(training only)

Indoor shot of intersection at night

Innenaufnahme einer Kreuzung bei Nacht

(a) Multimodal machine translation pipeline (no image during inference)

Image ↔ Lang 1

Lang 1 ↔ Lang 2

Multimodal

Multilingual

(b) Pre-trained Models