Bidirectional Decoupled Alignment Module

Identity Feature Stream

FFN · MHSA · Layer Norm

$\oplus$

V · K · Q

LayerNorm · GAP

AGM

ConCat · FFN · W_{id}

f_{id}^i

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

Clothing Feature Stream

FFN · MHSA · Layer Norm

$\oplus$

Q · K · V

Layer Norm · GAP

W_{clo}

f_{clo}^i

Semantic Cross-Attention

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

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 · $\oplus$ · Layer Norm · Mean Pool · FFN

f_{clo}^t

f_{id}^t

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