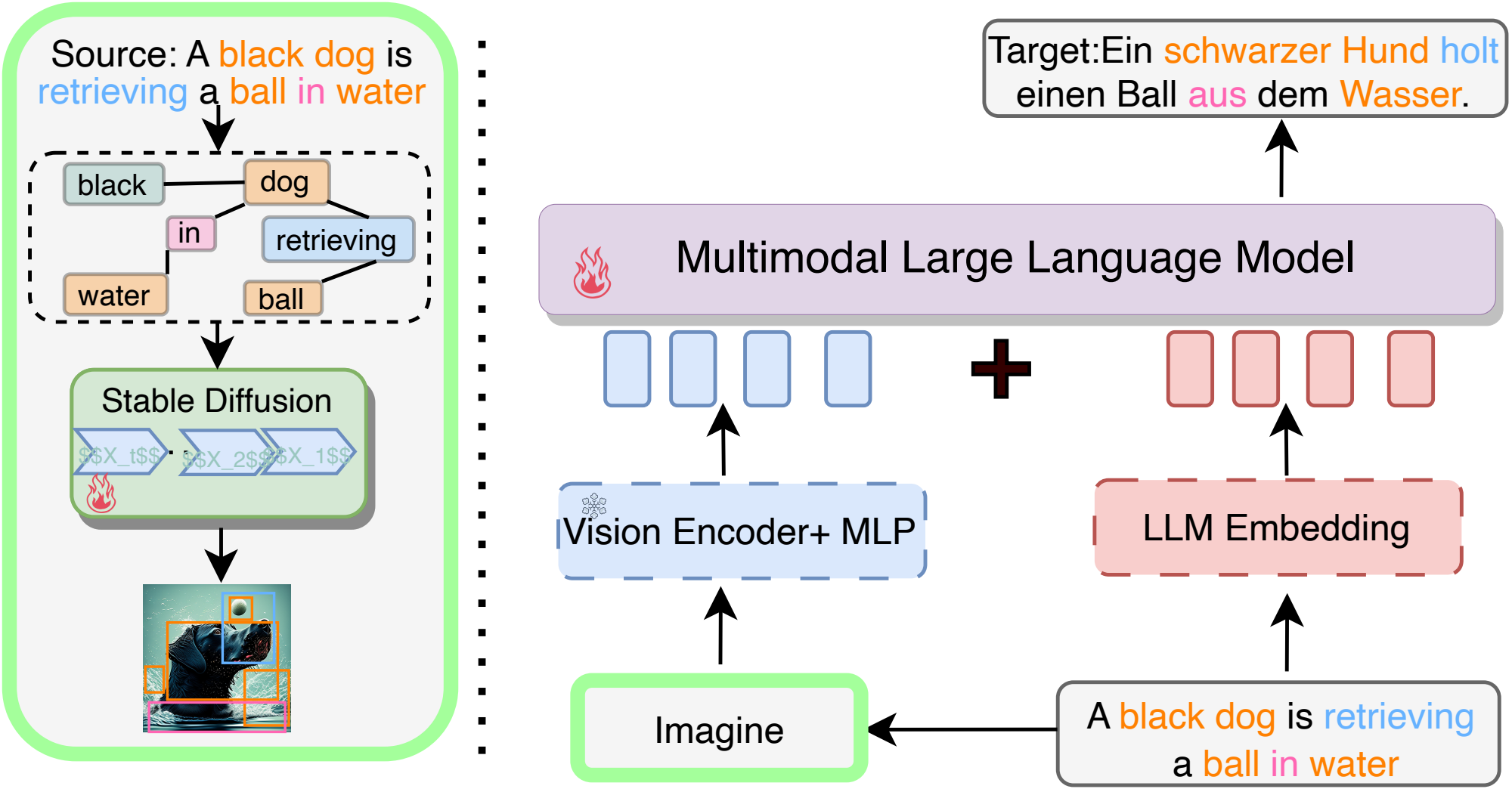
Source: A black dog is retrieving a ball in water
black
dog
in
retrieving
water
ball
Stable Diffusion
$$X_t$$
$$X_2$$
$$X_1$$
Target:Ein schwarzer Hund holt einen Ball aus dem Wasser.
Multimodal Large Language Model
Vision Encoder+ MLP
LLM Embedding
Imagine
A black dog is retrieving a ball in water