Streaming Verification Decoding Inference Framework

Input: $Ctx^{<t_j}$ ... Dec^1 ... Dec^k

$+$ New Frm^{t_j}

Decoding step^{t_i}

Forward: VLM 🚫 Decoder Layer

logits — PPL — Dec^k — PPL — logits

Decoding gate: output perplexity — Compare with α * decode threshold

Greater than ✔ / Less than ✘

Decoder

Update context: Dec^{k+1}

Append Dec^{k+1}

... Dec^k ... Dec^{k+1}

Swap Frm^{t_j} and Dec^k

... Dec^k

Frame-1 ~ Frame-5 Caption-1

Frame-6 ~ Frame-11 Caption-2

Frame-17 ~ Frame-21 Caption-4

Frame-by-Frame Input

VLM 🚫 Decoder Layer

Caption 1

Caption 2

Streaming Verification Decoding

logits logits logits logits logits logits logits logits

Decoder

Cap1 ✔ / Slient ✘ / Slient ✘ / Slient ✘ / Slient ✘ / Cap2 ✔ / Slient ✘ / Slient ✘ / Slient / Slient / Slient ... Cap4 / Slient / Slient / Slient / Slient