# a) Spatio-Procedural Topological Mapping

b) Dual-Stream Encoding

Cross-Modal Graph Encoding

G_t^s

Stop Node

Candidate Nodes : $\bigcirc$, $\{\bigcirc\}^{|\mathcal{C}_t|}$.

"Go up the stairs and stop at the top in front of a mirror."

Distance Matrix

Node Embeddings → Graph Transformer

Text Embeddings

Cross-Modal Route Encoding

Existed Routes in G_{t-1}^p

Initial Node

Newly Formed Routes in G_t^p

Initial Node

Stop Node

Route Representations → Temporal Transformer

Casual Mask
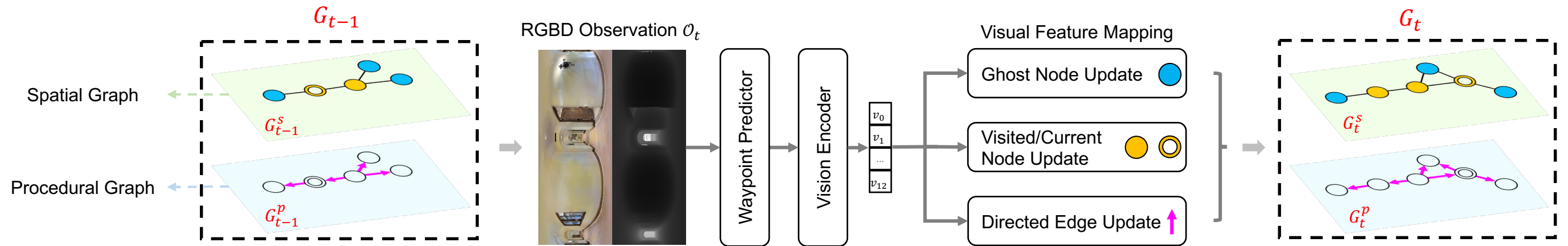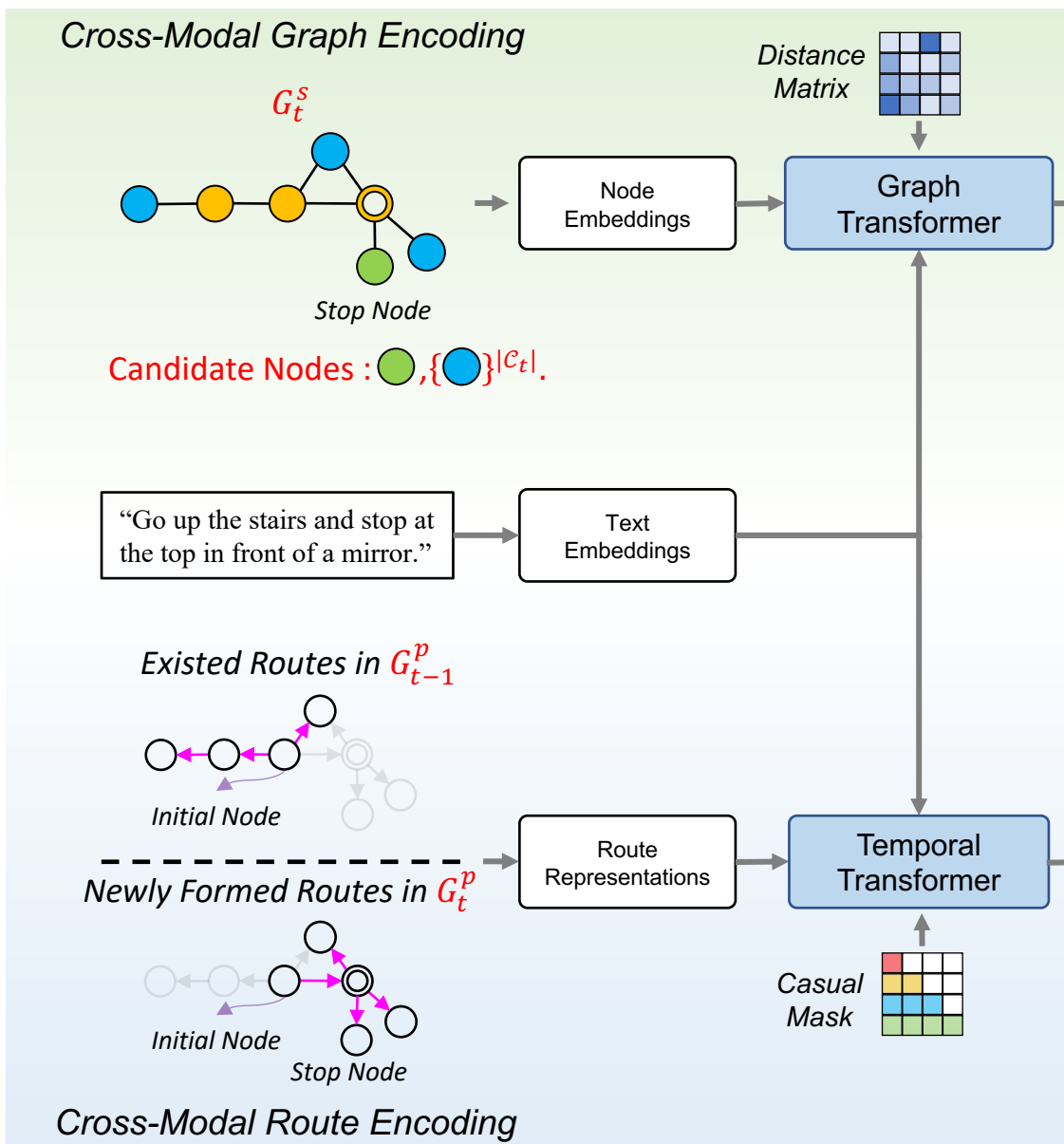
c) Hierarchical Policy Integration

Feature-Level Collaboration

X_t^g → Policy Head → S_t^g

$[x_0^g; x_0^r]$ → C → FFN → S

$1 - \alpha$

α

$\oplus$ X_t^f → Cross Attention → $\oplus$ → Self Attention → $\oplus$ → Policy Head → S_t^f

Text Embeddings

X_t^r → Policy Head → S_t^r

Logit-Level Fusion

$$S_t^r + S_t^g + S_t^f = \tilde{S}_t$$

$\mathcal{L}_{CE}$ $\mathcal{L}_{CE}$ $\mathcal{L}_{CE}$ $\mathcal{L}_{CE}$

Legend:
Graph Embeddings | Route Embeddings | Logits
Current Node | Visited Node | C Concatenation
Ghost Node | Stop Node | S Sigmoid Function