Diffusion-Based Garment Synthesis via Knowledge Graph-Driven Structural Cross-Modal Semantic Alignment

Miao-Yin Chen, Shu-Han Chuang
Department of Data Science, Soochow University, Taipei, Taiwan

OVERVIEW

AI-driven garment synthesis transforms fashion design.

Challenges

- Semantic drift
- Attribute confusion
- Regional inconsistencies

Our solution

- **Knowledge graph**-driven diffusion model for enhanced control and alignment.

Objective

- Improve semantic fidelity and structural precision in garment synthesis.

Text Prompt

Generate New Image

Problems

Challenges in Garment Synthesis

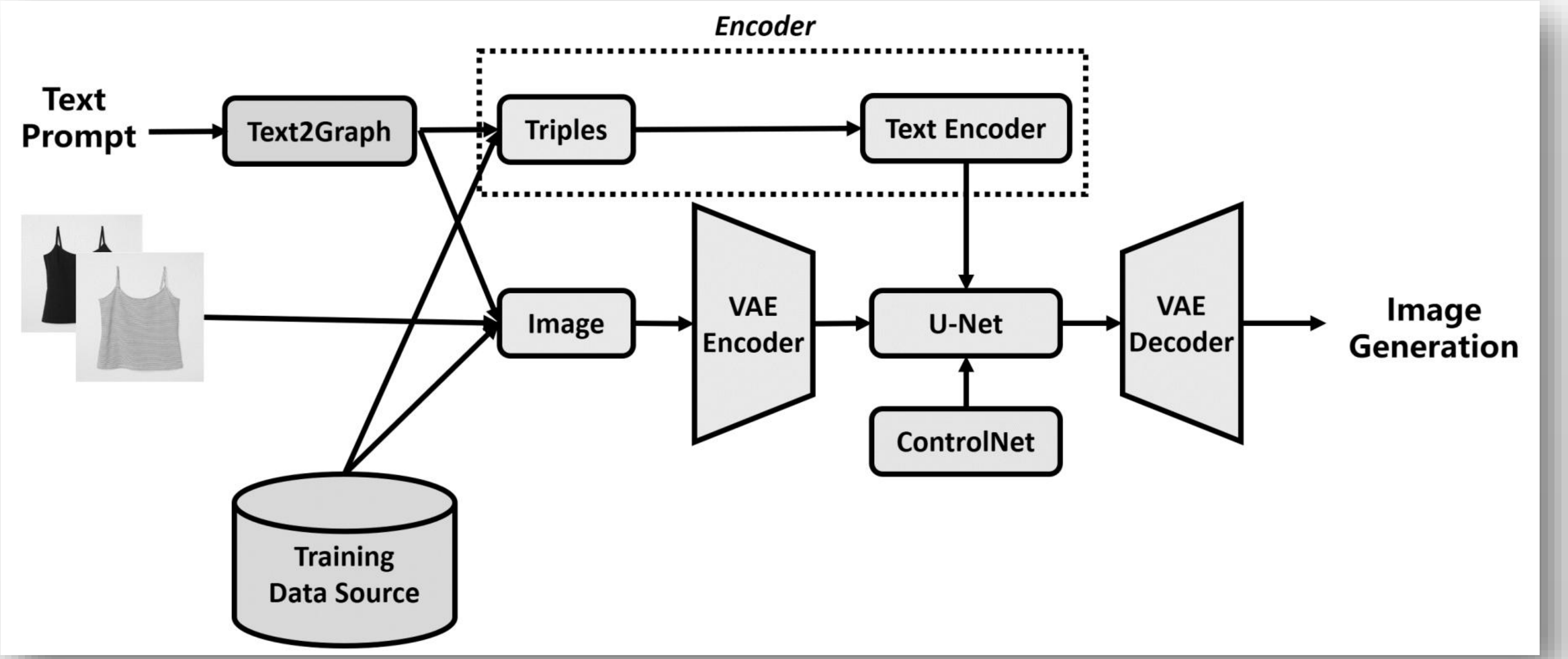
- Semantic drift — **Misalignment** between text prompts and visual outputs.
- Attribute confusion — **Incorrect assignment** of colors, patterns, or styles.
- Regional inconsistencies — Unintended modifications to garment **regions**.

Limitations of Existing Models

- Generative Adversarial Networks — Unstable training, model collapse, limited structural control.

Proposed Approach

We propose a **knowledge graph-driven diffusion model** to enhance control and alignment in garment synthesis



01. Text2Graph Module

Converts natural language into **semantic triples**
Use LLM to encode triples into text embedding

"a navy blue jacket with a red straight-point collar and a blue belted waist."

(jacket, has-color, navy blue),
(jacket, has-feature, collar),
(collar, has-style, straight-point),
(collar, has-color, red),
(jacket, has-feature, waist),
(waist, has-color, blue)

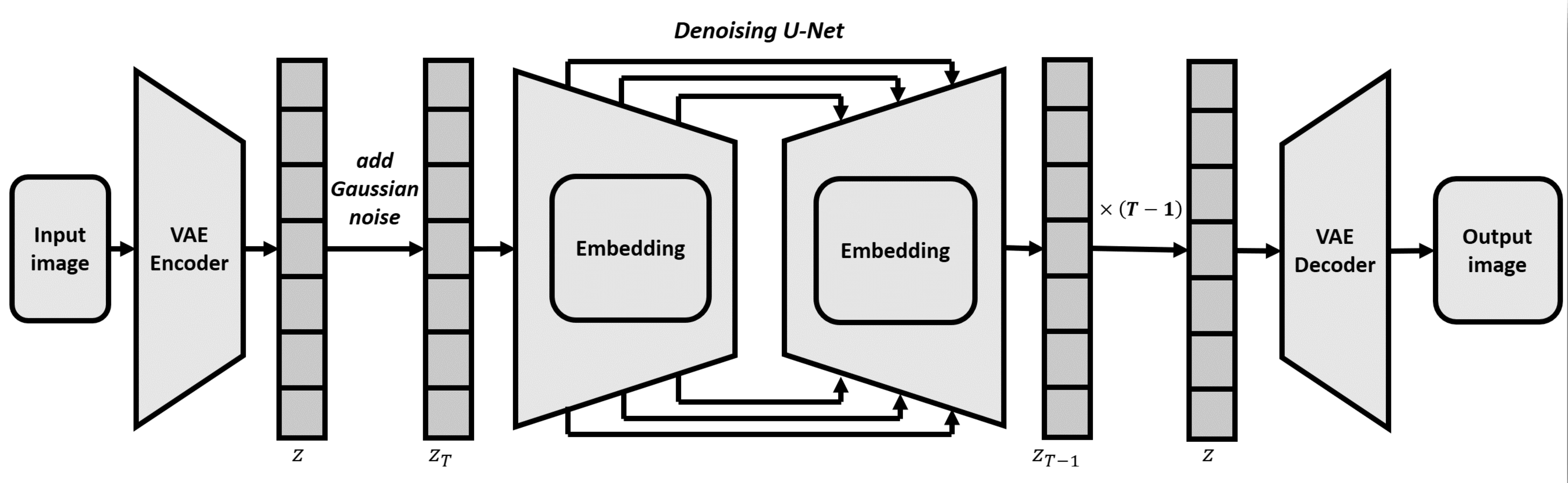
02. Stable Diffusion

Fine-tuned with **LoRA** for fashion-specific tasks.

03. ControlNet

Incorporates structural inputs (pose, sketches) to ensure the generated images maintain **spatial** association.

Diffusion-Based Garment Synthesis



- Forward phase:
  Adds Gaussian noise to images.
- Reverse phase:
  U-Net iteratively removes noise to reconstruct images.
- VAE:
  Compresses images into latent space for efficiency.

High-fidelity garment images with accurate textures, shapes, and colors.

Experiments

01. Datasets

H&M Personalized Fashion Recommendations dataset from Kaggle competition (approx. 105,000 images)

02. Parameters Setting

- Fine-tuned Stable Diffusion v1.4 with LoRA
- *AdamW* optimizer, learning rate of 1×10^{-6}, over 200,000 iterations
- Conducted on a single NVIDIA A6000 GPU
- Images resized to 512×512

03. Evaluation Metrics

CLIP Score → Text-image semantic alignment.

Fréchet Inception Distance (FID) → Visual realism

Inception Score (IS) → Image quality and diversity.

Results

Model	CLIP Score	FID Score	Inception Score
Stable diffusion	31.54	290.15	1.76
Fine-tuned Stable diffusion	28.56	275.35	1.89
KG-driven Stable diffusion	29.28	255.90	1.86

Higher CLIP Score = Better Text-Image Alignment

Lower FID = Better Visual Realism

Higher IS = Better Quality & Diversity

CLIP Score

- ✓ Our approach shows a modest improvement over the fine-tuned model, but lower than base model.
- ✓ Knowledge graph-driven text triple prompt are structurally concise but semantically sparse.

→ The original pretrained CLIP model was not trained on structured triples, it fails to fully understand input triples, even if the image is semantically correct.

Results

Model	CLIP Score	FID Score	Inception Score
Stable diffusion	31.54	290.15	1.76
Fine-tuned Stable diffusion	28.56	275.35	1.89
KG-driven Stable diffusion	29.28	255.90	1.86

Higher CLIP Score = Better Text-Image Alignment

Lower FID = Better Visual Realism

Higher IS = Better Quality & Diversity

FID Score

Our approach achieves a notably lower FID, indicating a closer alignment with the distribution of real images and reflecting superior visual realism.

Results

Model	CLIP Score	FID Score	Inception Score
Stable diffusion	31.54	290.15	1.76
Fine-tuned Stable diffusion	28.56	275.35	1.89
KG-driven Stable diffusion	29.28	255.90	1.86

Higher CLIP Score = Better Text-Image Alignment

Lower FID = Better Visual Realism

Higher IS = Better Quality & Diversity

Inception Score

Our approach attains a higher Inception Score, demonstrating improved image quality and greater intra-class diversity.
Both fine-tuned model and our approach achieve better IS than the original Stable Diffusion v1.4 baseline → Fine-tuning process may lead to better results.

Results

Input prompt:

“A navy blue jacket with red straight-point collar and blue belted waist”

(a)

Stable diffusion	Navy blue jacket ✓	Red straight-point collar ✗	Blue belted waist ✗
Fine-tuned Stable diffusion	Navy blue jacket ✓	Red straight-point collar ✗	Blue belted waist ✗
KG-driven Stable diffusion	Navy blue jacket ✓	Red straight-point collar △	Blue belted waist ✓

Results

Input prompt:

"a blue shirt with brown collar"


(b)

Stable diffusion	Blue shirt	Brown collar
Fine-tuned Stable diffusion	Blue shirt	Brown collar
KG-driven Stable diffusion	Blue shirt	Brown collar

Results

Input prompt:

"Black cotton shirt with chest patch pocket, round neck, and featuring logo patch at the chest,"

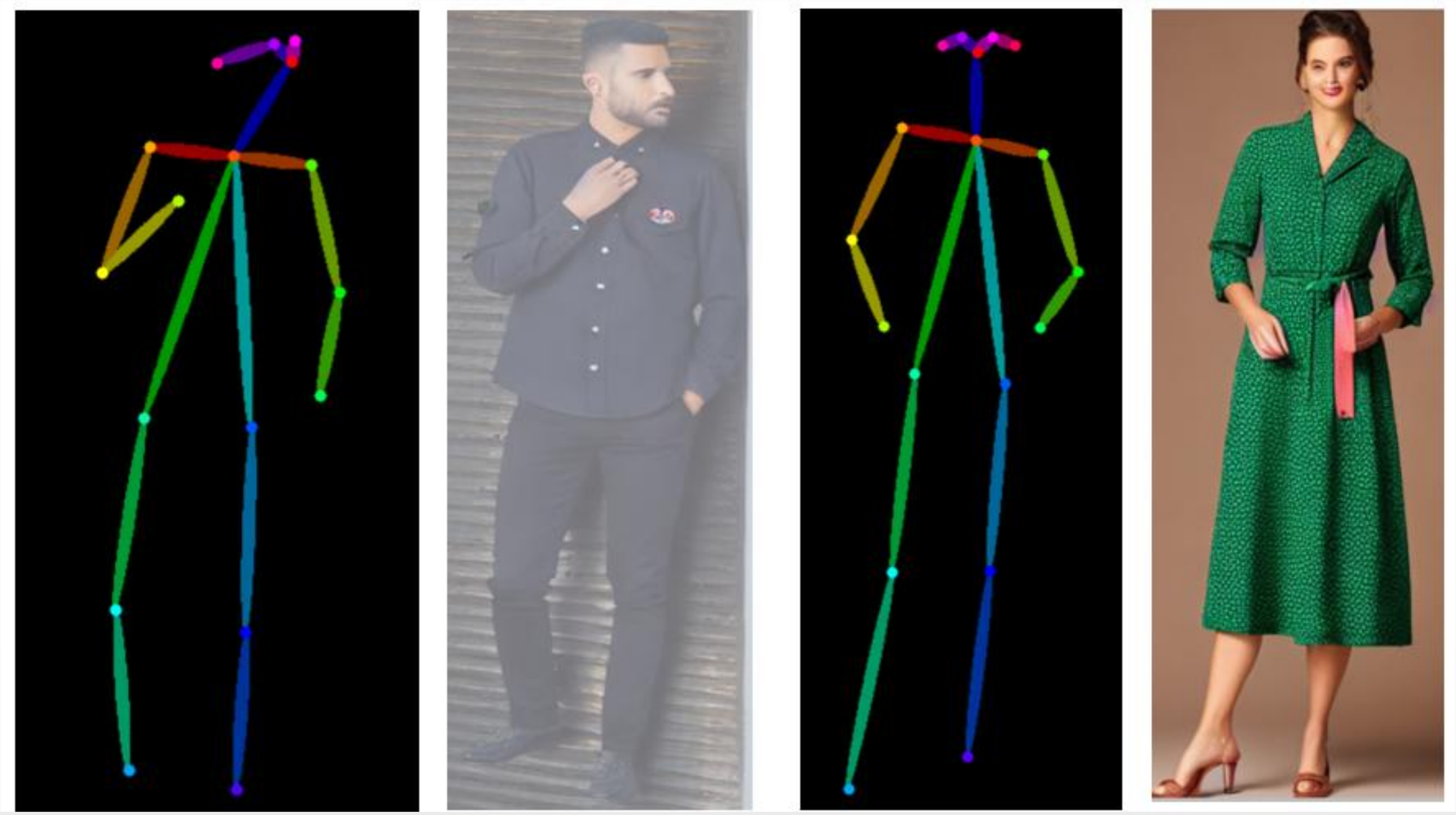
Stable diffusion		Black cotton shirt ✓	Chest patch pocket ✗	Round neck ✗	Logo patch ✓
Fine-tuned Stable diffusion		Black cotton shirt ✓	Chest patch pocket △	Round neck ✓	Logo patch ✓
KG-driven Stable diffusion		Black cotton shirt ✓	Chest patch pocket ✓	Round neck ✓	Logo patch ✓

(c)

Pose-controlled garment synthesis

Our approach combines ***OpenPose*** with knowledge graphs for precise pose and attribute alignment.

By using ***OpenPose***, it can accurately replicate the exact body pose of the input image.



Conclusion

Key Achievements

Knowledge graph-driven framework enhances semantic fidelity and controllability.

Superior performance in visual realism and text-image alignment.

Limitations

Sensitivity to noisy text inputs.

Challenges with long, ambiguous prompts.

Thank you