BATCLIP: Bimodal Online Test-Time Adaptation for CLIP

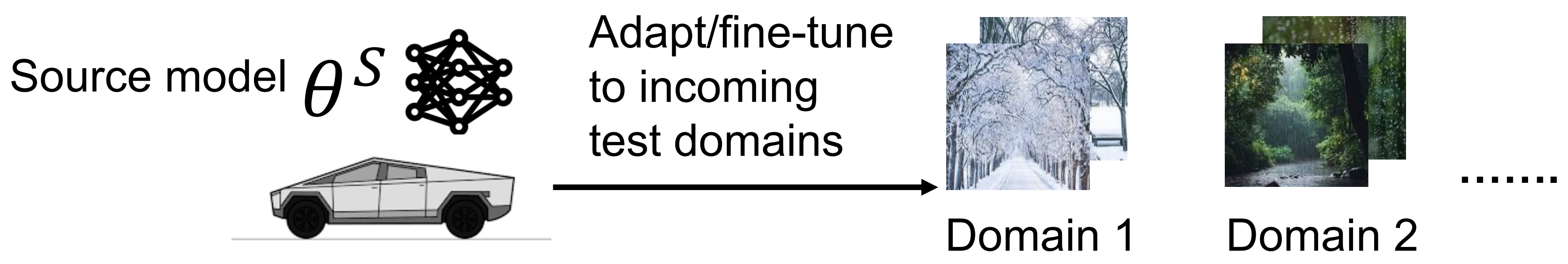
Sarthak Kumar Maharana[1], Baoming Zhang[1], Leonid Karlinsky[2], Rogério Schmidt Feris[2], Yunhui Guo[1]

UT Dallas[1], MIT-IBM Watson AI Lab[2]

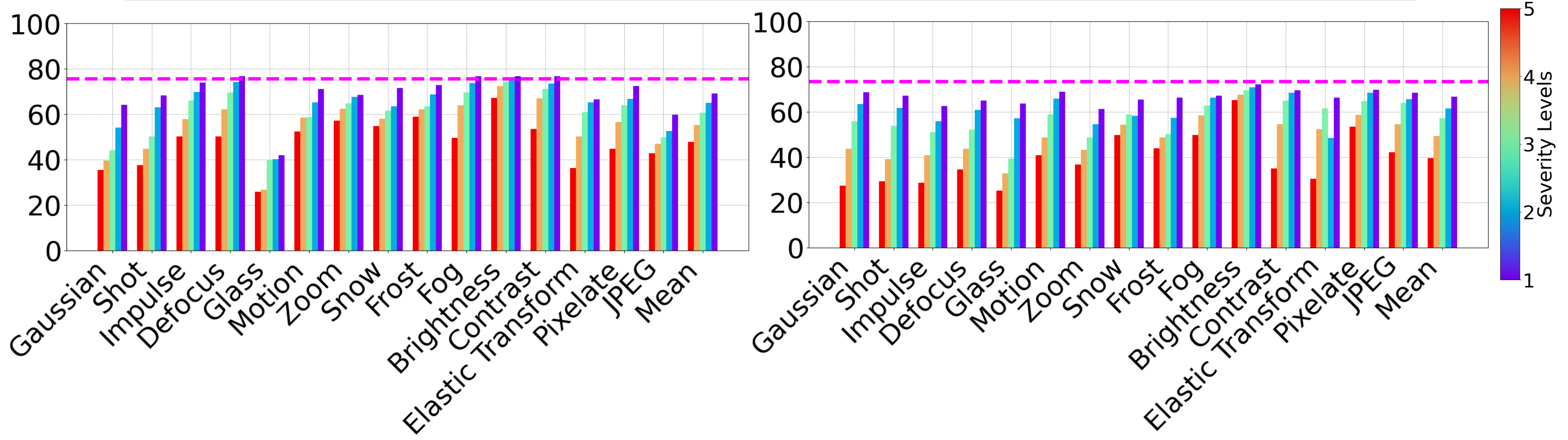
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Problem Overview and Motivation

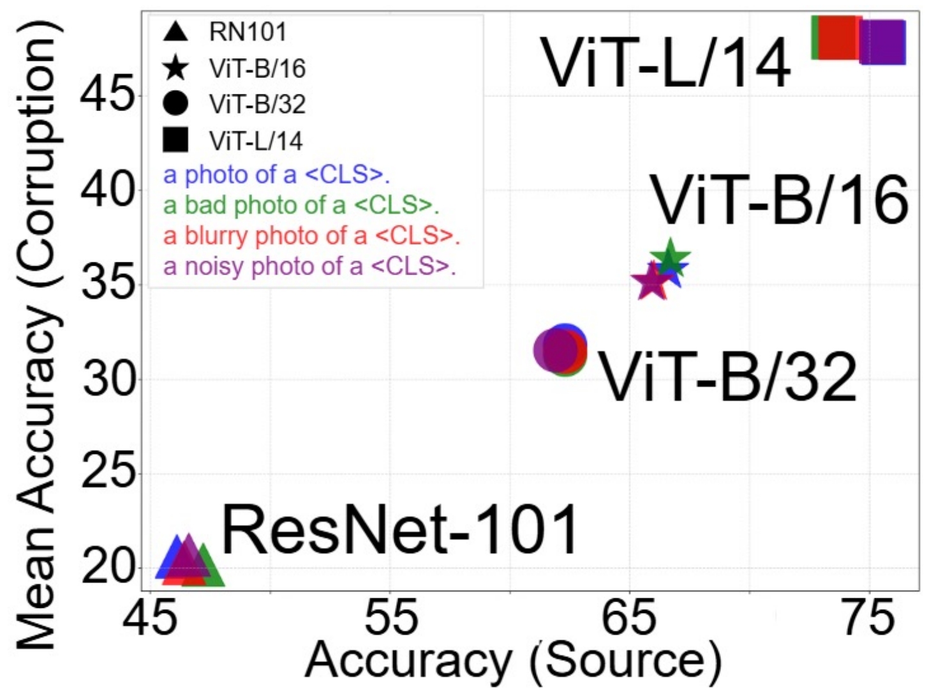
- **Online test-time adaptation (TTA)** involves pre-trained/source model adaptation to incoming **unlabeled test data** to minimize the source-target domain distribution gap.
  - *Single forward pass* to preserve privacy.
  - No access to the pre-training/source dataset.



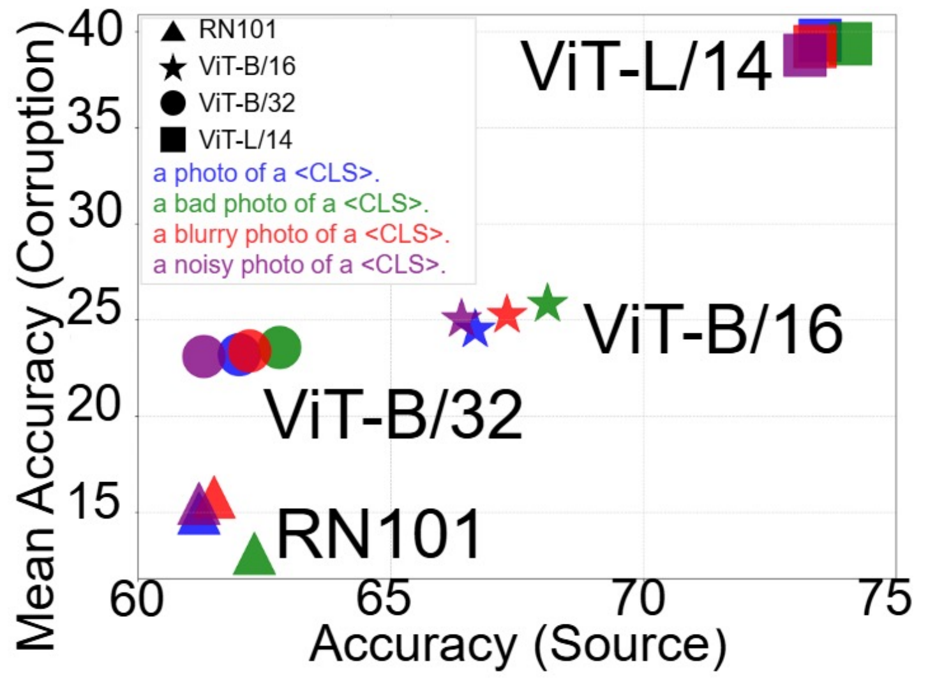
Are zero-shot CLIP features transferable to "new" domain shifts/corruptions? NO!



Across backbones, CLIP is very sensitive to severe visual corruptions (up, ViT-L/14); at test-time, "relevant" prompting doesn't help (bottom). So, there's a need for minimal adaptation as text and visual features are independent.



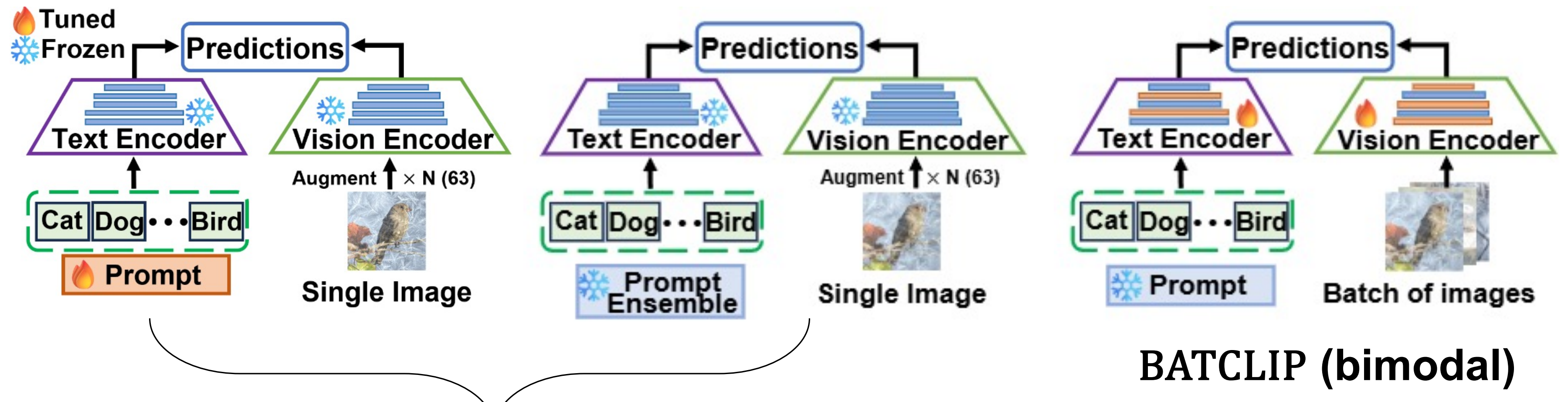
CIFAR-100C



ImageNet-C

Proposed Methodology

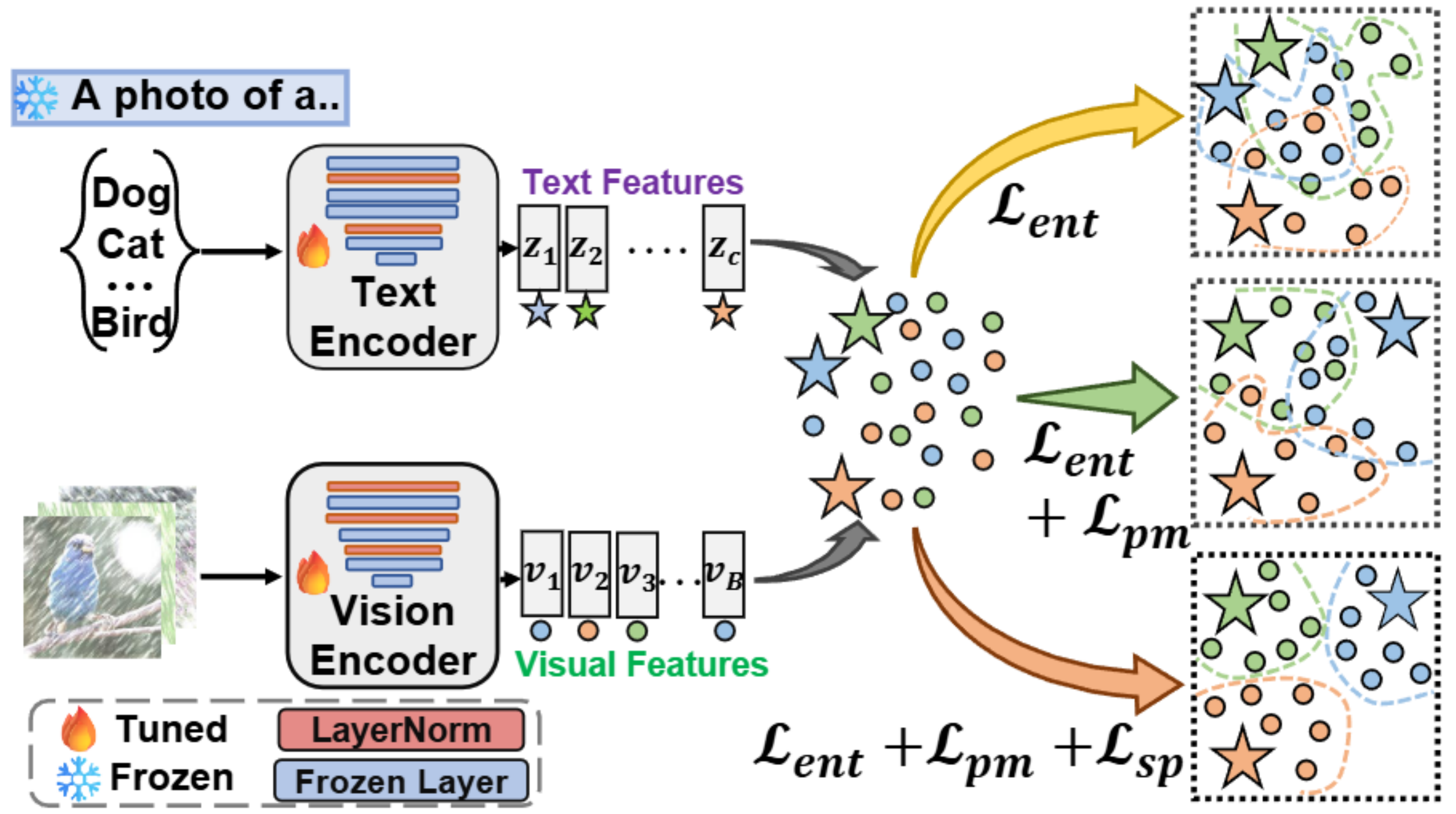
- Existing online TTA approaches using CLIP vs ours.



Prior unimodal CLIP TTA approaches

Unimodal TTA approaches obtain suboptimal alignment between the modality features post adaptation.

- Ours for **online** CLIP (or any contrastively pre-trained VLM) TTA



$$\mathcal{L}_{ent} = -\sum_{c} p(l_c) \log p(l_c)$$

Entropy minimization

$$\mathcal{L}_{pm} = \frac{1}{C} \sum_{c} \bar{v}_c \cdot \hat{z}_c$$

Maximize the projection of a class prototype onto normalized text feature.

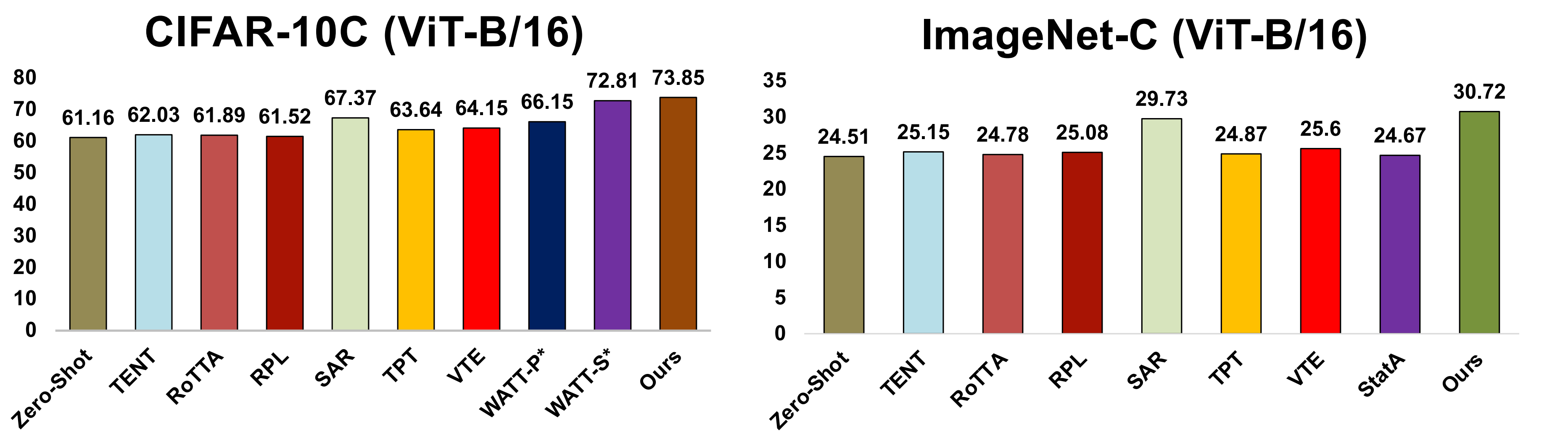
$$\mathcal{L}_{sp} = \sum_{l \in C} \sum_{c \in C} \mathbb{1}[l \neq c](1 - \cos(\bar{v}_c, \bar{v}_l))$$

Maximize inter-class separability of prototypes to obtain good on-the-fly decision boundaries and generalization.

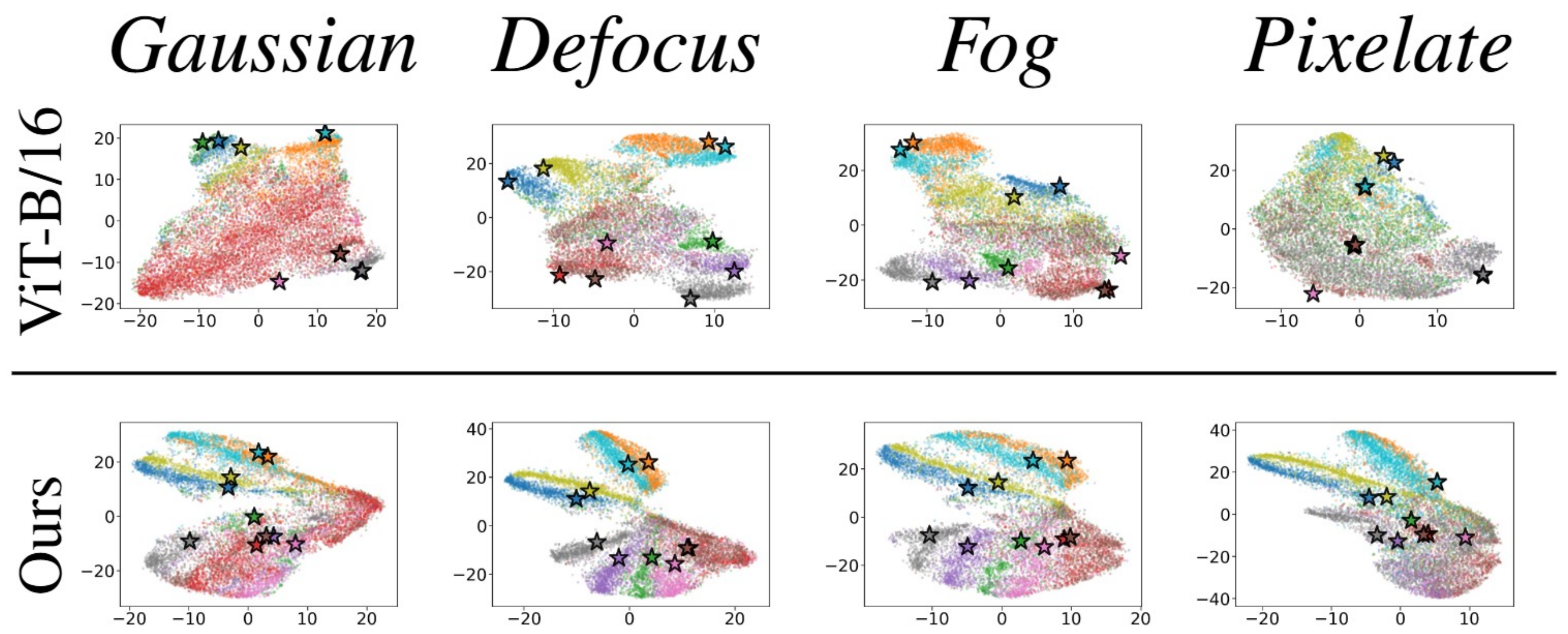
Overall, BATCLIP maximizes vision-text alignment and inter-class separability for online TTA under severe domain shifts.

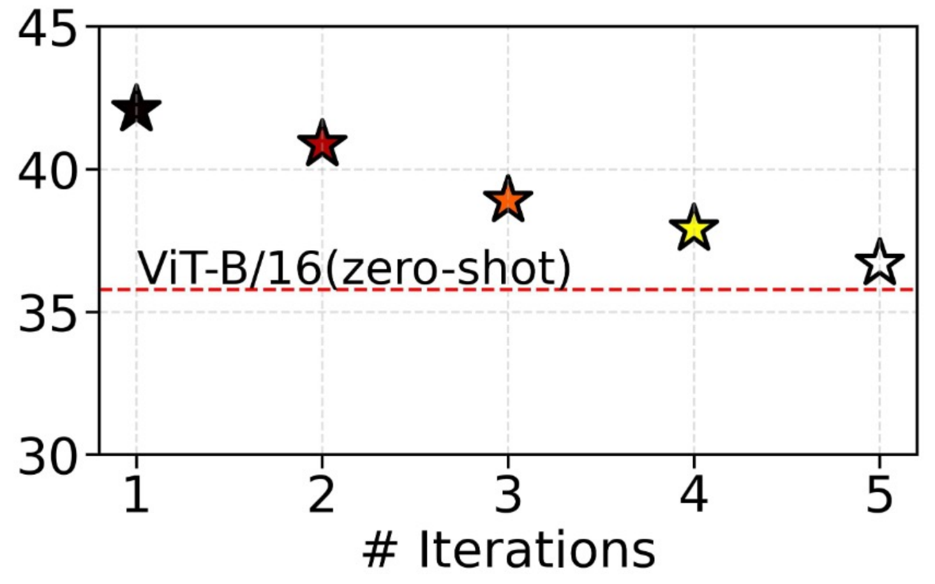
Results and Discussions

- **Mean accuracy** (%) across 15 domains/tasks of image corruptions.

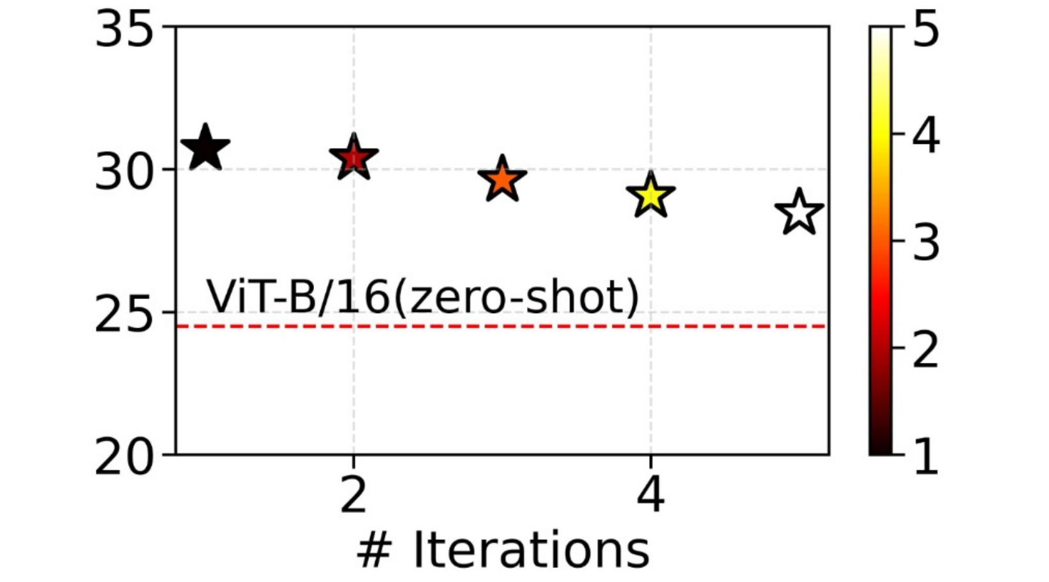


BATCLIP yields more discriminative visual features that exhibit stronger alignment with their corresponding text features – with just one adaptation step.





CIFAR-100C



ImageNet-C

Adaptation for multiple iterations on a single test batch >> zero-shot CLIP.

For more results on complex domains including lighting conditions, camera types →

*Leonid Karlinsky is with Xero.