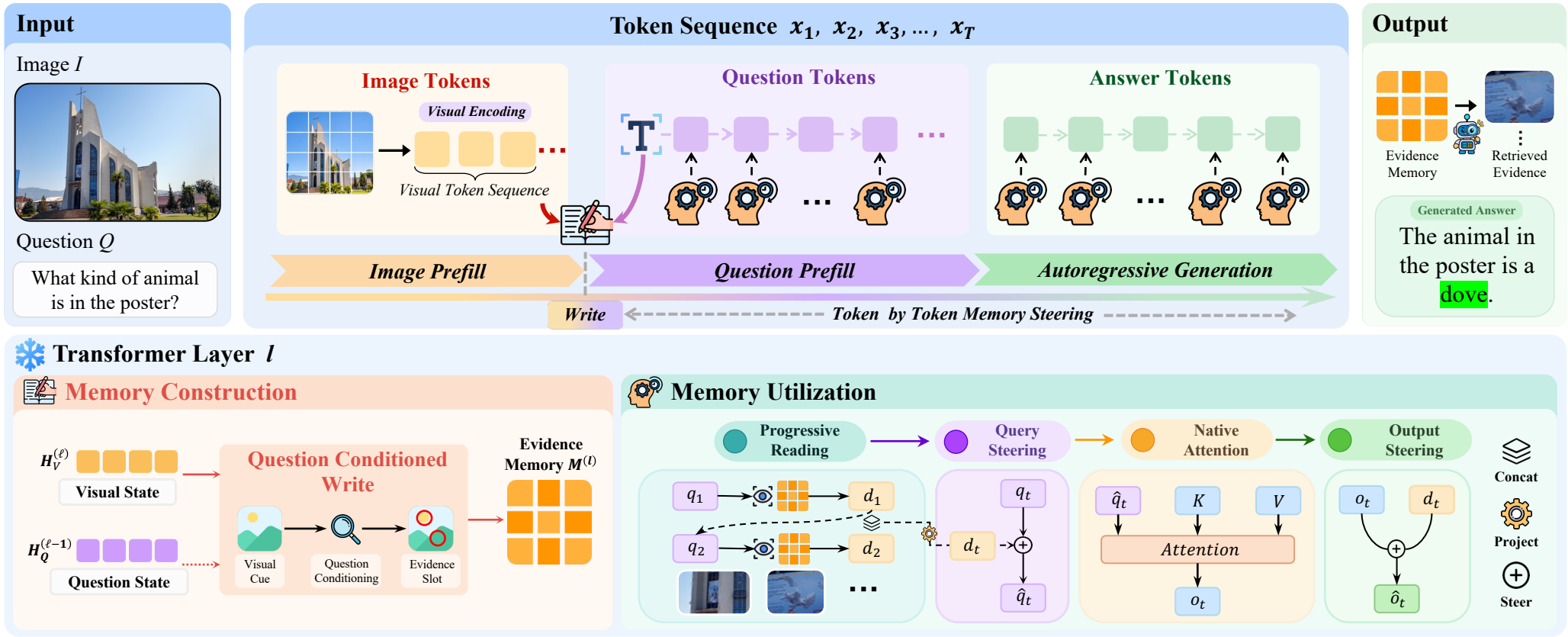

Input
Image I
Question Q
What kind of animal is in the poster?
Token Sequence $x_1, x_2, x_3, \ldots, x_T$
Image Tokens
Visual Encoding
Visual Token Sequence
Question Tokens
Answer Tokens
Image Prefill
Question Prefill
Autoregressive Generation
Write
Token by Token Memory Steering
Output
Evidence Memory
Retrieved Evidence
Generated Answer
The animal in the poster is a dove.
Transformer Layer l
Memory Construction
$H_V^{(\ell)}$
Visual State
$H_Q^{(\ell-1)}$
Question State
Question Conditioned Write
Visual Cue
Question Conditioning
Evidence Slot
Evidence Memory $M^{(l)}$
Memory Utilization
Progressive Reading
Query Steering
Native Attention
Output Steering
q_1
q_2
d_1
d_2
d_t
q_t
$\hat{q}_t$
K
V
Attention
o_t
$\hat{o}_t$
Concat
Project
Steer