(a) Semantic Anchor Disambiguation

z_{text}^{init}

P_{cat} : P_{dog} : P_{table}

Prompt:
A **cat** and a **dog** resting on a **table**

T5

Subject Anchors

A_{cat} A_{dog} A_{table}

Subject List:
cat   dog   table

Ambiguation Scale Calculation

Ambiguation Scale
s_{cat}
s_{dog}
s_{table}

TimeStep t

Semantic Disambiguation Interpolation

A_{cat} + P_{cat} → s_{cat}
A_{dog} + P_{dog} → s_{dog}
A_{table} + P_{table} → s_{table}

+ Positive pair   − Negative pair

Disambiguation Pathway
$\Delta_{A_{cat}}$
A_{cat}
A_{dog}
A_{table}
P_{cat}^*
Guidance
P_{cat}
$+ \omega(t) \cdot s_{cat} \cdot \Delta_{A_{cat}}$

z_{text}^{init}
P_{cat}^* : P_{dog}^* : P_{table}^*

z_{video}

Concat

(b) Dynamic Layout Fusion Attention

Model-adaptive Layout Mask Generation

L^{adapt}

L^{fuse}

Q_{cat}
Q_{dog} Query
Q_{table}

Top K
Treshold

$\oplus$

LLM Layout Mask Generation

LLM Plan

cat dog
cat dog
table

L^{prior}

K

t-t attn		t-v attn
v-t attn		v-v attn

Q

Subject-aware Attention Mask

Attention Contol

MM-DIT Block : MM-DIT Block
× N

T Steps

3D VAE Decoder