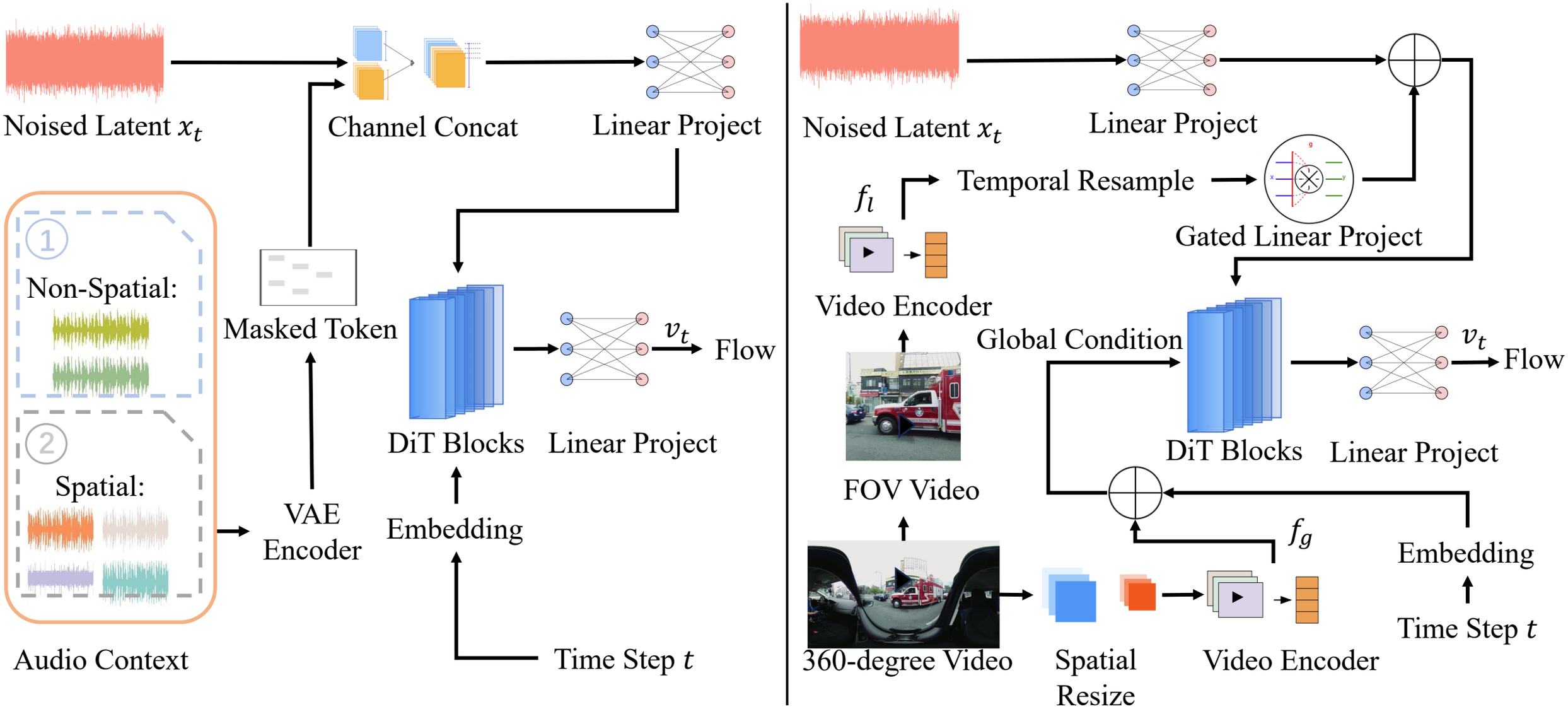

(a) Self-Supervised Coarse-to-Fine Pre-training

(b) Spatial-aware Supervised Fine-tuning