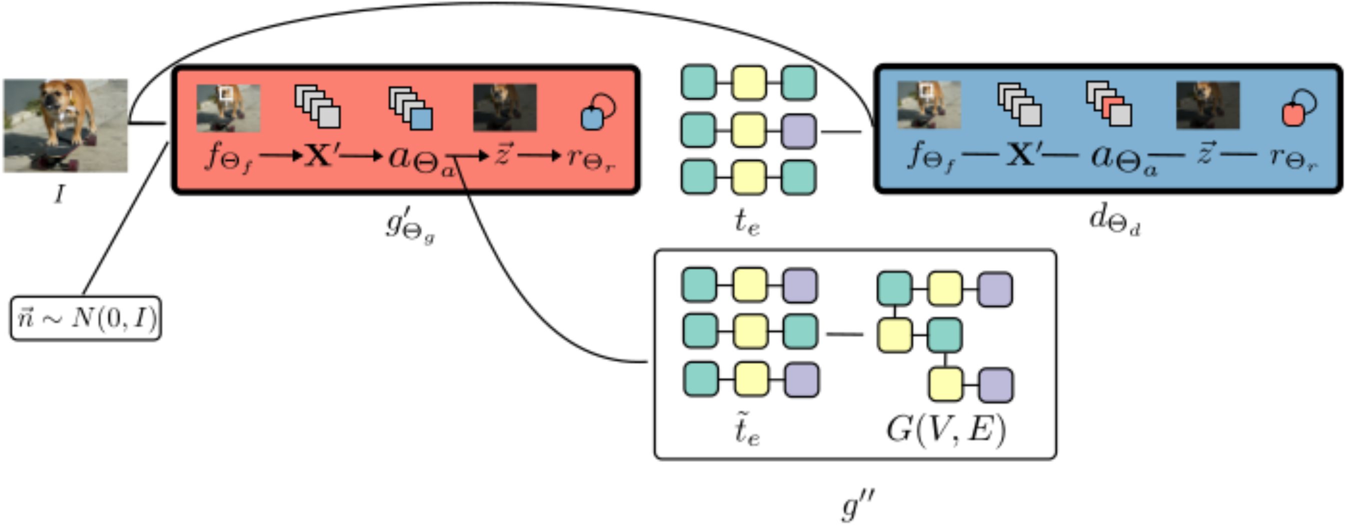

Figure 2: Our architecture: Input images I and noise $\vec{n}$ are first fed to the generator g_{Θ_g} (red), which processes the image using a CNN f_{Θ_f}, generates image features $\mathbf{X}'$, passes those to an attention mechanism a_{Θ_a} that generates a dynamic image representations $\vec{z}$ and attention vectors $\vec{\alpha}$. $\vec{z}$ is fed to an LSTM that produces triples $\tilde{t}_e$. During training, these triples are passed along with ground truth triples t_e to the discriminator d_{Θ_d} (blue) that contains the same components as the generator. The discriminator only produces a score, however. During test time all $\tilde{t}_e$s and αs are passed to g'' which resolves the triples into a graph $G(V, E)$.