

文本部分

儿童与猫共骑
童车，旁边停
着经典 ...

标记化

37
42
74
26
...

加噪

M
42
M
59
...

集中注意
力掩码



加噪步

{0,1,...,T} 采样
→ t

图像部分



CLIP
模型

[CLS]

长度预测任务

37 42 74 26 ...

Transformer

(a) DDCap训练过程

长度预测: N ← [CLS]

CLIP
模型



首佳推理策略

初始化
N 个掩码
标记

M
M
M
M
...

Transformer

11
22
74
33
...

重新
加噪

M
M
74
M
...

Transformer

...

37
42
74
26
...

t 采样步

(b) DDCap推理过程