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

Miao-Yin Chen
Department of Data Science
Soochow University
Taipei, Taiwan
miaoyin0923@gmail.com

Shu-Han Chuang
Department of Data Science
Soochow University
Taipei, Taiwan
shuhan0819@gmail.com

***Abstract*— Recent advancements in generative modeling have substantially expanded creative possibilities within AI-assisted garment synthesis. Despite these developments, achieving fine-grained control over visual outputs using only natural language prompts remains challenging, particularly in spatial consistency and precise attribute manipulation. In this study, we propose a framework for knowledge graph-driven image synthesis that leverages structured semantic triples—comprising a head entity, relation, and tail entity. By fine-tuning a pre-trained Stable Diffusion model with these semantic structures, we guide the generation process to achieve improved cross-modal alignment and consistent structural conditioning. Our model is trained on a curated subset of the H&M Personalized Fashion Recommendations dataset, containing approximately 100,000 product images paired with detailed textual descriptions. The experimental results demonstrate that our approach achieves superior image-text alignment and enhanced visual fidelity, as evidenced by evaluating Fréchet Inception Distance (FID), Inception Score (IS), CLIP Score, and human evaluation. By combining structured semantic knowledge graph with state-of-the-art generative modeling, our method advances the development of interpretable and controllable systems for next-generation digital fashion applications.**

Keywords—Diffusion-Based Garment Synthesis, Knowledge Graph, openpose, Cross-Modal Semantic Alignment

I. INTRODUCTION

Traditional customized garment synthesis remains prohibitively complex and resource-intensive, often rendering it inaccessible to the broader consumer market due to high production costs and limited scalability. However, recent advancements in image generation and synthesis—driven by the emergence of powerful generative models—have significantly transformed this landscape. AI-assisted design tools are enabling a new paradigm of human where designers can interact with generative systems through intuitive modalities such as natural language prompts or hand-drawn sketches to produce virtual prototypes interactively and in real time [1]. This convergence of generative technology and user-centered design is not only reshaping consumer expectations but also redefining creative workflows within the fashion industry, paving the way for more accessible, scalable, and personalized garment design solutions.

Historically, generative approaches to garment synthesis have been predominantly driven by Generative Adversarial Networks (GANs), which yielded significant progress in tasks such as attribute disentanglement and human figure generation [2-6]. Despite these advancements, GAN-based methods often suffer from critical limitations, including unstable training dynamics, mode collapse, and limited flexibility in handling explicit structural constraints. These challenges have catalyzed a paradigm shift toward diffusion-based generative models, particularly Denoising Diffusion Probabilistic Models (DDPMs), which offer improved training stability, fidelity, and controllability in image synthesis tasks [7-9]. Nevertheless, two recurrent challenges persist in fashion-oriented image generation: semantic drift and incomplete attribute mapping. Semantic drift refers to deviations between the intended textual description and the generated visual output, often resulting in attribute confusion—where visual features such as color, pattern, or style are inaccurately rendered or incorrectly assigned to garment regions. Similarly, regional inconsistencies occur when irrelevant or unintended areas of the image are modified, detracting from the overall structural coherence. These issues typically stem from inadequate modeling of the linguistic–visual alignment, resulting in weak semantic compositionality and reduced interpretability of complex design specifications. While state-of-the-art diffusion models excel at generating high-resolution, aesthetically pleasing imagery, they often struggle to preserve fine-grained visual details that are critical in the fashion domain. Furthermore, exclusive reliance on textual prompts impose inherent limitations in specificity, precision, and expressiveness, thereby constraining the model's performance in addressing the nuanced requirements of design-centric tasks.

Despite the substantial progress achieved by recent generative models, several critical challenges continue to hinder the effectiveness of AI-driven fashion synthesis. Among the most prominent issues is cross-modal semantic misalignment, which results in discrepancies between generated images and their corresponding textual descriptions. Furthermore, limited semantic compositionality often leads to fragmented or incomplete visual representations, reducing the coherence of complex garment features. Additional challenges include garment part leakage, where unintended or extraneous garment components appear erroneously, and attribute confusion, wherein visual properties such as color, texture, or pattern are inaccurately assigned [10,11]. Moreover, regional inconsistencies—manifesting as misplaced or distorted garment regions—disrupt the overall visual fidelity and undermine the structural integrity of the synthesized output [10,11]. Collectively, these limitations compromise both the semantic fidelity and reliability of generative models in fashion contexts, often yielding outputs that deviate from the intended design semantics expressed in the input prompts. To address these issues, prior research has proposed

consistency loss functions that leverage attention-based masking to enforce pixel-level stability in regions deemed irrelevant to the core semantic content [11]. In response to these persistent challenges, we propose a diffusion-based framework augmented with knowledge graph-driven semantics for cross-modal garment synthesis. Knowledge graph-based semantic bundling mechanism may improve both the interpretability and controllability of the generation process which enhances alignment between linguistic descriptions and visual structures.

II. METHOD

A. Sytsem architectrue and workflow

Our proposed framework synergistically integrates high-level semantic reasoning with low-level structural control by leveraging knowledge graph-augmented text prompts. At its core is a pre-trained diffusion model, fine-tuned via a parameter-efficient paradigm that minimizes computational overhead while improving adaptability across downstream tasks. A schematic overview of the architecture is illustrated in Fig. 1. The generation pipeline comprises four primary stages: semantic text encoding, latent space construction, iterative denoising, and final image decoding. During the semantic encoding phase, we introduce a Text2Graph language module, which transforms natural language prompts into structured semantic triples [12, 13]. These triples encapsulate fine-grained attribute–object relationships, allowing the model to effectively disentangle and represent the compositional semantics of complex fashion descriptions. The generated triples are subsequently processed by a shared text encoder, producing dense, contextually enriched latent embeddings. These latent representations are refined through the iterative denoising process inherent to diffusion-based generative models. However, the conventional U-Net backbone employed in diffusion architectures often struggles to preserve detailed structural information—an essential requirement in fashion image synthesis. To address this limitation, we incorporate ControlNet as an architectural extension that conditions the U-Net on auxiliary structural inputs, such as pose estimations, edge maps, or designer-provided sketches. This structural conditioning improves spatial fidelity, ensuring that generated outputs adhere to the prescribed geometric layout while maintaining semantic alignment with the input prompts. By uniting structured semantic representation, precise structural conditioning, and interactive design control, our framework advances the capabilities of fashion-oriented generative modeling, offering a controllable, semantically-aware solution for garment synthesis and virtual design applications.

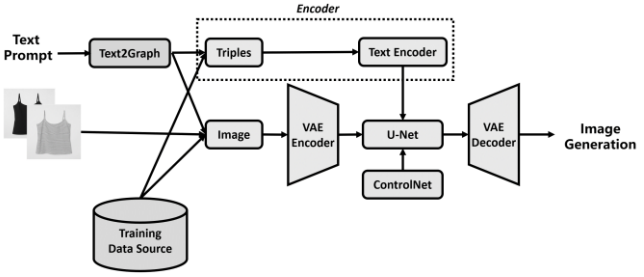

Fig. 1. The workflow of our proposed approach

B. Knowledge Graphs and Semantic Modeling

knowledge graphs have emerged as a powerful tool for enhancing semantic understanding in vision–language tasks. A knowledge graph represents entities as nodes and their relationships—such as has-color, has-feature, or made-of—as labeled edges, capturing structured, contextualized semantics that transcend flat textual descriptions. This graph-based formalism enables generative models to reason over interconnected concepts, resolve ambiguities, and interpret complex attribute compositions with greater precision. Within the context of large language models (LLMs), the process of Text2Graph refers to the transformation of unstructured natural language into structured representations, such as semantic or knowledge graphs. This involves extracting semantic triples of the form (head entity, relation, tail entity). For instance, a descriptive prompt like "*a navy blue jacket with a red straight-point collar and a blue belted waist*" can be parsed into structured triples such as (*jacket, has-color, navy blue*), (*jacket, has-feature, collar*), (*collar, has-style, straight-point*), (*collar, has-color, red*), (*jacket, has-feature, waist*),and (*waist, has-color, blue*). These triples offer semantic control, supporting accurate attribute-to-region alignment during image synthesis. Then, the process begins with semantic parsing of the input triples using a pre-trained vision–language encoder, such as Contrastive Language–Image Pre-training (CLIP) [14]. CLIP plays a pivotal role in bridging the semantic gap between language and vision by embedding both modalities into a shared latent space. These semantically enriched latent representations serve as guidance signals during the generation process.

C. Diffusion-Based Garment Synthesis

Stable Diffusion represents a significant advancement in generative modeling, particularly in the domain of text-to-image synthesis. Its primary objective is to translate natural language prompts into high-resolution, photorealistic visual outputs with high semantic fidelity. Stable Diffusion adopts a two-phase generative framework comprising a forward and a reverse process. During the forward diffusion phase, structured image data is progressively corrupted by the addition of Gaussian noise, effectively transforming it into a noise distribution. In the reverse denoising phase, a neural network—typically a U-Net architecture—is trained to iteratively reconstruct the original image distribution by learning to remove noise at each time step. This iterative refinement process enables the model to capture fine-grained visual details, including subtle textures and nuanced structural elements, making it particularly well-suited for high-fidelity synthesis tasks such as garment generation. The detailed workflow of the diffusion-based garment synthesis pipeline is illustrated in Fig. 2, highlighting the stages of semantic conditioning, noise injection, and progressive image reconstruction.

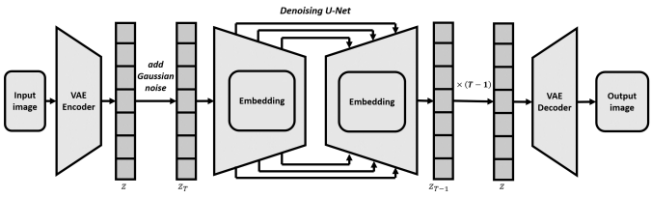

Fig. 2. The workflow of diffusion-based garment synthesis

The image generation process in Stable Diffusion begins with the Variational Autoencoder (VAE) encoder, which

compresses the input image into a lower-dimensional latent representation z. This step significantly reduces computational complexity while preserving essential visual features such as shape, texture, and color. To simulate the forward diffusion process, Gaussian noise is progressively added to the latent representation, producing a noisy version z_T that serves as the input for the reverse denoising phase [15]. At the core of the denoising process lies the U-Net architecture, a symmetric encoder–decoder network equipped with embedding layers and skip connections [16]. These skip connections are crucial for maintaining spatial coherence and contextual consistency across different resolution scales, allowing the network to recover fine-grained details while maintaining global structural alignment. Through a series of iterative refinement steps, the U-Net gradually removes noise from the latent representation, reconstructing a cleaner version that approximates the original latent distribution. Following denoising, the refined latent code is passed through the VAE decoder, which reconstructs the final image from the latent space. The VAE framework ensures that the latent space is structured and semantically meaningful, allowing the model to generate high-fidelity outputs that preserve salient visual characteristics. Additionally, the diffusion process can be conditioned on auxiliary inputs such as textual prompts, pose skeletons, or hand-drawn sketches, thereby enabling semantically guided image synthesis. This conditioning mechanism enhances the model's ability to generate visually coherent outputs that align closely with user intent.

D. Fintuned Diffusion-Based Garment Synthesis

To facilitate effective domain adaptation for fashion-specific generative tasks, we fine-tune a pre-trained Stable Diffusion model using Low-Rank Adaptation (LoRA)—a parameter-efficient fine-tuning technique that introduces task-specific modifications while largely preserving the original model weights [17,18]. This approach substantially reduces training overhead and enhances generalization across a broad range of fashion synthesis applications. In the subsequent stage of the generation pipeline, ControlNet is jointly trained alongside the LoRA-adapted diffusion model to incorporate explicit structural conditioning [19]. This conditioning imposes geometric constraints derived from auxiliary inputs, such as pose estimations or designer-provided sketches, thereby ensuring structural fidelity without compromising semantic coherence. The resulting hybrid architecture enables controllable, high-fidelity garment image synthesis that faithfully integrates both textual descriptions and structural guidance. By combining LoRA-based fine-tuning with ControlNet's geometric constraint enforcement, the framework advances automated fashion design capabilities, significantly enhancing controllability, visual realism, and creative flexibility. This supports a wide range of applications including interactive design ideation and virtual prototyping workflows.

III. Experiments and Results

A. Data source

We utilize a curated subset of the H&M Personalized Fashion Recommendations dataset from Kaggle competition (Available: https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations). This dataset serves as a valuable resource for advancing research in fashion informatics, particularly in the integration of structured semantic metadata with visual garment representations. It contains detailed metadata for approximately 105,000 unique fashion items, covering a diverse range of garment categories, styles, and attributes. This rich metadata is essential for modeling fine-grained semantic relationships within the fashion domain, enabling applications such as knowledge graph construction and the generation of semantically enriched text-to-image prompts. Moreover, each fashion item is accompanied by a high-resolution, front-facing product image, which functions as a reliable visual ground truth for training and evaluating multimodal vision-language models. These images facilitate the extraction of salient visual features critical for various tasks, including fashion retrieval, image synthesis, and attribute prediction.

B. The paramters setting in finetuned Stable Diffusion

The proposed approach leverages Stable Diffusion v1.4 as the foundational generative model. This pre-trained diffusion network is fine-tuned on a curated fashion image dataset to facilitate localized, text-driven garment editing with enhanced semantic alignment and visual fidelity, thereby better capturing domain-specific characteristics. Fine-tuning is performed over 200,000 iterations using the *AdamW* optimizer, with a conservative learning rate of 1×10^{-6} to preserve pre-trained knowledge while enabling effective domain adaptation. To address GPU memory limitations and improve training stability, gradient accumulation is employed over four steps, resulting in an effective batch size of one. All experiments are conducted on a single *NVIDIA* A6000 GPU, with both input and output images uniformly resized to a resolution of 512 × 512 pixels.

C. The Preformance Evaluation of our approach

To evaluate the effectiveness of the proposed framework, we perform comprehensive comparisons against state-of-the-art methods using a suite of quantitative metrics. Our evaluation is based on three widely adopted benchmarks: Fréchet Inception Distance (FID) [20], Inception Score (IS) [21], and CLIP Score [22]. FID quantifies the similarity between the distributions of real and generated image features, which are extracted using a pre-trained Inception v3 network. A lower FID score indicates that the synthesized images closely match the statistical properties of real images, reflecting superior visual fidelity and realism. The Inception Score assesses both the visual quality and diversity of generated outputs by analyzing the predicted class distributions from the Inception v3 model. Higher IS values suggest that the images are not only visually convincing but also exhibit sufficient category diversity. The CLIP Score measures semantic alignment between the generated image and the input textual prompt. This is computed as the cosine similarity between multimodal embeddings obtained from the CLIP model, with higher scores indicating stronger consistency between visual content and linguistic description. Together, these metrics provide a comprehensive assessment of model performance, capturing both low-level image fidelity and high-level semantic accuracy in the generated fashion images.

As summarized in Table I, the proposed framework consistently outperforms baseline models across all evaluation metrics. Compared to the original Stable Diffusion v1.4 framework, its fine-tuned variant, and our knowledge graph-driven Stable Diffusion model, our approach achieves a notably lower FID score of 25.90, indicating a closer alignment with the distribution of real images and reflecting superior visual realism. Additionally, it attains a higher

Inception Score (IS) of 1.86, demonstrating improved image quality and greater intra-class diversity. Notably, both the fine-tuned model and our approach also achieve better IS than the original Stable Diffusion v1.4 baseline. Fine-tuning process may lead to better composition or object generation. While the CLIP Score of our approach is 29.28 that 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. Due to the original CLIP model was not trained on structured triples and expects context-rich and descriptive inputs, it fails to fully understand or align structured input triples with image content, even if the image is semantically correct.

TABLE I. THE PREFORMANCE COMPARISIONS BETWEEN OUR PROPOSED MODEL AND STAT-OF-THE-ART MODELS

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

D. The human-centered evaluation of our approach

Beyond automated metrics, a human-centered evaluation was conducted to assess the practical performance of the proposed framework in garment synthesis tasks. To investigate the effectiveness of the structural semantic consensus guidance, we perform an ablation study involving several model variants. Fig. 3 highlights two critical challenges in garment synthesis: attribute confusion and regional inconsistency, and demonstrates how our framework effectively mitigates these issues. In Fig. 3(a), the problem of regional inconsistency is illustrated using the input prompt: "*a navy blue jacket with red straight-point collar and blue belted waist*". Both the baseline and fine-tuned models incorrectly alter the color of the blue belt to red, indicating a failure to maintain the integrity of visual regions unrelated to the specified manipulation. In contrast, our proposed model achieves precise generation of fine-grained visual details by incorporating structural semantic consensus guidance. The knowledge graph-driven approach fosters robust spatial alignment between garment parts and their corresponding textual attributes, enforcing accurate part-level semantic correspondence and thereby improving visual coherence and fidelity in garment synthesis.

An example of attribute confusion is illustrated in Fig. 3(b), where visual attributes are either incorrectly assigned or omitted. Given the detailed prompt, "*a blue shirt with brown collar*," the baseline model erroneously associates the colors "blue" and "brown" with incorrect garment components. While both the fine-tuned model and our proposed framework demonstrate marked improvements, our approach significantly reduces attribute confusion. These models enhance fine-grained control and semantic accuracy, resulting in more faithful alignment between textual descriptions and visual outputs.

Fig. 3(c) demonstrates the full capabilities of our proposed framework in generating realistic and semantically faithful fashion images. Given a detailed prompt such as "*Black cotton shirt with chest patch pocket, round neck, and featuring logo patch at the chest*," baseline models capture the garment's overall shape, color, and logo but fail to accurately render subtle yet critical elements like the neckline style and patch pocket. Although the fine-tuned model reproduces these features more accurately, the design of the patch pocket appears distorted and incomplete. In contrast, our knowledge graph-driven approach successfully replicates all specified details with a higher degree of semantic fidelity, visual coherence, and realism, thereby producing outputs that are both visually convincing and faithful to the input description.

Fig. 3. Results of our proposed model on the garment synthesis task for several challenging examples that require precise generation of fine-grained visual details. From top to bottom: the first row shows outputs from the base model, the second row shows results from the fine-tuned model, and the third row presents outputs from our proposed model.

By structuring semantic attributes within a knowledge graph, the proposed method explicitly encodes relationships between entities, thereby addressing common challenges in generative modeling, including attribute ambiguity and spatial inconsistency. This graph-based representation enables fine-grained control during image synthesis by preserving the compositional integrity of attribute–object associations. Furthermore, the semantically enriched graph facilitates stronger cross-modal alignment between textual and visual modalities, enhancing the model's capacity to interpret and generate complex, design-oriented content. Consequently, this approach substantially advances generative modeling performance in domains demanding high semantic fidelity and structural precision, such as digital fashion.

E. The results of Pose-controlled garment synthesis

As illustrated in Fig. 4, the first column displays the input skeleton pose extracted using *OpenPose*, while the third column visualizes the corresponding human body keypoints with color-coded markers indicating joint connections and limb orientations. The second and fourth columns present paired examples, including the fashion images synthesized by our knowledge graph-guided, fine-tuned Stable Diffusion model, which is conditioned jointly on the input pose and the accompanying textual garment description. In this example, the model successfully synthesizes the garment, accurately reflecting both the prescribed body pose and the detailed semantic attributes specified in the input. The integration of *OpenPose*-based pose control within the knowledge graph-augmented diffusion framework markedly enhances

controllability, structural precision, and semantic fidelity of the generated outputs, thereby advancing the state of the art in pose-aware, text-guided garment generation.

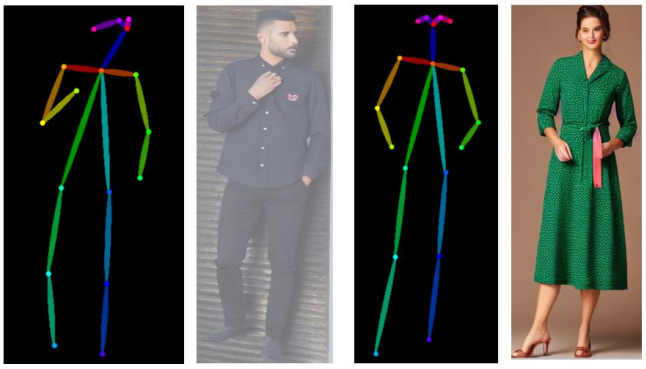

Fig. 4. Pose-controlled garment synthesis results using our proposed knowledge graph-driven fine-tuned Stable Diffusion model.

IV. Conclusion

Our work explores advanced methodologies for garment synthesis by leveraging a knowledge graph-driven, fine-tuned Stable Diffusion model. The core objective is to enhance controllability and semantic alignment in image generation through the structured transformation of text prompts into knowledge graphs. This approach has proven effective in encoding domain-specific semantic and relational knowledge, where entities are represented as nodes and their relationships as edges. Such structured representations facilitate greater semantic fidelity and fine-grained control throughout the generative process. Our experimental results demonstrate the promise of our method in producing visually coherent and semantically grounded garment images. However, several limitations persist. One notable challenge is the model's sensitivity to noisy or imprecise textual inputs, which can impair the alignment between descriptive language and generated visual features. Moreover, the inherent ambiguity and implicit knowledge embedded in natural language often complicate the accurate extraction of structured graphs from text. An additional complexity arises when processing longer textual inputs, which necessitates effective strategies for segmenting and aggregating semantic content without compromising coherence. To address these limitations, future work will explore advanced graph embedding techniques aimed at improving the robustness of the model against textual variability and enhancing its ability to generalize across diverse linguistic and structural patterns.

References

[1] A. Baldrati et al., "Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing," arXiv preprint arXiv:2304.02051, 2023. [Online]. Available: https://arxiv.org/abs/2304.02051

[2] T. Xu et al., "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," arXiv preprint arXiv:1711.10485, 2017. [Online]. Available: https://arxiv.org/abs/1711.10485

[3] M. Tao et al., "DF-GAN: A simple and effective baseline for text-to-image synthesis," arXiv preprint arXiv:2008.05865, 2022.

[4] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 12174–12182.

[5] H. Zhang et al., "Cross-modal contrastive learning for text-to-image generation," arXiv preprint arXiv:2101.04702, 2022. [Online]. Available: https://arxiv.org/abs/2101.04702

[6] M. Zhu et al., "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," arXiv preprint arXiv:1904.01310, 2019. [Online]. Available: https://arxiv.org/abs/1904.01310

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," arXiv preprint arXiv:2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[8] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," arXiv preprint arXiv:2105.05233, 2021. [Online]. Available: https://arxiv.org/abs/2105.05233

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10674–10685.

[10] S. Zhu et al., "Be your own Prada: Fashion synthesis with structural coherence," arXiv preprint arXiv:1710.07346, 2017.

[11] X. Zhang et al., "DiffCloth: Diffusion Based Garment Synthesis and Manipulation via Structural Cross-modal Semantic Alignment," arXiv preprint arXiv:2308.11206, 2023.

[12] A. Hur, N. Janjua, and M. Ahmed, "A Survey on State-of-the-art Techniques for Knowledge Graphs Construction and Challenges ahead," in 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 2021, pp. 99-103, doi: 10.1109/AIKE52691.2021.00021.

[13] A. Hur, N. Janjua, and M. Ahmed, "Unifying context with labeled property graph: A pipeline-based system for comprehensive text representation in NLP," Expert Systems with Applications, vol. 239, p. 122269, 2024, doi: 10.1016/j.eswa.2023.122269.

[14] H.-Y. Chen, Z. Lai, H. Zhang, X. Wang, M. Eichner, K. You, M. Cao, B. Zhang, Y. Yang, and Z. Gan, "Contrastive Localized Language-Image Pre-Training," arXiv preprint arXiv:2410.02746, 2024. [Online]. Available: https://arxiv.org/abs/2410.02746

[15] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," Foundations and Trends® in Machine Learning, vol. 12, no. 4, pp. 307–392, Nov. 2019

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS, vol. 9351, Springer, 2015, pp. 234–241

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[18] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," arXiv preprint arXiv:2208.12242, 2022. [Online]. Available: https://arxiv.org/abs/2208.12242

[19] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 3813–3824

[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," Advances in Neural Information Processing Systems, vol. 30, 2017.

[21] S. Barratt and R. Sharma, "A note on the inception score," arXiv preprint arXiv:1801.01973, 2018. [Online]. Available: https://arxiv.org/abs/1801.01973

[22] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," arXiv preprint arXiv:2104.08718, 2021. [Online]. Available: https://arxiv.org/abs/2104.08718