v_1

v_n

Visual Encoder

$H(s_1)$

$F(v_1)$

$H(s_n)$

$F(v_n)$

Textual Encoder

A man is jumping high up and spreading his legs.

s_1

A woman and a man are hugging each other indoors.

s_n

Visual-Textual Concept Decoder

(Un)likelihood Training

man

woman

...

suit

hat

...

jump

...

$\hat{G}(v_1)$

man

woman

...

outdoors

indoors

...

hug

...

$\hat{G}(v_n)$

man

woman

...

male

...

jump

legs

...

$\hat{G}(s_1)$

man

woman

...

outdoors

...

female

embrace

...

$\hat{G}(s_n)$
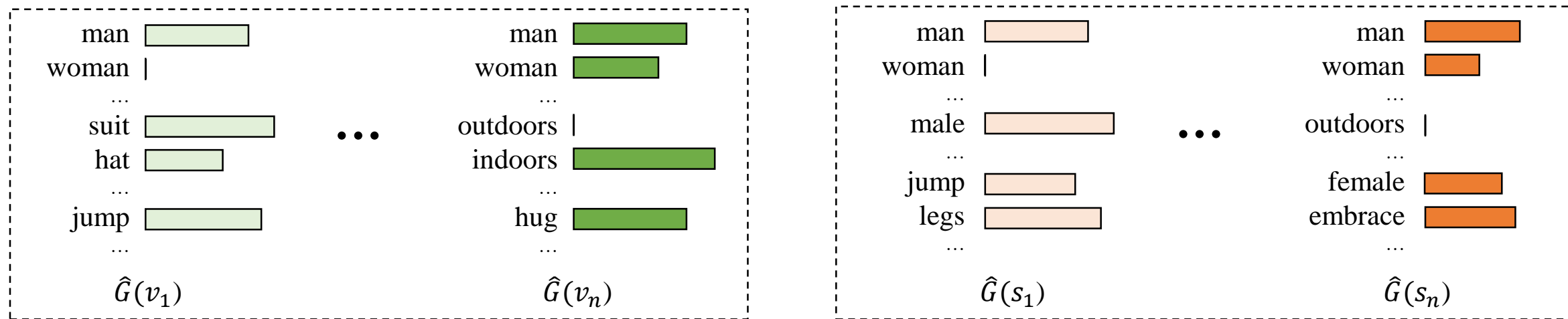
Visual embedding's interpretation

Textual embedding's interpretation