Bidirectional Decoupling Alignment Module

Identity Feature Stream

FFN → MHSA → Layer Norm

Semantic Cross-Attention

Clothing Feature Stream

V K Q

Q K V

LayerNorm GAP

Layer Norm GAP

AGM

ConCat FFN W_{id} W_{clo}

f_{id}^i

$\min \mathcal{L}_{HSIC}$

f_{clo}^i

$\min \mathcal{L}_{aln}$

Image Encoder

Fusion

Clothes Caption:
"Dress, Hat, Shorts, Sleeve, Pants, Jacket, Coat, Skirt, Sweater, Sleeve, Cap, Scarf, Belt, Collar, Neck, Floral"

Person Caption:
"She wears a purple long sleeved, ankle length dress. There is a pattern on the dress."

Text Encoder → MHSA → Layer Norm → Mean Pool → FFN

f_{clo}^t

f_{id}^t

MHSA Multi-Head Self-Attention **FFN** Feed-Forward Network **AGM** Adaptive Gating Mechanism