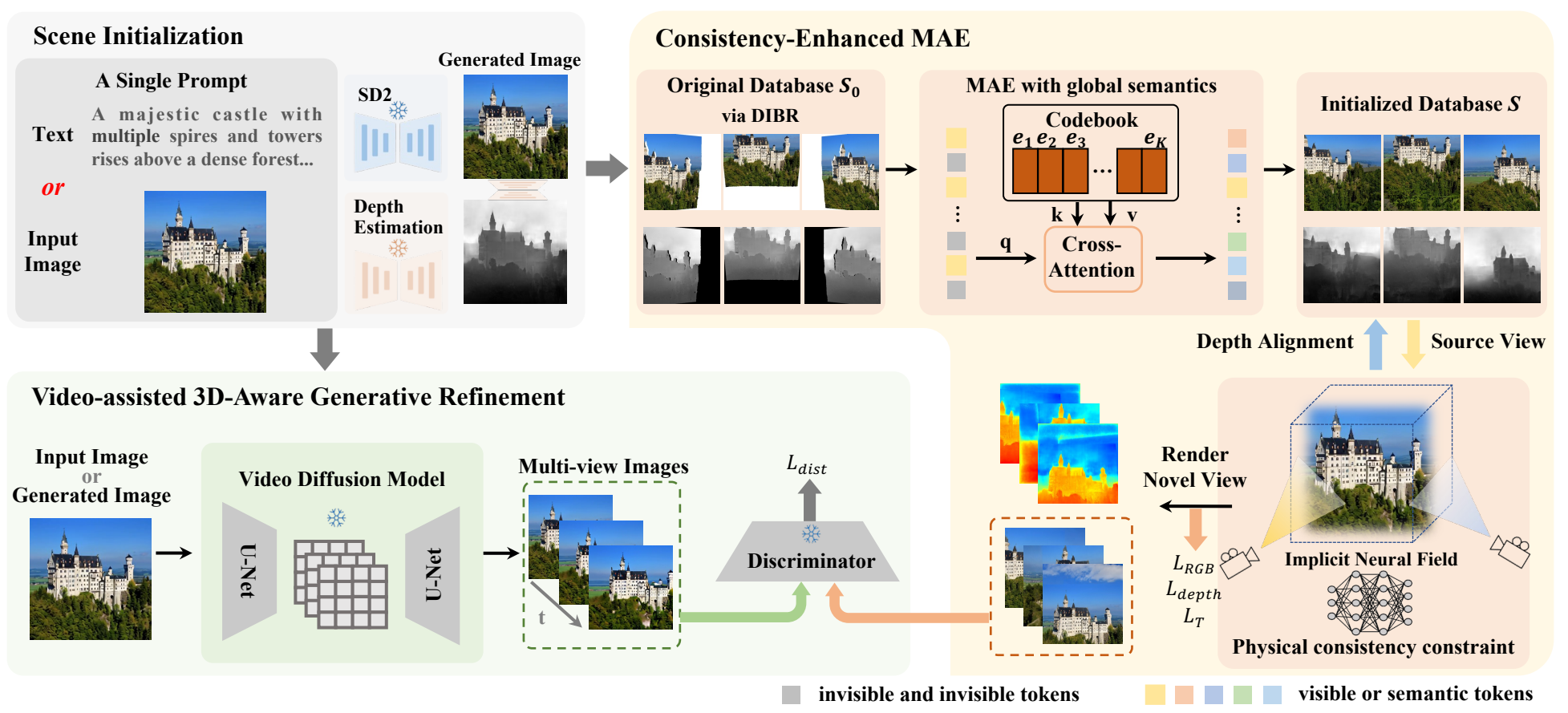
Scene Initialization
A Single Prompt
Text
A majestic castle with multiple spires and towers rises above a dense forest...
or
Input Image
SD2
Generated Image
Depth Estimation
Consistency-Enhanced MAE
Original Database S_0
via DIBR
MAE with global semantics
Codebook
e_1 e_2 e_3 ... e_K
k
v
q
Cross-Attention
Initialized Database S
Depth Alignment
Source View
Video-assisted 3D-Aware Generative Refinement
Input Image
or
Generated Image
Video Diffusion Model
U-Net
U-Net
Multi-view Images
t
L_{dist}
Discriminator
Render Novel View
L_{RGB}
L_{depth}
L_T
Implicit Neural Field
Physical consistency constraint
invisible and invisible tokens
visible or semantic tokens