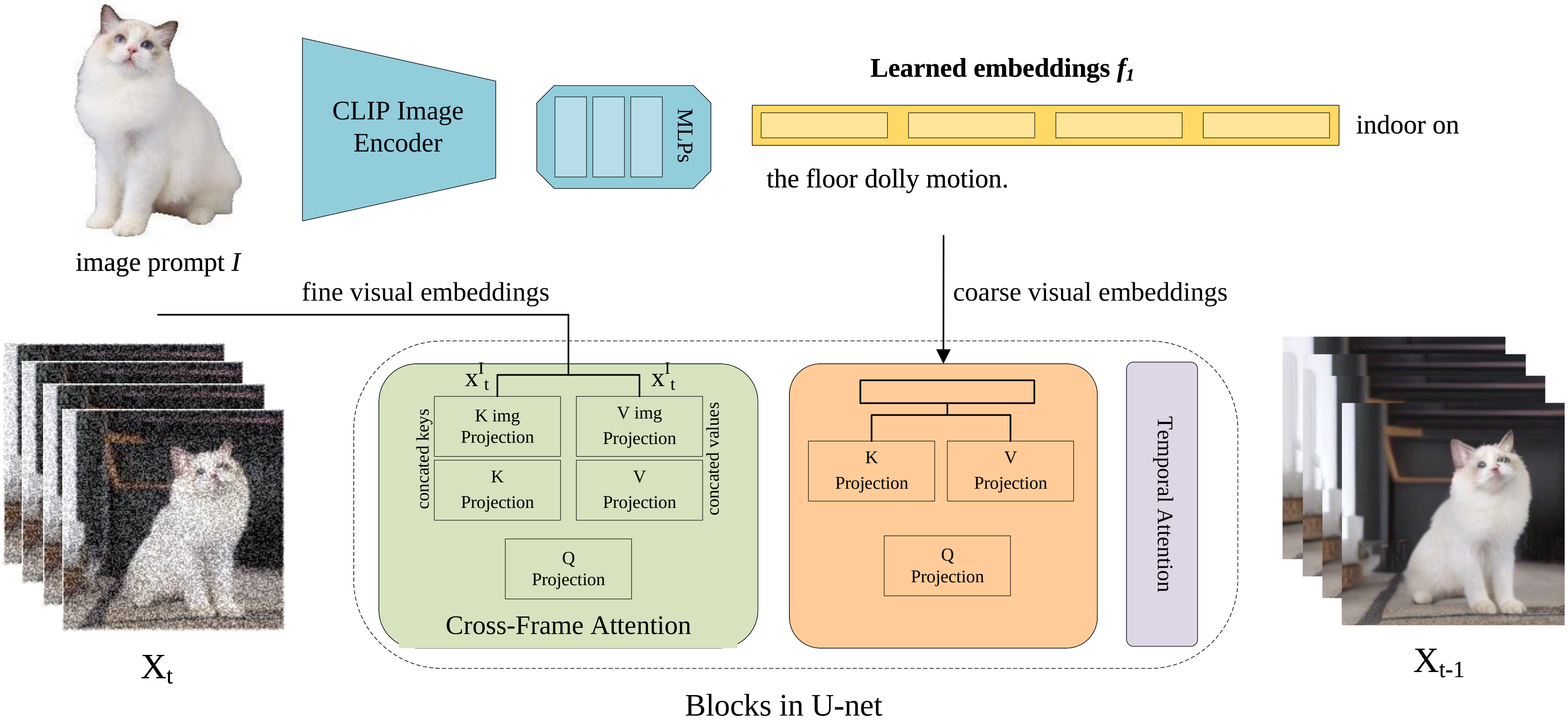
image prompt *I*
CLIP Image Encoder
MLPs
Learned embeddings f_1
indoor on
the floor dolly motion.
fine visual embeddings
coarse visual embeddings
x^I_t
x^I_t
concated keys
K img Projection
V img Projection
concated values
K Projection
V Projection
Q Projection
Cross-Frame Attention
K Projection
V Projection
Q Projection
Temporal Attention
X_t
X_{t-1}
Blocks in U-net