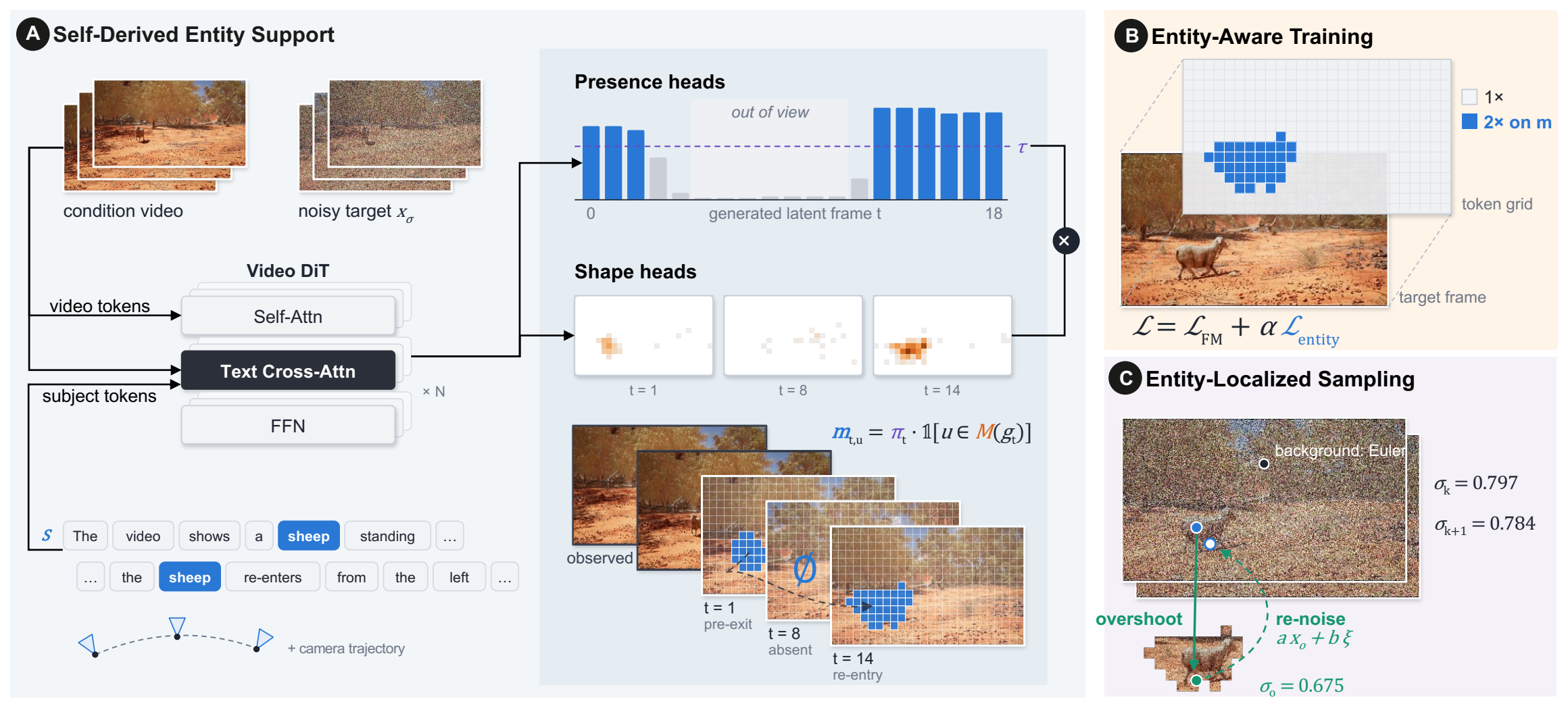

A Self-Derived Entity Support
condition video
noisy target x_σ
Video DiT
video tokens
Self-Attn
Text Cross-Attn
subject tokens
FFN
× N
s
The
video
shows
a
sheep
standing
…
…
the
sheep
re-enters
from
the
left
…
+ camera trajectory
Presence heads
out of view
τ
0
generated latent frame t
18
Shape heads
t = 1
t = 8
t = 14
$m_{t,u} = \pi_t \cdot \mathbb{1}[u \in M(g_t)]$
observed
t = 1
pre-exit
t = 8
absent
t = 14
re-entry
B Entity-Aware Training
1×
2× on m
token grid
target frame
$\mathcal{L} = \mathcal{L}_{FM} + \alpha \mathcal{L}_{entity}$
C Entity-Localized Sampling
background: Euler
$\sigma_k = 0.797$
$\sigma_{k+1} = 0.784$
overshoot
re-noise
$a x_o + b \xi$
$\sigma_o = 0.675$