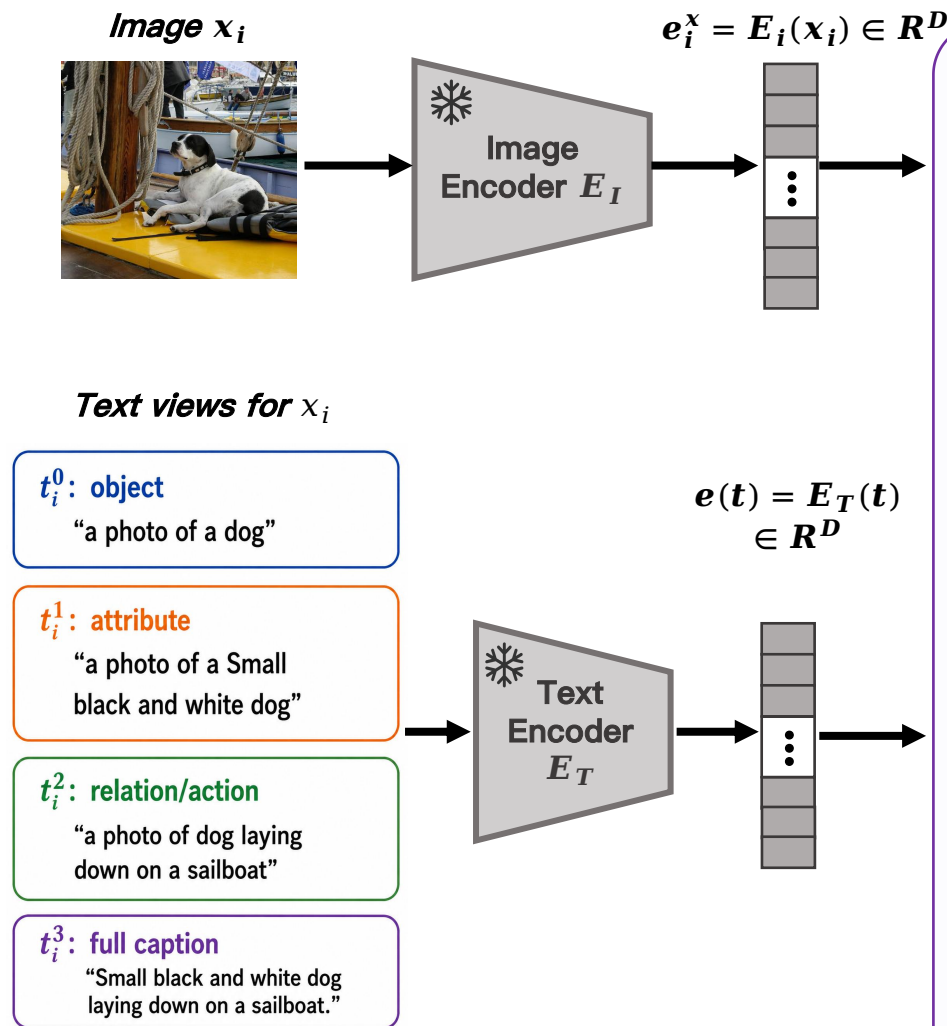
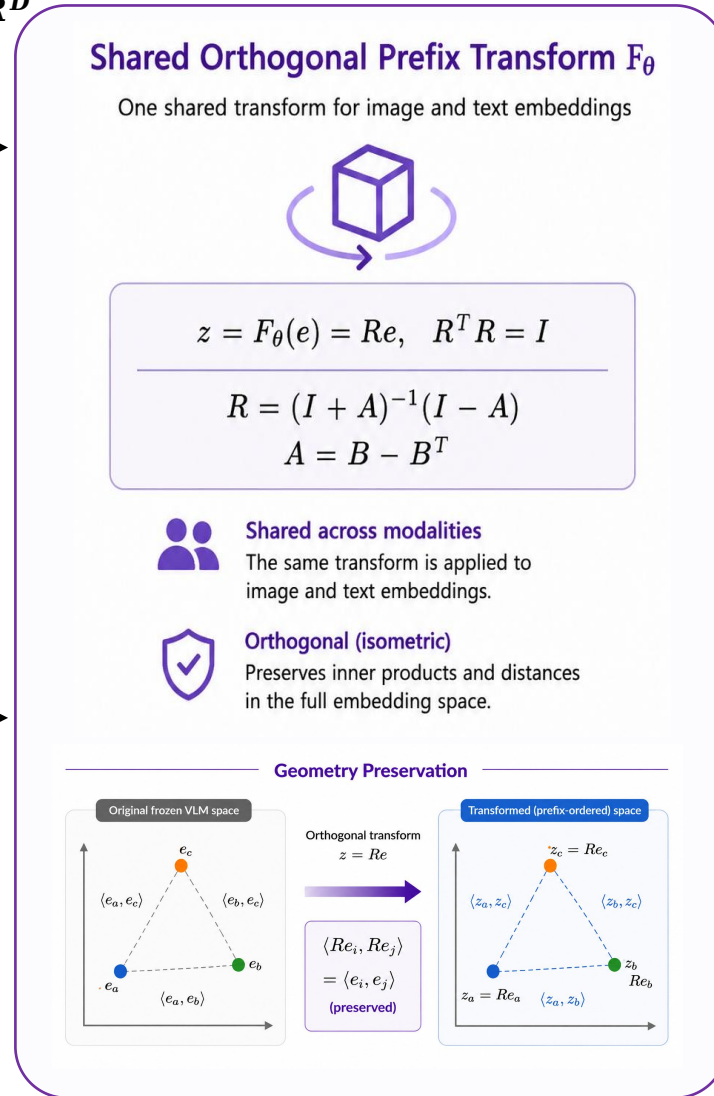


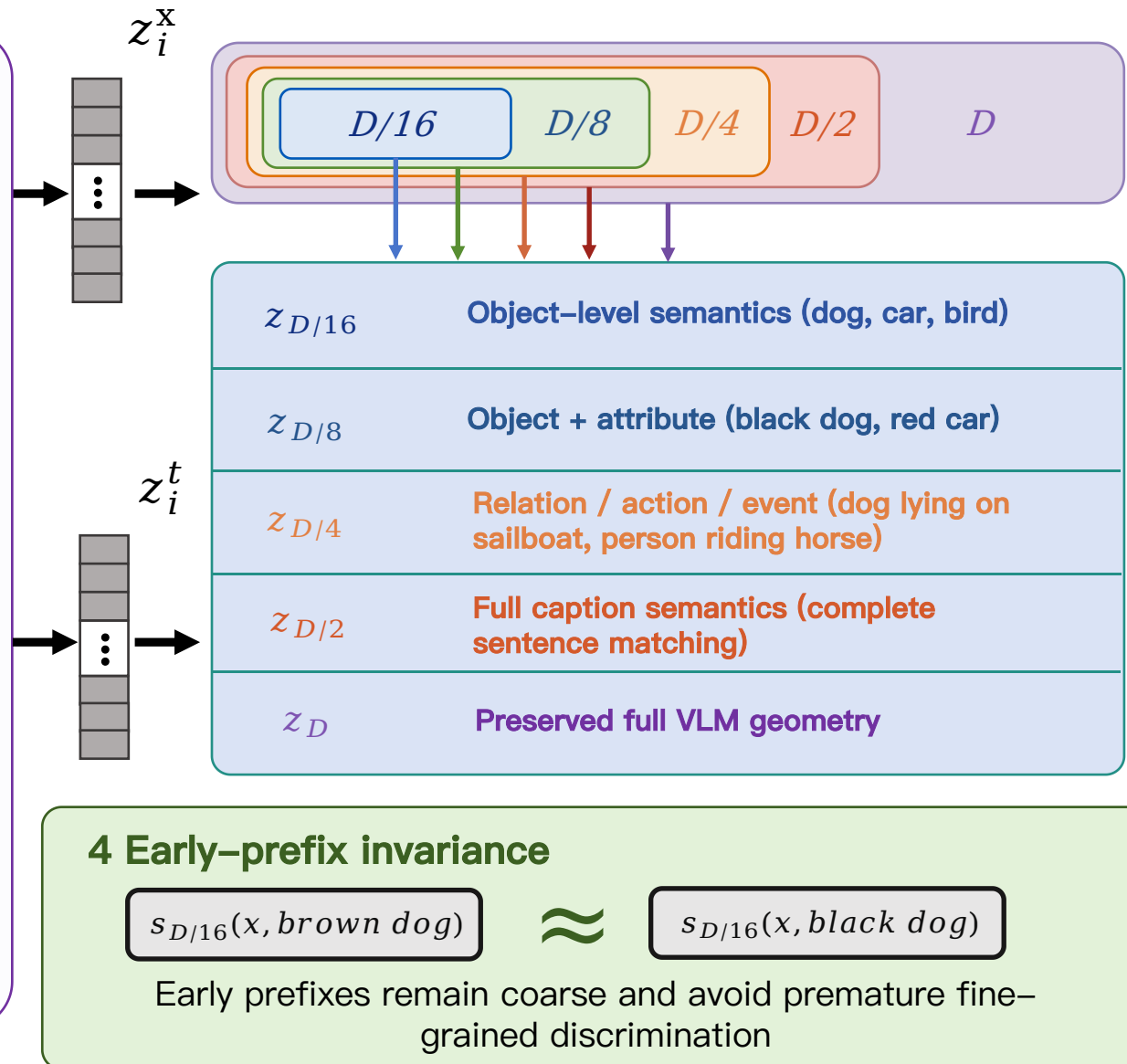
① Frozen VLM inputs and Encoders



② Shared Orthogonal Prefix Transform



③ Semantic Matryoshka Prefix Ladder



④ Training Objectives

1 Multi-Granularity Alignment

$z_{D/16} \leftrightarrow t^0$
dog

$z_{D/8} \leftrightarrow t^1$
small black and white dog

$z_{D/4} \leftrightarrow t^2$
dog laying down on a sailboat

$z_{D/16} \leftrightarrow t^0$
small black and white dog laying down on a sailboat

2 Cumulative Retention

$D/16$ [Blue bar]

$D/8$ [Blue bar, Green bar]

$D/4$ [Blue bar, Green bar, Orange bar]

$D/2$ [Blue bar, Green bar, Orange bar, Red bar]

Longer prefixes contain previous coarse semantics.

3 selective hard-negative ranking

Positive (match) ✓

Small black and white dog laying down on a sailboat

Negative examples (marked with X):

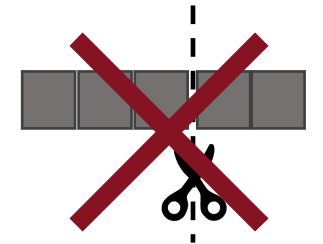
- object negative:** Small black and white cat laying down on a sailboat.
- attribute negative:** Small brown and white dog laying down on a sailboat.
- relation negative:** Small brown and white dog laying down under a sailboat.
- action negative:** Small black and white dog standing on a sailboat.
- order negative:** A sailboat laying down on a small black and white dog.
- full negative:** A small boat in the water and a person.

Earliest prefix via $\kappa(r)$

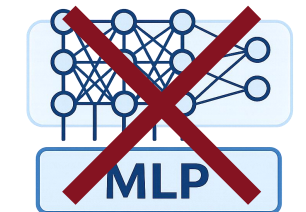
$\kappa(\text{object}) = D/16$
 $\kappa(\text{attribute}) = D/8$
 $\kappa(\text{relation/action}) = D/4$
 $\kappa(\text{order/action}) = D/4$
 $\kappa(\text{full}) = D/2$

Not the following:

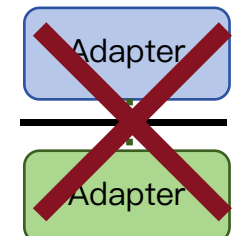
Direct truncation



Generic MLP adapter



Separate image/text adapters



Our goal:
Prefix length = semantic