B : Cross-Modal Bias
Z : Visual Confounder
V : Video Feature
T : Text Feature
M : Visual Mediator
Y : Label

(a) Original SCM

Textual back-door causal intervention

(b) SCM after textual back-door causal intervention

Visual front-door causal intervention

(c) SCM after visual front-door causal intervention

Discover Causal Structure

(d) Deconfounded SCM

V (Video)

T (Text): Make scrambled egg

Y (Label): Are the video and text a match pair?
"matched" / "unmatched"

Cross-Modal Bias

Unmatched ✕ T *Matched*

stirfry egg
add salt & pepper

Visual Confounder

(background) (attire)

(lighting) ...

Debiased Textual Embeddings

Sequential Causality

butter pan → crack egg

Co-occurring Causality

put egg to plate
stirfry egg

Deconfound Visual Embeddings

take eggs
butter pan
crack egg
stirfry egg
put egg to plate