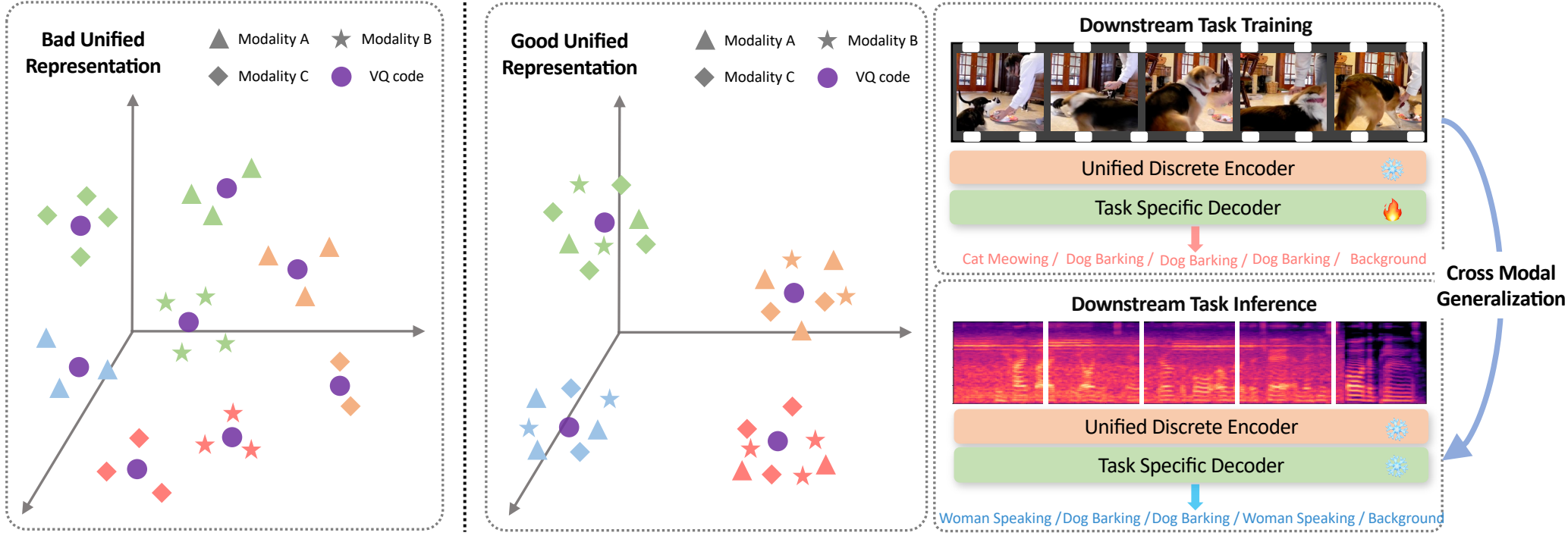
Bad Unified Representation
Modality A
Modality B
Modality C
VQ code
Good Unified Representation
Modality A
Modality B
Modality C
VQ code
Downstream Task Training
Unified Discrete Encoder
Task Specific Decoder
Cat Meowing / Dog Barking / Dog Barking / Dog Barking / Background
Cross Modal Generalization
Downstream Task Inference
Unified Discrete Encoder
Task Specific Decoder
Woman Speaking / Dog Barking / Dog Barking / Woman Speaking / Background