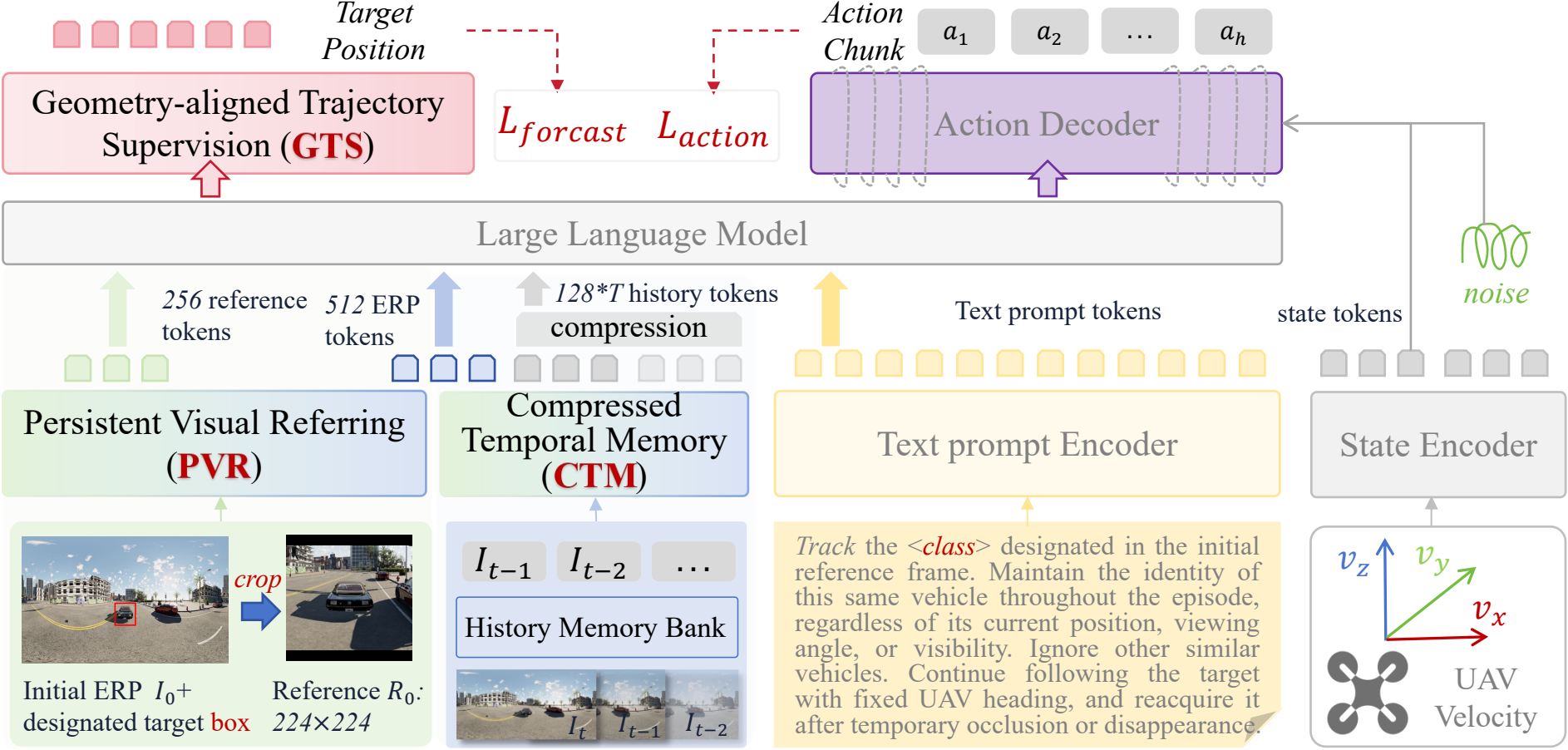

Target Position
Action Chunk
a_1
a_2
...
a_h
Geometry-aligned Trajectory Supervision (GTS)
$L_{forcast}$
L_{action}
Action Decoder
Large Language Model
256 reference tokens
512 ERP tokens
128*T history tokens
compression
Text prompt tokens
state tokens
noise
Persistent Visual Referring (PVR)
Compressed Temporal Memory (CTM)
Text prompt Encoder
State Encoder
crop
Initial ERP I_0+ designated target box
Reference R_0: 224×224
I_{t-1}
I_{t-2}
...
History Memory Bank
I_t
I_{t-1}
I_{t-2}
Track the <class> designated in the initial reference frame. Maintain the identity of this same vehicle throughout the episode, regardless of its current position, viewing angle, or visibility. Ignore other similar vehicles. Continue following the target with fixed UAV heading, and reacquire it after temporary occlusion or disappearance.
v_z
v_y
v_x
UAV Velocity