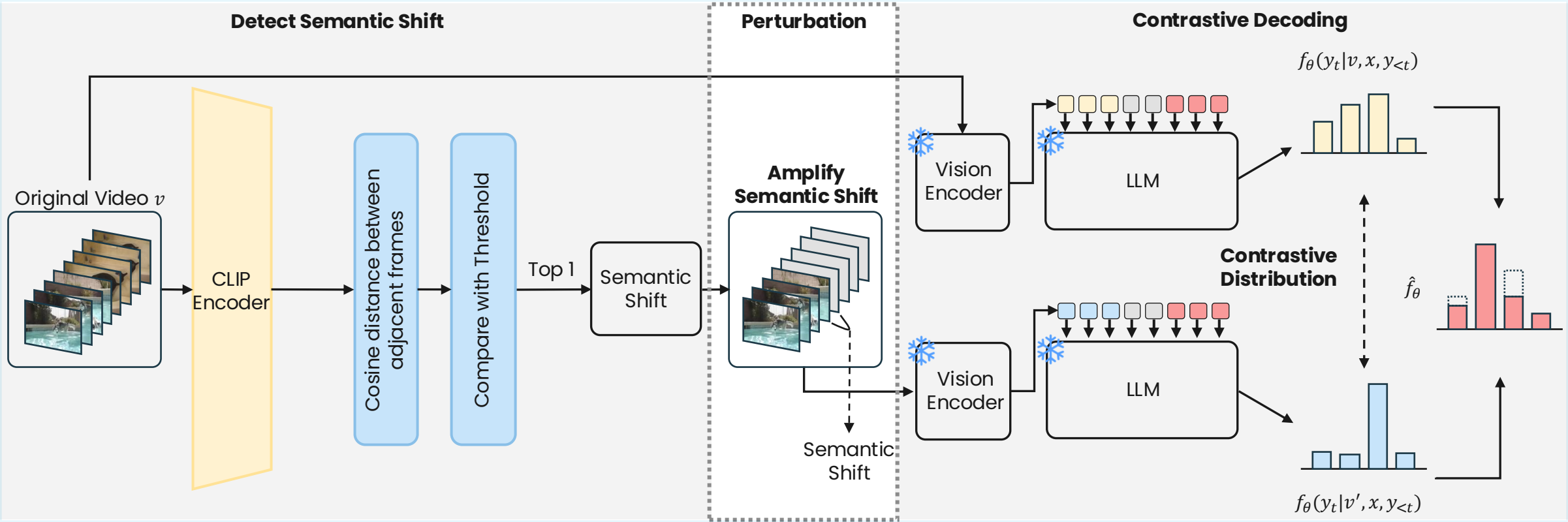
Detect Semantic Shift
Perturbation
Contrastive Decoding
Original Video $v$
CLIP Encoder
Cosine distance between adjacent frames
Compare with Threshold
Top 1
Semantic Shift
Amplify Semantic Shift
Semantic Shift
Vision Encoder
LLM
Vision Encoder
LLM
$f_\theta(y_t|v,x,y_{<t})$
Contrastive Distribution
$\hat{f}_\theta$
$f_\theta(y_t|v',x,y_{<t})$