Causal-Adapter: Taming Text-to-Image Diffusion for Counterfactual Image Generation

Lei Tong, Zhihua Liu, Chaochao Lu, Dino Oglic,

Tom Diethe, Philip Teare, Sotirios A. Tsaftaris, Chen Jin

Paper: arxiv.org/abs/2509.24798

Website: https://leitong02.github.io/causaladapter/

Why Counterfactual Images?

Visual “what-if” reasoning can support controllable editing and scientific interpretation.

Human faces

What would this person look like if…

they were younger or older?

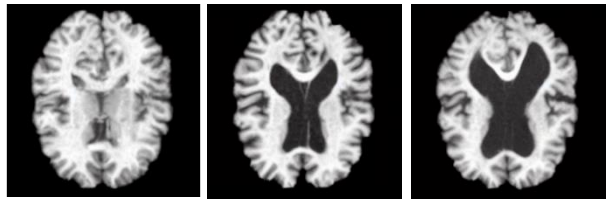
the gender attribute changed?

Goal: change what should change, preserve identity-specific details.

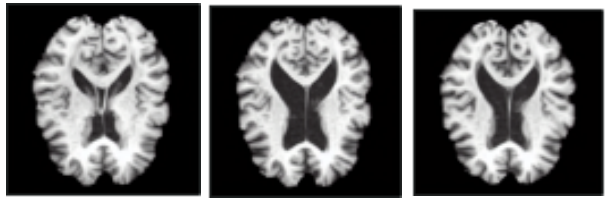
Medical imaging

What would this MRI look like if…

ventricular volume increased?



different slice class?

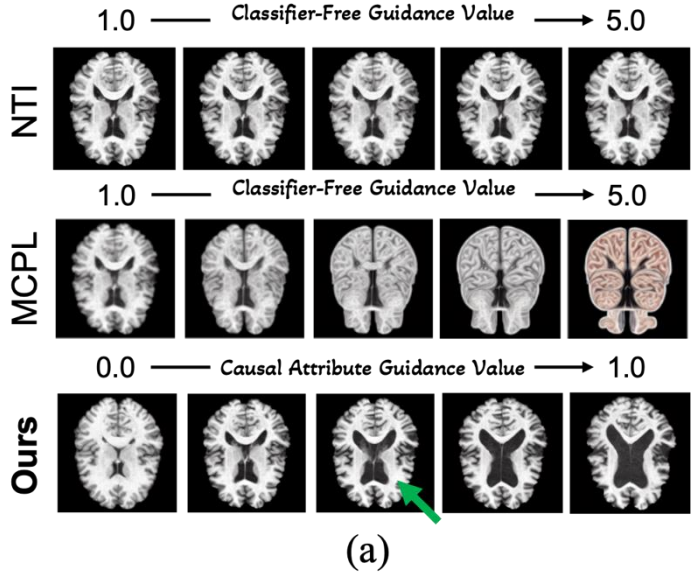


Goal: produce localized anatomical changes while preserving patient-specific structure.

T2I Editing Is Not Sufficient for Faithful Counterfactual Generation

Strong generative priors are helpful, but prompt-driven editing lacks precise causal control.

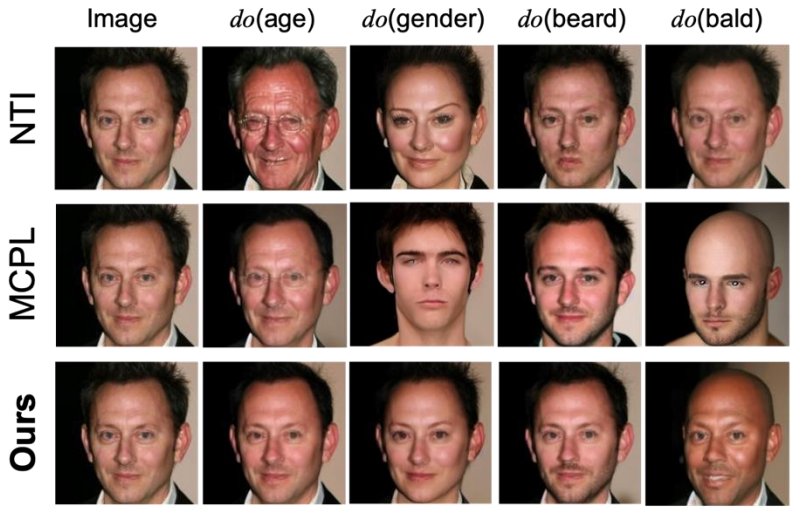
Key problem **A prompt can describe an edit, but it does not define a causal mechanism.**



(a)

1 Numerical attributes are hard to encode

Text encoders are built for natural language, not precise continuous values such as age, brain volume, ventricular volume, or slice index.



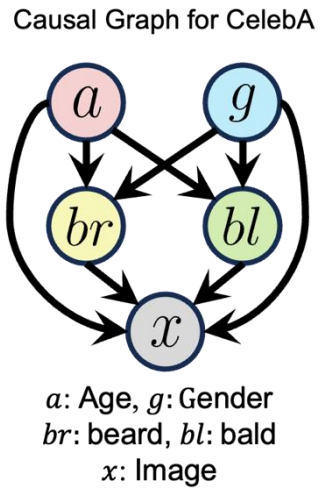
(b)

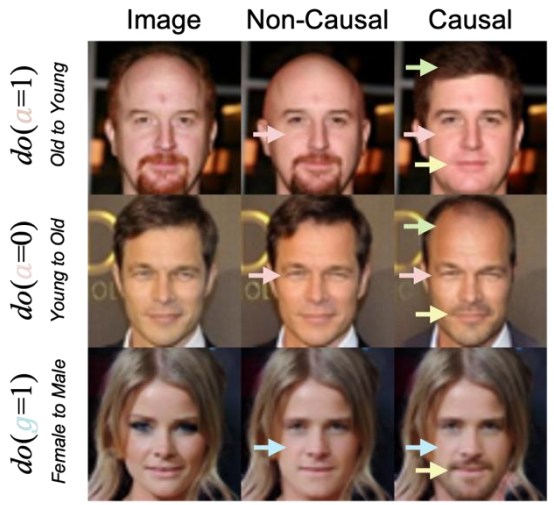
2 No explicit SCM and attribute entanglement

Prompt editing may change the target attribute, but also unintended details such as beard, hair, or identity, because causal relations are not explicitly modeled.

What Makes a Counterfactual Faithful?

Not every visually plausible edit is causally faithful.





Non-causal editing may only change the target attribute.

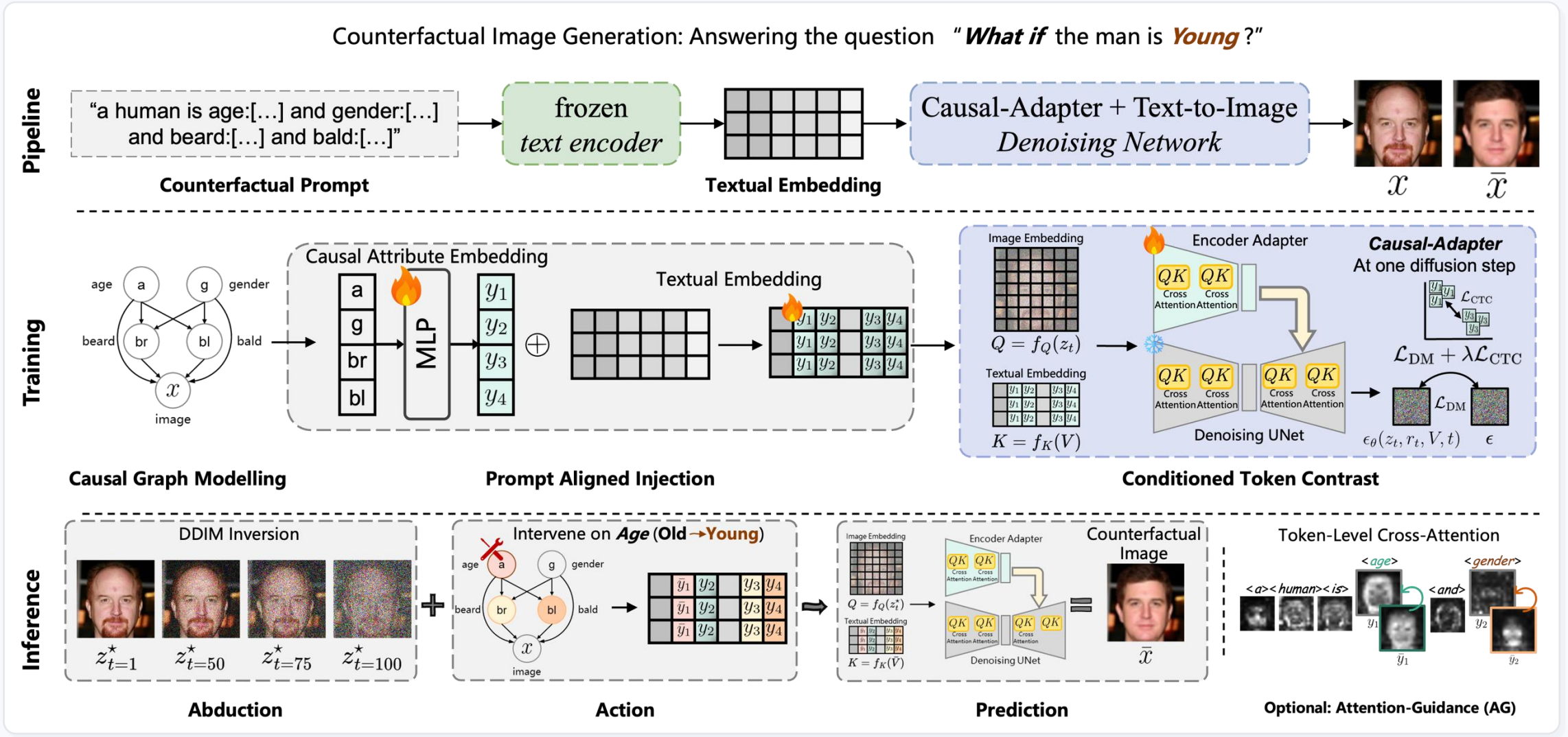
Causal editing propagates the intervention through related attributes.

1. **Effectiveness**
   Does the target intervention succeed?
2. **Causal propagation**
   Do dependent attributes change consistently?
3. **Minimality**
   Are unrelated attributes minimally affected?
4. **Identity preservation**
   Does the image remain the same subject?

Faithful counterfactual generation = reasoning + generation

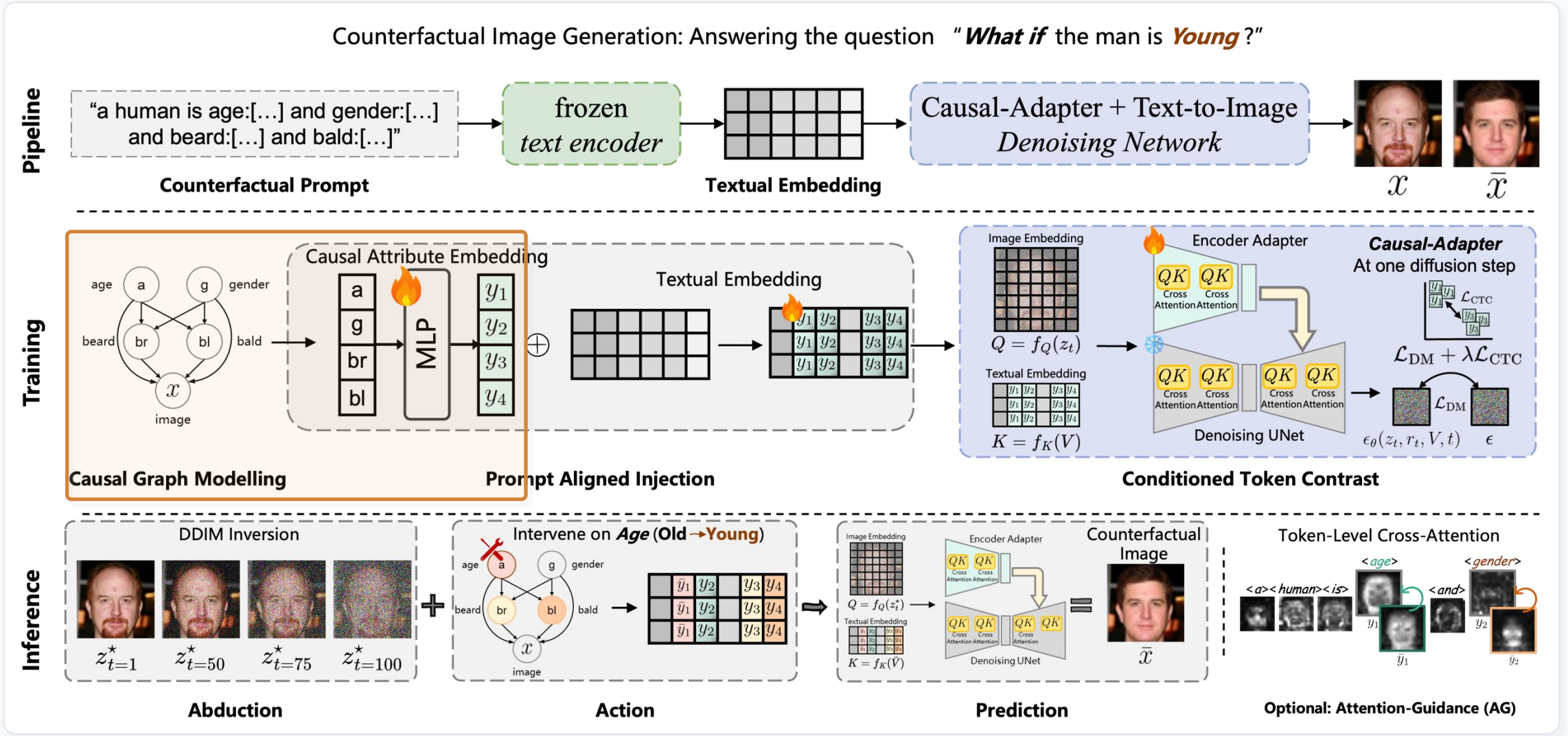
Method: Causal-Adapter

Use a frozen T2I diffusion model, but inject structured causal semantics through a lightweight adapter.



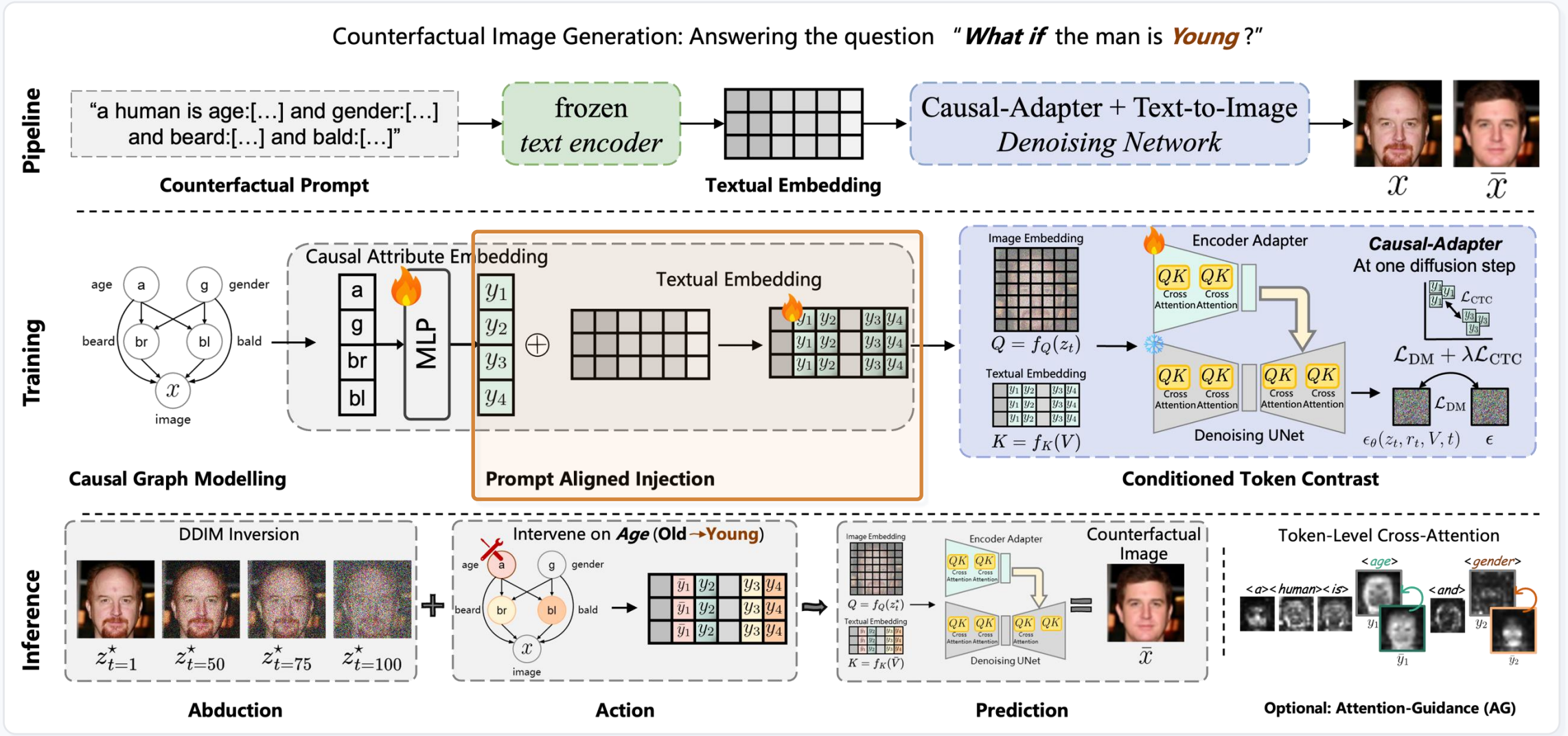
Method: Causal-Adapter

Use a frozen T2I diffusion model, but inject structured causal semantics through a lightweight adapter.



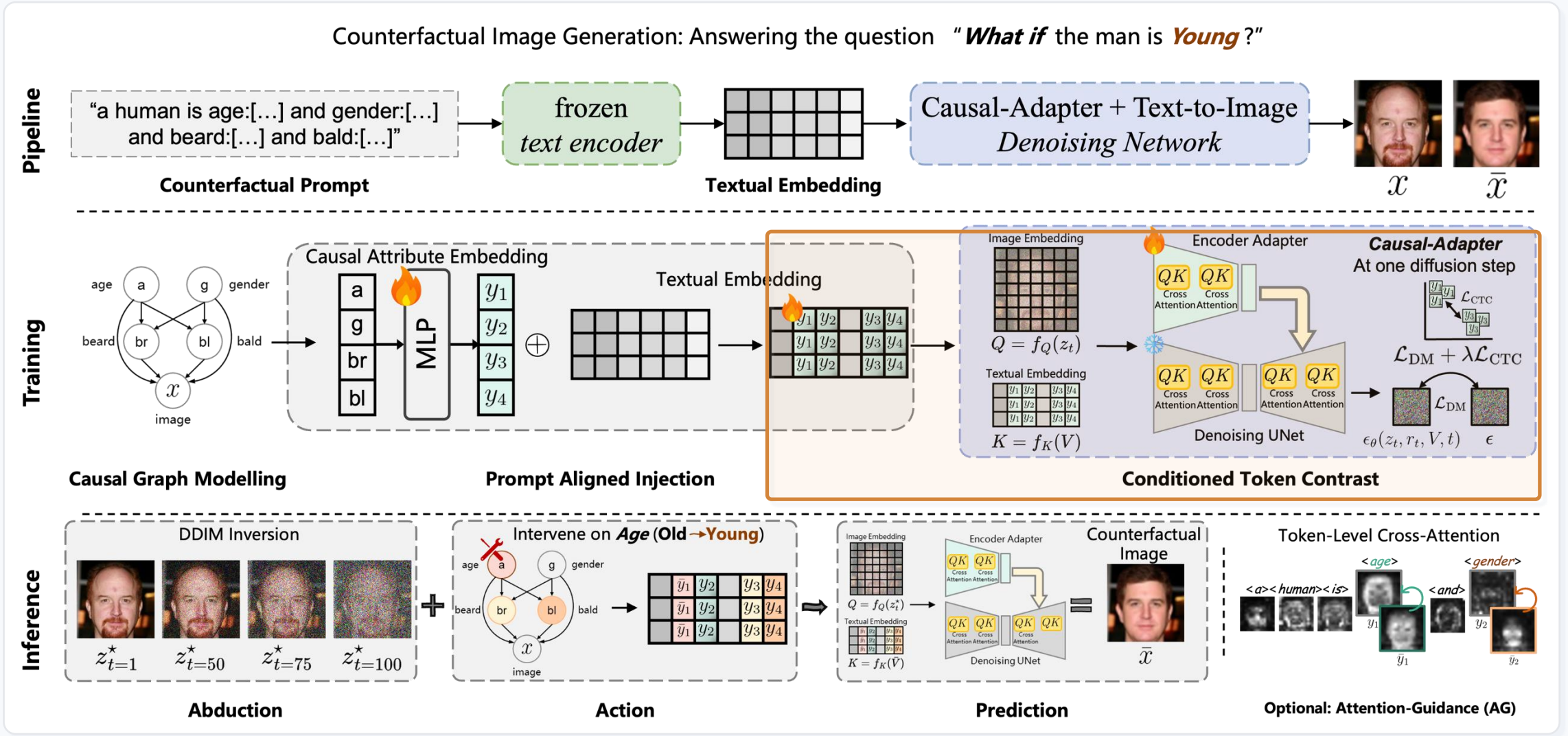
Method: Causal-Adapter

Use a frozen T2I diffusion model, but inject structured causal semantics through a lightweight adapter.



Method: Causal-Adapter

Use a frozen T2I diffusion model, but inject structured causal semantics through a lightweight adapter.



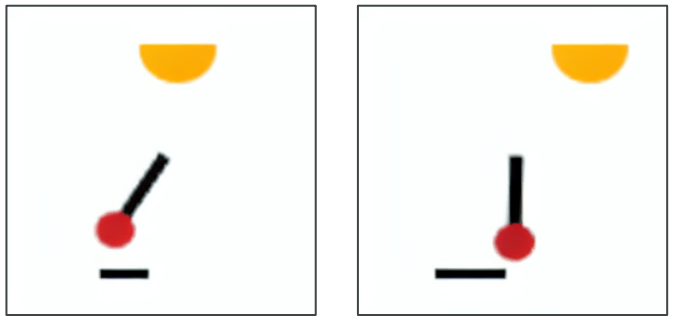
Results: Faithful Changes Across Domains

Causal-Adapter improves counterfactual effectiveness, realism, and identity preservation.

Physics

Pendulum

do(Pendulum) *do*(Light)



Pendulum/light interventions propagate to shadow attributes in accordance with physical laws.

Faces

CelebA

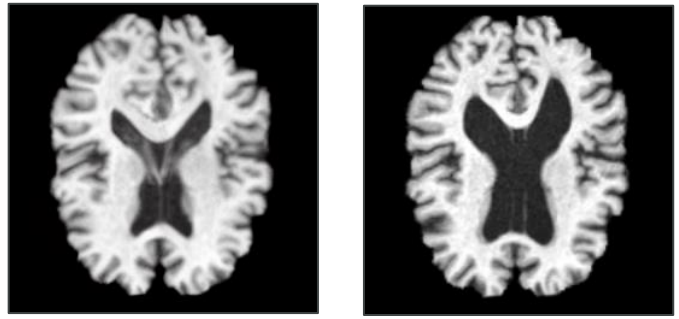
do(Age) *do*(Gender)

Identity-preserving face edits update causally related attributes (beard, bald) while maintaining identity.

Medical images

ADNI brain MRI

do(Patient Age) *do*(Ventricular)



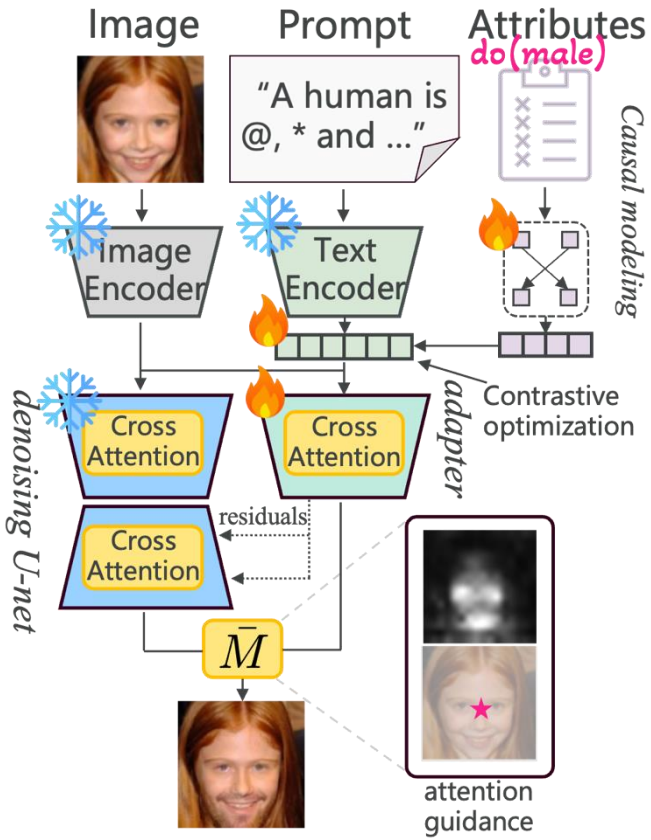
Anatomical edits remain localized and realistic, while preserving subject-specific structure.

Across datasets, Causal-Adapter turns prompt-driven editing into causally faithful counterfactual generation.

Takeaway

An off-the-shelf T2I diffusion model can be tamed with causal semantic attributes to generate faithful counterfactual images.

Causal-Adapter Skeleton



Modular. Causal. Faithful.

Key Highlights

Significant Gains
+50% intervention effectiveness; +87% image quality (FID) improvement over prior methods.

Low Compute
Fine-Tune in 10 hours on a single NVDIA A10G (24GB).

Model-Agnostic
Works with SD1.5, SD 3, FLUX.1 and future T2I backbones.

Precise Attention
Better semantic-spatial alignment in diffusion latents via attention guidance.

Causal Graph Support
Supports learning causal graph from scratch when none is provided.

Open-Source
Code, data, and fine-tuned models will be publicly available.