IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

# Multi-Subject Text-to-Video Generation Using AI with Subject and Motion Control

**Mr. Hase Chaitanya Dnyaneshwar[1], Mr. Barsale Sham Bhagwan[2], Mr. Mulay Varun Vinayak[3], Ms. Pathan Sabiha Riyajkhan[4], Prof. Muneshwar R. N.[5]**
Department of Information Technology
Amrutvahini College of Engineering, Sangamner, India

**Abstract:** *The field of Artificial Intelligence has enabled automatic generation of multimedia content from textual descriptions. Among these advancements, text-to-video generation has gained significant attention due to its ability to transform written input into dynamic visual sequences. However, most existing systems are limited to simple scenarios involving a single subject and often fail when multiple subjects are involved in the same scene. Issues such as identity confusion, inconsistent motion, and lack of temporal stability reduce the quality of generated videos.*

*This paper presents a multi-subject text-to-video generation system that incorporates subject-aware representation and motion control mechanisms. The system processes user-defined textual prompts and generates coherent video sequences using a diffusion-based deep learning model. Each subject is handled independently to preserve identity, while motion parameters are applied to ensure smooth transitions across frames. A user-friendly interface is developed to allow real-time interaction and visualization.*

*The proposed approach improves visual consistency, maintains subject separation, and produces more realistic motion compared to conventional methods. The system can be applied in areas such as content creation, animation, education, and digital media production.*

**Keywords:** Text-to-Video Generation, Artificial Intelligence, Deep Learning, Diffusion Models, Motion Control, Multi-Subject Systems.

## I. INTRODUCTION

Artificial Intelligence has transformed the way digital content is created by enabling machines to interpret human language and generate corresponding visual outputs. One such application is text-to-video generation, where a system converts textual descriptions into short video clips. This technology reduces the need for manual video production and opens new opportunities in automated content creation.

Although recent models have demonstrated promising results, they are generally designed for simple scenes involving a single object or subject. When multiple subjects are present, these systems often fail to maintain consistency in appearance, positioning, and interaction. This leads to visual artifacts such as blending of subjects and irregular motion patterns.

To overcome these challenges, this work focuses on developing a system capable of generating videos with multiple subjects while preserving their identities and controlling their motion. The proposed approach aims to produce stable and visually coherent outputs suitable for real-world applications.

## II. LITERATURE SURVEY / RELATED WORK

This section reviews existing research work related to text-to-video generation and multi-subject modeling techniques. Various approaches based on diffusion models, transformers, and deep learning architectures have been studied to understand their strengths and limitations. The comparison of these methods helps identify research gaps and motivates the development of the proposed system. A summary of relevant studies is presented in Table 1.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

**Table 1. Comparative Summary of Literature Review**

| Sr. No. | Paper | Source | Technique | Observations |
|---|---|---|---|---|
| 1 | Hong Chen et al., Video Dreamer[1] | IEEE Transactions on Multimedia, 2025 | Multi- subject text-to-video diffusion with disentangled fine tuning | Improved subject separation and visual consistency, but requires high computational resources |
| 2 | J. Xing et al., Make-your-video [2] | IEEE TVCG, 2025 | Textual and structural guided video diffusion | Enabled customized video generation with structural constraints; identity preservation |
| 3 | U. Singer et al., Make-A-Video [3] | ICLR, 2023 | Text-conditioned diffusion with temporal attention | Successfully generated short videos from text prompts, though multi-subject consistency and long-term temporal stability were limited. |
| 4 | T. Romb --ch et al., Latent Diffusion Models[4] | arXiv, 2022 | Latent space diffusion for high-resolution synthesis | Reduced computational cost while maintaining image quality; lacks direct temporal mode--ling for video generation |
| 5 | S. Esser et al., Taming Trans formers[5] | CVPR,2021 | Transformer-based image synthesis with learned code books | Provided a scalable frame- work for high-resolution image generation, forming the basis for later video diffusion models. |

**Comparative Analysis**

The studies reviewed demonstrate significant progress in text-to-video generation but reveal several key limitations:

1. Limited Multi-Subject Handling: Most existing systems are designed for single-subject scenarios and struggle to manage multiple subjects simultaneously.
2. Identity Preservation Issues: Models often fail to maintain consistent identity of subjects across frames, leading to visual confusion and blending.
3. Inadequate Motion Control: Current approaches lack precise mechanisms to control subject motion, resulting in unrealistic or unstable movements.
4. Temporal Inconsistency: Generated videos frequently suffer from flickering and lack of smooth transitions between frames.

From a methodological view, diffusion and transformer-based models produce high-quality visuals but require high computational resources and lack fine-grained control over subject interactions. While recent approaches improve visual quality, they still fail to achieve a balance between identity preservation, motion control, and temporal stability.

**Research Gap and Problem Identification**

Despite advancements in text-to-video generation, several gaps remain.

1. Existing systems lack efficient mechanisms to handle multiple subjects while preserving their individual identities.
2. There is limited focus on ensuring temporal consistency across long video sequences.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

3. High computational complexity restricts real-time implementation and scalability.

Hence, the identified problem statement is:

To design and develop an AI-based multi-subject text-to-video generation system that ensures identity preservation, provides effective motion control, maintains temporal consistency, and generates high-quality videos efficiently for real-world applications.

## III. PROBLEM STATEMENT

Text-to-video generation using Artificial Intelligence has emerged as a powerful tool for automated visual content creation. However, generating videos from textual descriptions becomes significantly challenging when multiple subjects are involved in a single scene. Issues such as identity confusion, inconsistent motion, and lack of temporal stability frequently occur in generated videos.

Existing text-to-video generation systems are limited due to:

- Dependence on models that are primarily optimized for single-subject video generation.
- Inability to preserve distinct identities of multiple subjects across frames.
- Lack of effective motion control mechanisms, leading to unrealistic or unstable movements.
- Temporal inconsistency, resulting in flickering and discontinuity in generated video sequences.
- High computational requirements, making real-time implementation difficult.
- Limited user control over subject appearance, positioning, and interactions.

As a result, current systems fail to produce realistic and coherent multi-subject videos, limiting their applicability in areas such as digital content creation, animation, and storytelling.

Therefore, there is a need to design and develop an intelligent, scalable, and efficient text-to-video generation system that can handle multiple subjects while ensuring identity preservation, controlled motion, temporal consistency, and improved user flexibility for real-world applications.

## IV. PROPOSED SYSTEM

The proposed system is an AI-Based Multi-Subject Text-to-Video Generation System that integrates diffusion-based deep learning models with subject-aware representation and motion control mechanisms within a Python-based framework and interactive web interface.

**Key Features of the Proposed System**

**1. Multi-Subject Text-to-Video Generation**

Architecture

The system uses a diffusion-based model to generate video frames from textual input.

Subject-aware embeddings are used to maintain distinct identity for each subject.

The architecture ensures better visual consistency compared to traditional models.

**2. Subject Identity Preservation**

Each subject in the input prompt is processed independently.

Unique embeddings are assigned to avoid identity mixing.

Ensures consistent appearance of subjects across all frames

**3. Motion Control Mechanism**

Motion parameters are applied to each subject individually.

Temporal attention mechanisms ensure smooth motion transitions.

Reduces unnatural or unstable movements in generated videos.

## IV. SYSTEM ARCHITECTURE

The proposed Multi-Subject Text-to-Video Generation System follows a layered and modular architecture to ensure efficient processing, scalability, and high-quality video generation with subject and motion control.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

**Architecture Overview**

The system consists of four major layers:

**1. Presentation Layer (User Interface)**

- Developed using Gradio-based web interface
- Provides interactive platform for users to input text prompts and configure parameters
- Allows:
  - Entering scene description (text input)
  - Defining multiple subjects and their attributes
  - Setting motion parameters
  - Viewing generated video output

**2. Application Layer (Processing Layer)**

- Implemented using Python
- Acts as a bridge between user interface and deep learning model
- Handles:
- Input validation and preprocessing
- Prompt encoding and tokenization
- Sending processed data to generation model

**3. Model Layer (Deep Learning Layer)**

- Implemented using diffusion-based model (WAN 2.1)
- Core components:

Text Encoder → Converts input text into embedding

Diffusion Model → Generates latent video frames

Motion Control Module → Controls subject movement

Temporal Attention → Maintains frame consistency

- Functions:

Multi-subject identity preservation

Frame-by-frame video generation

Motion-aware video synthesis

**4. Output Layer (Video Generation & Storage)**

Uses Variational Auto Encoder (VAE) for decoding

Stores and displays:

Generated video

Video metadata (duration, resolution)

Temporal Attention → Maintains frame consistency

Provides download option for users



DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

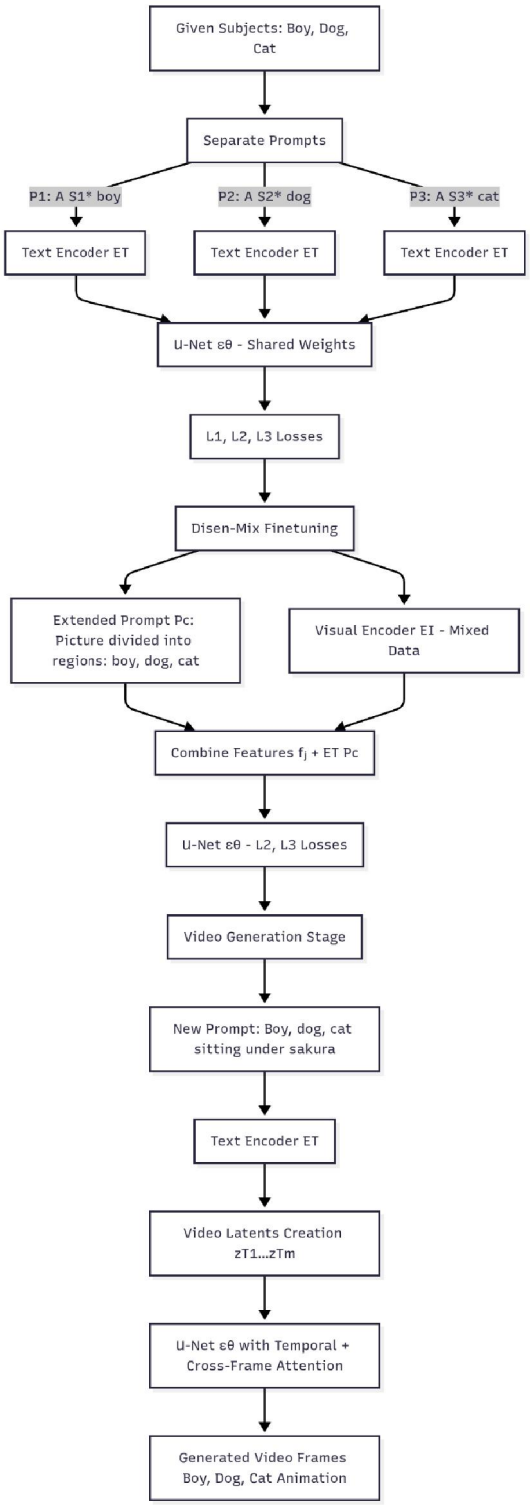



**Fig. 1 System Architecture**


DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

## VI. METHODOLOGY

The methodology follows a structured deep learning pipeline from text input to final video generation using diffusion-based models.

**Step 1: Input Collection**

Text prompt describing the scene

- Multiple subjects (e.g., boy, dog, cat)
- User-defined parameters (motion, style, duration)

**Step 2: Prompt Processing**

- Split input into subject-specific prompts
- Clean and structure textual data
- Prepare prompts for encoding

**Step 3: Text Encoding**

- Convert text prompts into embedding using Text Encoder
- Generate semantic representation for each subject
- Capture relationships between subjects

**Step 4: Feature Extraction**

- Use U-Net architecture for extracting spatial features
- Learn contextual relationships between subjects
- Apply loss functions to improve feature learning

**Step 5: Disentanglement & Feature Refinement**

- Apply Disen-Mix fine-tuning
- Separate subject-specific features
- Enhance identity preservation and clarity

**Step 6: Video Generation**

- Use diffusion model to generate latent frames
- Iteratively refine frames through de-noising
- Generate sequence of video frames

**Step 7: Temporal Consistency & Motion Control**

- Apply temporal attention mechanisms
- Ensure smooth transitions between frames
- Control motion of individual subjects

**Step 8: Frame Decoding & Video Assembly**

- Decode latent frames using Variational Auto Encoder (VAE)
- Combine frames into continuous video
- Maintain resolution and frame rate

**Step 9: Output Generation**

- Display generated video to user

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

- Provide download option
- Show output metadata (duration, quality)

### 6.1 SYSTEM WORKFLOW

The complete workflow of the proposed system is as follows:

1. User inputs text prompt describing the scene
2. System identifies multiple subjects from input
3. Prompt is split into subject-specific descriptions
4. Text encoding converts prompts into embedding
5. Feature extraction using U-Net architecture
6. Disentanglement of subject-specific features
7. Diffusion model generates latent video frames
8. Motion control applied to individual subjects
9. Temporal consistency maintained across frames
10. Frames decoded and assembled into final video
11. Output video displayed and available for download

### 6.2 MATHEMATICAL MODEL

Let:

- P = Input text prompt
- Si = Individual subject descriptions
- E = Text embedding
- X = Feature vector
- Zt = Latent representation at time t
- V = Output video
- **Feature Vector Representation:**

$$X=[E1,E2,E3,\ldots,En]$$

- **Normalization:**

$$X\_norm=(X-X\_min)/(X\_max-X\_min)$$

- **Convolution Operation**
- **(Feature Extraction):**

$$F=X\_norm*W+b$$

Where:

W = Weight matrix

b = Bias

F = Feature map

- **Residual Learning:**

$$H(x) = F(x) + x$$

- **Diffusion Process:**

$$Zt = Zt\text{-}1+ \varepsilon t$$

- **Denoising Process:**

$$Zt\text{-}1 = Zt - \varepsilon t$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

- **Softmax Function:**

p(yi) = (e_z i)/( ∑ _j▒e^zj )

- **Final Output:**

V = Decoder(Zt)

Video = {f1, f2, f3, ..., fn}

### 6.3 ALGORITHM FOR TEXT-TO-VIDEO GENERATION

**Input :** Text prompt with multiple subjects

**Output:** Generated video with motion consistency

**Step 1:** Accept input text prompt from user

**Step 2:** Identify and extract multiple subjects from the prompt

**Step 3:** Split input into subject-specific descriptions

**Step 4**: Convert text into embedding using text encoder

**Step 5:** Perform feature extraction using U-Net Architecture

**Step 6:** Apply disentanglement to separate subject features

**Step 7**: Generate latent frames using diffusion model

**Step 8:** Apply motion control to each subject

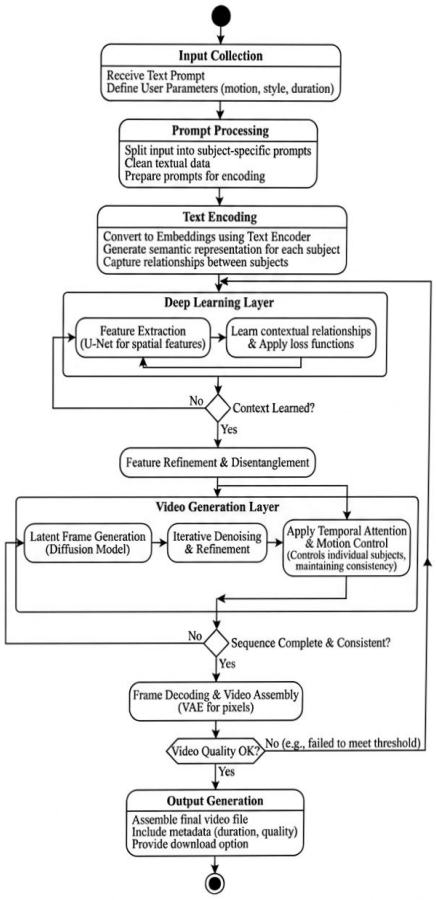
**Step 9:** Ensure temporal consistency across frames

**Step 10:** Decode latent frames into visual frames using VAE

**Step 11:** Combine frames to generate final video

**Step 12:** Display and provide output video to user

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2




## VIII. IMPLEMENTATION

The proposed Multi-Subject Text-to-Video Generation System is implemented by combining deep learning techniques with an interactive user interface to achieve efficient and high-quality video synthesis. The system is designed to support real-time interaction while maintaining scalability and performance.

The user interface is developed using Gradio, which enables users to provide textual input describing the scene and view the generated video output. The interface is simple and responsive, allowing users to define multiple subjects and control generation parameters effectively.

The processing layer is built using Python and is responsible for handling input preprocessing, prompt segmentation, and coordination between different components of the system. It ensures that the input data is properly structured before being passed to the deep learning model.

The core functionality of the system is driven by a diffusion-based deep learning model integrated with a U-Net architecture. This layer utilizes libraries such as TensorFlow / PyTorch, NumPy, and related tools for model execution. The model transforms textual embedding into latent representations and progressively generates video frames while maintaining spatial and temporal coherence.

To improve subject clarity, text encoding techniques are applied to convert input prompts into meaningful embedding. Additionally, feature disentanglement methods are used to separate and preserve the identity of multiple subjects throughout the video sequence.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

The output generation process is handled using a Variational Auto Encoder (VAE), which decodes latent representations into visual frames. These frames are then combined sequentially to form the final video output, which is displayed within the interface and can be downloaded by the user.

Supporting tools such as Python-based visualization libraries are used for testing and evaluation. The system is deployed on a Windows environment with standard hardware requirements, including an Intel i5 processor and 8GB RAM, with optional GPU support to enhance processing speed.

## VIII. RESULTS AND DISCUSSION

The implementation of the proposed multi-subject text-to-video generation system shows noticeable improvement in generating coherent and visually consistent videos compared to conventional approaches. The system was evaluated based on output quality, subject identity preservation, motion smoothness, and temporal consistency.

Experimental observations indicate that the proposed approach produces better results than basic text-to-image or single-subject video generation models discussed in the literature. The integration of subject-aware processing and diffusion-based generation improves the overall visual quality and reduces common issues such as subject mixing and distortion.

The system effectively maintains distinct identities for multiple subjects throughout the video sequence. The use of disentanglement techniques ensures that each subject retains its unique characteristics across frames. Additionally, the incorporation of temporal attention mechanisms contributes to smoother transitions and reduces flickering effects.

Another important outcome is the improved motion representation. The motion control component allows more stable and realistic movement of subjects compared to traditional models that lack explicit motion handling. This results in more natural and visually appealing video outputs.

The system also demonstrates consistent performance when handling increasingly complex prompts involving multiple subjects. It is observed that combining structured prompt processing with deep learning techniques enhances the reliability and scalability of the system.

Overall, the proposed model provides a more effective solution for multi-subject text-to-video generation by improving identity preservation, motion stability, and temporal coherence, making it suitable for practical applications such as content creation, animation, and visual storytelling.

## IX. FUTURE SCOPE

While the proposed multi-subject text-to-video generation system demonstrates effective performance, there is significant scope for further improvement and expansion.

One possible enhancement is the incorporation of deeper contextual understanding of text prompts. By leveraging advanced Natural Language Processing techniques, the system can better interpret complex descriptions, emotions, and interactions among multiple subjects, leading to more realistic and meaningful video outputs.

Another area of improvement lies in motion modeling. Future work can focus on generating more accurate and physically consistent movements, particularly in scenes involving interactions between multiple entities. Integrating advanced temporal learning methods or physics-aware constraints can enhance realism in generated videos.

The inclusion of explaining features can also be beneficial. Introducing Explainable AI mechanisms would allow users to understand how the system interprets input prompts and generates corresponding visual outputs, thereby increasing transparency and user confidence.

Further enhancements can include providing users with more control over the generation process. Features such as customization of subject appearance, scene layout, lighting, and camera movement can make the system more versatile for applications like animation, storytelling, and media production.

Scalability can be improved by deploying the system on cloud-based platforms, enabling faster processing and wider accessibility. This would allow multiple users to interact with the system simultaneously without performance degradation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

Additionally, adopting continuous learning strategies can help the model adapt to new data and evolving requirements. Exploring more advanced architectures, such as transformer-based models, may further improve generation quality and efficiency.

In the long term, the system can be developed into a comprehensive AI-driven video generation platform capable of supporting diverse applications across entertainment, education, marketing, and virtual environments.

## X. CONCLUSION

The proposed multi-subject text-to-video generation system demonstrates the effective application of deep learning techniques for automated video synthesis. By integrating diffusion-based models with subject-aware processing, the system addresses key limitations of traditional text-to-video approaches, particularly in handling multiple subjects within a single scene.

The system processes complex textual inputs and generates coherent video outputs while preserving subject identity, motion consistency, and visual quality. The use of advanced architectures such as U-Net and diffusion models improves feature representation and enables stable frame generation across time.

The interactive interface enhances usability by allowing users to provide custom prompts and obtain video outputs in a streamlined manner. This makes the system suitable for practical applications such as content creation, animation, and visual storytelling.

Experimental observations indicate that the proposed approach produces more consistent and realistic results compared to simpler or single-subject generation methods. The modular design of the system also allows future enhancements and easy integration with advanced models and platforms.

In conclusion, the project highlights the growing potential of Artificial Intelligence in creative media generation. By enabling automated and customizable video creation from textual descriptions, the system contributes to advancements in digital content production and opens new possibilities for intelligent multimedia applications.

## REFERENCES

[1] Alec Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, 2021.
[2] Robin Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models,” CVPR, 2022.
[3] William Peebles and Saining Xie, “Scalable Diffusion Models with Transformers,” ICCV, 2023.
[4] Olaf Ronneberger et al., “U-Net: Convolutional Networks for Biomedical Image Segmentation,” MICCAI, 2015.
[5] Kaiming He et al., “Deep Residual Learning for Image Recognition,” CVPR, 2016.
[6] Ashish Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems, 2017.
[7] Jonathan Ho et al., “Denoising Diffusion Probabilistic Models,” NeurIPS, 2020.
[8] Prafulla Dhariwal and Alex Nichol, “Diffusion Models Beat GANs on Image Synthesis,” NeurIPS, 2021.