Counterfactual Training Impacts on Transformer-Based VQA Efficiency Under Adversarial Conditions

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the effect of counterfactual training on the inference efficiency and throughput of transformer-based VQA architectures under adversarial perturbations. Videos often capture objects, their visible properties, their motion, and the interactions between different objects. Objects also have physical properties such as mass, which the imaging pipeline is unable to directly capture. 20 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CRIPP-VQA: Counterfactual Reasoning about Implicit Physical Properties via Video Question Answering. Research question: What is the effect of counterfactual training on the inference efficiency and throughput of transformer-based VQA architectures under adversarial perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

14 papers retrieved. 20 claims extracted; 0 independently verified. Quality review score: 3.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The task is to predict the answer (a) given an input video (v) and a question (q).	×	0.07
Each video v contains m number of objects randomly selected from the set O = {o1, o2, ..., on}.	×	0.06
Object oi has several associated properties (i.e., oi = (mi, ci, si, ti, li, vi)), where color (ci), shape (si), texture	×	0.05
The evaluation metrics used are per-option (PO) and per-question (PQ) accuracy.	×	0.06
Per-option accuracy refers to the option-wise whether all options are correctly predicted or not.	×	0.01
Each planning task involves performing an action over objects within a video.	×	0.06
TDW is used to re-simulate the models' predictions on the original video to check whether the given planning goal is ach	×	0.04
The intended functionality of MONet is to decompose individual objects into separate masks.	×	0.03
The predicted masks by MONet contain areas corresponding to more than one object.	×	0.01
Replacing MONet with Mask-RCNN in Aloe (Aloe*) leads to more reliable object detection.	×	0.01
Three different state-of-the-art models are considered for the video question answering task: MAC, HCRN, and Aloe.	×	0.10
MAC is modified by performing channel-wise feature concatenation of each frame to adapt to video inputs.	×	0.02
HCRN uses a hierarchical strategy to learn the relation between the visual and textual data.	×	0.04
Aloe is a transformer-based model designed for object trajectory-based complex reasoning over synthetic datasets.	×	0.06
Aloe uses MONet for obtaining object features by performing an unsupervised decomposition of each frame into objects.	×	0.03
Aloe takes frame-wise object features to predict the answers to the input question, using the [CLS] token and employs a	×	0.05
The model predicts the most frequent label to analyze textual biases.	×	0.03
Blind-BERT is a pretrained text-only QA model that takes only questions as input to predict the answer and ignores the v	×	0.05
The demo page contains several examples of the CRIPP-VQA dataset.	×	0.11
Table 8 shows the types of questions asked in the CRIPP-VQA dataset.	×	0.12

References

- http://arxiv.org/abs/2211.03779v1
- http://arxiv.org/abs/1909.09192v1
- http://arxiv.org/abs/2103.15670v3