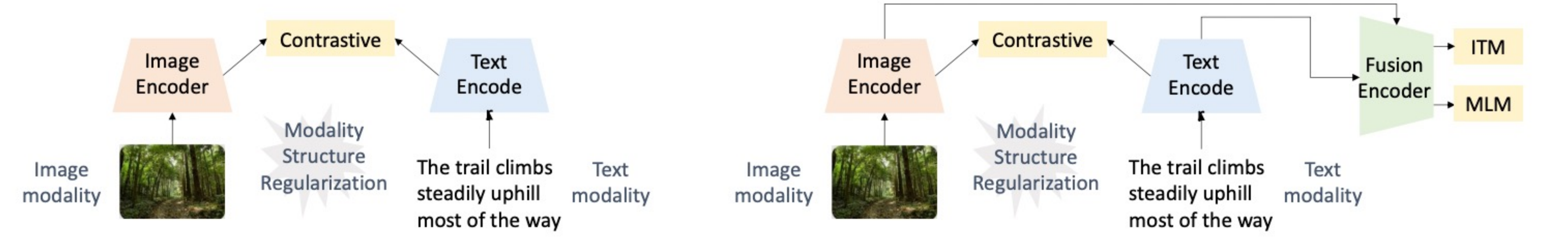


Geometric consistency on image-text features

Geometric consistency on augmented features

Experiments

- Our methods are general regularizations that can be applied for many multi-modal frameworks
- We evaluate our method on two popular vision-language pre-training frameworks
- Two-tower-based models (e.g. CLIP)
- Fusion-based models (e.g. ALBEF)



(a) Two-tower-based models. (b) Fusion-based models.

Experiments setup

- **For two-tower-based models:**
  - text-specific encoder : BERT
  - image-specific encoder : Resnet50
  - text augmentation: EDA
  - image augmentation: random augmentations
  - pre-training data: CC3M

- **For fusion-based models:**
  - text-specific encoder : BERT
  - image-specific encoder : ViT
  - fusion-encoder: BERT
  - text augmentation: momentum model
  - image augmentation: random augmentations + momentum model
  - pre-training data: CC3M, VG, SBU, COCO

Experiments on two-tower-based models

- Zero-shot transfer

Table 1. Zero-shot TopK classification accuracy (%) on CIFAR10, CIFAR100 and ImageNet1K.

Method	CIFAR10			CIFAR100			ImageNet1K		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
CLIP [48]	44.95	72.58	88.3	15.05	29.51	37.53	16.72	28.61	34.38
CyCLIP [18]	43.22	71.43	83.22	15.09	27.39	34.35	17.77	30.06	36.20
OURS_{Sep}	46.61	**81.21**	**92.44**	19.37	36.66	46.26	20.21	33.25	39.60
OURS_{Br}	43.15	72.77	86.72	14.22	26.46	33.28	**20.45**	**33.56**	39.28
OURS_{GC}	**56.36**	80.47	90.27	**22.70**	**41.66**	**51.78**	20.25	33.50	**39.91**

Experiments on two-tower-based models

- Natural distribution shifts
- Distribution shifted benchmarks of ImageNet1K
- Standard benchmark to evaluate the robustness of models

ImageNet1K		
Top1	Top3	Top5
16.72	28.61	34.38
17.77	30.06	36.20
20.21	33.25	39.60
20.45	**33.56**	39.28
20.25	33.50	**39.91**

Table 2. Zero-shot TopK classification accuracy (%) on Natural Distribution Shifts.

Method	ImageNetV2			ImageNetSketch			ImageNet-A			ImageNet-R		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
CLIP [48]	14.11	25.76	31.80	8.61	16.47	21.13	2.81	7.31	11.32	19.07	31.99	39.03
CyCLIP [18]	15.25	26.59	32.15	8.30	16.18	20.77	3.27	8.45	13.07	19.85	33.35	40.35
OURS_{Sep}	16.78	28.97	35.68	9.22	17.86	23.00	3.45	9.88	15.81	22.06	35.65	43.01
OURS_{Br}	17.02	29.39	35.53	10.34	18.39	23.05	3.01	7.50	11.45	20.40	32.43	38.45
OURS_{GC}	**17.37**	**29.84**	**36.65**	**10.90**	**20.77**	**26.11**	**3.87**	**11.36**	**16.76**	**23.85**	**37.90**	**45.03**

Experiments on two-tower-based models

- Linear Probing
- Fit a linear classifier on learned models

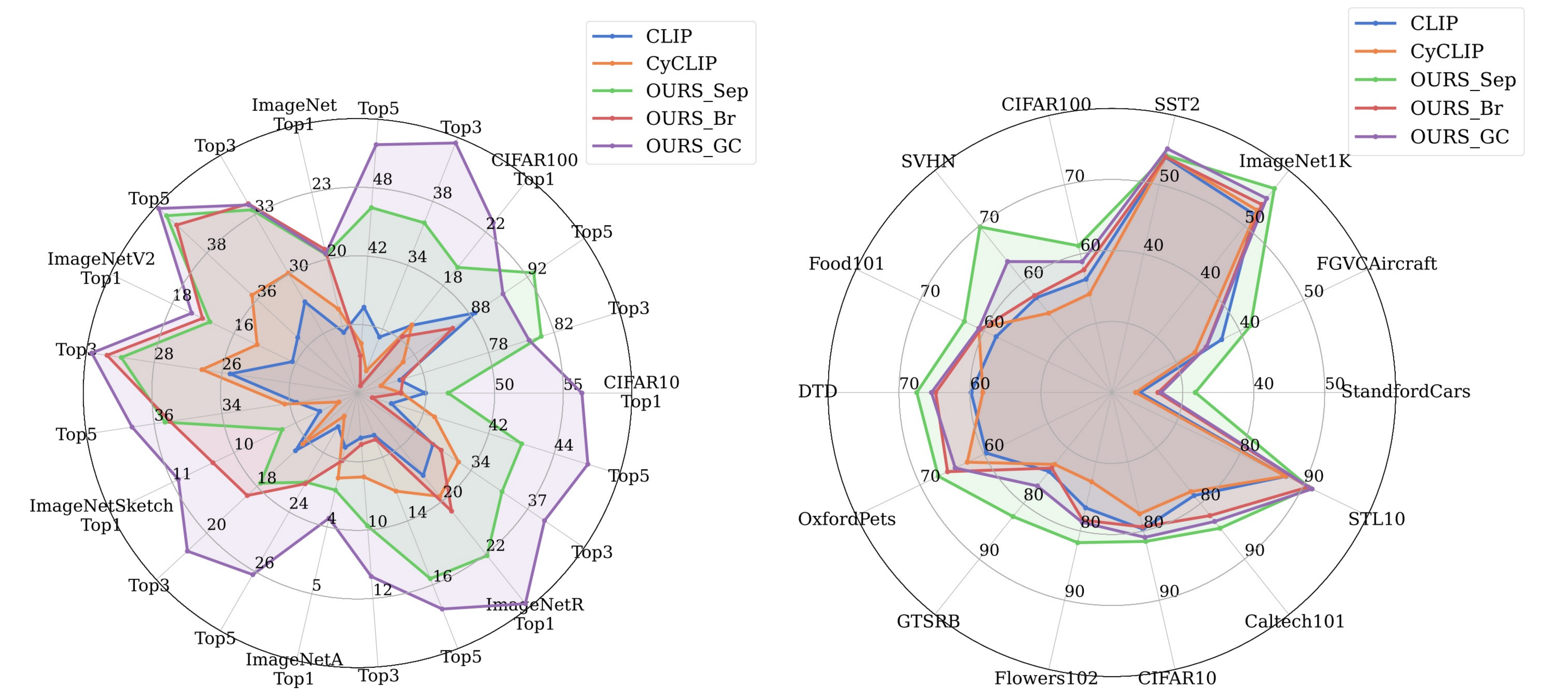
Table 3. Linear probing Top1 classification accuracy (%) on visual benchmarks.

	Caltech101	SVHN	STL10	CIFAR10	CIFAR100	DTD	FGVCAircraft	OxfordPets	SST2	Food101	GTSRB	StanfordCars	Flowers102	ImageNet1K	Average
CLIP [48]	78.57	57.07	87.22	79.74	56.36	59.84	37.17	59.66	53.98	58.11	74.21	23.96	76.66	52.10	61.05
CyCLIP [18]	77.86	54.29	87.61	77.53	54.23	58.19	33.00	62.63	54.81	60.82	72.95	23.36	72.89	52.83	60.14
OURS_{Sep}	**84.45**	**69.82**	90.96	**81.51**	**61.19**	**67.50**	**41.70**	**67.16**	54.26	**63.08**	**82.35**	**31.76**	**81.69**	**56.73**	**66.73**
OURS_{Br}	82.18	57.46	90.69	79.42	57.72	64.84	34.74	65.71	54.04	60.52	73.61	26.50	78.44	53.87	62.84
OURS_{GC}	83.23	63.58	**91.31**	80.92	58.89	65.43	34.83	64.51	**55.19**	60.80	76.84	26.95	78.76	54.96	64.01

Visualization of results on zero-shot transfer and natural distribution shift

Visualization of results on linear probing

Visualization of results



(a) Zero-shot Performance. (b) Linear Probing Performance.

Latent feature structure visualization

Experiments on fusion-based models

- Fusion-based models are more powerful to learn the cross-modality interactions

- We evaluate on vision-language tasks:

  ✓ Visual Question Answering (VQA)
  ✓ Natural Language for Visual Reasoning (NVLR^2)
  ✓ Visual Entailment (VE)

Experiments on fusion-based models

Table 4. Downstream tasks performance on fusion-based models.

Method	VQA		NLVR^2		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
ImageBERT [32]	70.80	71.00	67.40	67.00	-	-
LXMERT [56]	72.42	72.54	74.90	74.50	-	-
12-in-1 [37]	73.15	-	- 78.87	-	76.95	
UNITER [7]	72.70	72.91	77.81	77.85	78.59	78.28
OSCAR [33]	73.16	73.44	78.07	78.36	-	-
VILLA [16]	73.59	73.67	78.39	79.30	79.47	79.03
ViLT [26]	70.94	-	75.24	76.21	-	-
ViCHA [53]	73.55	-	78.14	77.00	79.20	78.65
ALBEF [31]	73.38	73.52	78.36	79.54	79.69	79.91
CODIS [13]	73.15	73.29	78.58	**79.92**	79.45	80.13
OURS_{All}	74.12	74.16	**80.18**	79.80	79.62	**80.23**
OURS_{Sep}	73.52	73.59	79.05	79.76	**79.95**	79.61
OURS_{Br}	**74.26**	**74.36**	78.70	79.36	79.86	79.95
OURS_{GC}	73.90	73.87	78.96	79.53	79.82	80.16

Summary

- We study the impact of modality alignment with empirical and theoretic analysis

- We propose three regularizations to construct latent feature structures
  - intra-modality regularization via deep feature separation
  - inter-modality regularization via Brownian bridge
  - intra-inter-modality regularization via geometric consistency

- We demonstrate improved performance on both two-tower-based models and fusion-based models on a variety of tasks

Thank you

qianjian@amazon.com

Appendix

- Let X_0 and X_1 denote the inputs from two modalities and Y denote the task label. A quantitative measure of “usefulness” of a modality could be defined as:

$$S(X_i) := \inf_{h:x \to y} E[l_{CE}(h(x_i), Y)]$$

- From information theory:

$$S(X_i) = H(Y \mid X_i)$$

- Hence we could use $I(X_i; Y) = H(Y) - H(Y \mid X_i)$ as a measure of the utility of one modality.

Appendix

Theorem 3.1. For a pair of modality encoders $g_T(\cdot)$ and $g_V(\cdot)$, if the multi-modal features $Z_T = g_T(X_T)$ and $Z_V = g_V(X_V)$ are perfectly aligned in the feature space, i.e., $Z_T = Z_V$, then $\inf_h \mathbb{E}_p[\ell_{\mathrm{CE}}(h(Z_T, Z_V), Y)] - \inf_{h'} \mathbb{E}_p[\ell_{\mathrm{CE}}(h'(X_T, X_V), Y)] \geq \Delta_p$.

Proof of Theorem 3.1. Consider the joint mutual information $I(Z_T, Z_V; Y)$. By the chain rule, we have the following decompositions:

$$\begin{aligned} I(Z_T, Z_V; Y) &= I(Z_T; Y) + I(Z_V; Y \mid Z_T) \\ &= I(Z_V; Y) + I(Z_T; Y \mid Z_V). \end{aligned}$$

However, since Z_T and Z_V are perfectly aligned, $I(Z_V; Y \mid Z_T) = I(Z_T; Y \mid Z_V) = 0$, which means $I(Z_T, Z_V; Y) = I(Z_V; Y) = I(Z_T; Y)$. On the other hand, by the celebrated data-processing inequality, we know that

$$I(Z_T; Y) \leq I(X_T; Y), \quad I(Z_V; Y) \leq I(X_V; Y).$$

Hence, the following chain of inequalities holds:

$$\begin{aligned} I(Z_T, Z_V; Y) &= \min\{I(Z_T; Y), I(Z_V; Y)\} \\ &\leq \min\{I(X_T; Y), I(X_V; Y)\} \\ &\leq \max\{I(X_T; Y), I(X_V; Y)\} \\ &\leq I(X_T, X_V; Y), \end{aligned}$$

where the last inequality follows from the fact that the joint mutual information $I(X_T, X_V; Y)$ is at least as large as any one of $I(X_T; Y)$ and $I(X_V; Y)$. Therefore, due to the variational form of the conditional entropy, we have

$$\begin{aligned} &\inf_h \mathbb{E}_p[\ell_{\mathrm{CE}}(h(Z_T, Z_V), Y)] - \inf_{h'} \mathbb{E}_p[\ell_{\mathrm{CE}}(h'(X_T, X_V), Y)] \\ &= H(Y \mid Z_T, Z_V) - H(Y \mid X_T, X_V) \\ &= I(X_T, X_V; Y) - I(Z_T, Z_V; Y) \\ &\geq \max\{I(X_T; Y), I(X_V; Y)\} - \min\{I(X_T; Y), I(X_V; Y)\} \\ &= \Delta_p. \end{aligned}$$

■