

Visual Grounding

How many cups appear in this video?

Q-Gate

$w_{grounding}$ w_{match} $w_{context}$

Σ

baseline

or

static *unimodal*

keyframes

VLM → "Five"

Visual Attribute

What is the color of the tall drinking glass initially held by the woman?

Q-Gate

$w_{grounding}$ w_{match} $w_{context}$

Σ

baseline

or

static *unimodal*

keyframes

VLM → "Blue"

Narrative Reasoning

According to this video, what is the purpose of adding fresh produce to the water?

Q-Gate

$w_{grounding}$ w_{match} $w_{context}$

Σ

baseline

or

static *unimodal*

"subtle flavor try adding fresh produce"

subtitle

VLM → "To infuse the water with a refreshing and subtle flavor"