Edge-Enabled Generative AI Services and AIGC Networks: Semantic and Security Perspectives

Dusit Tao Niyato

https://arxiv.org/pdf/2303.16129.pdf

“Life is not about finding yourself
Life is about creating yourself”

George Bernard Shaw

What is Artificial Intelligence-Generated Content (AIGC)

- Professionally-generated content (PGC)
  - generators are individuals or organizations with professional skills, knowledge, and experience
- User-generated content (UGC)
  - digital material generated by users including website visitors and social media users
- AIGC
  - an automated method for generating, manipulating, and modifying valuable and diverse data using AI algorithms creatively

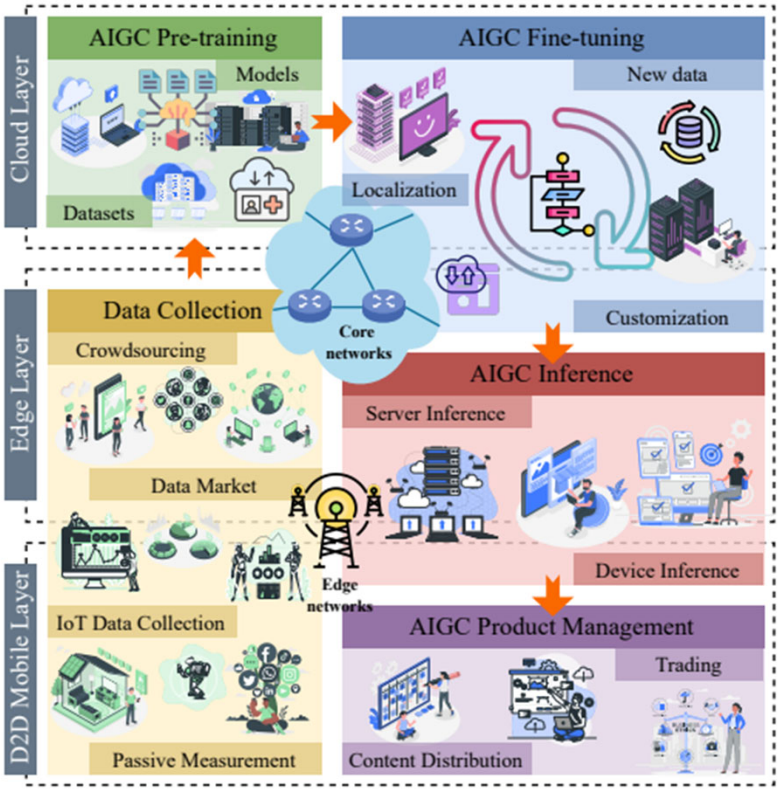


Fig. 1: The overview of mobile AIGC networks, including the cloud layer, the edge layer, and the D2D mobile layer. The lifecycle of AIGC services, including data collection, pretraining, fine-tuning, inference, and product management, is circulated among the core networks and edge networks.

AI-generated Texts

- Recent advancements in Natural Language Generation (NLG) technology have led to AI-generated text that is nearly indistinguishable from human-written text

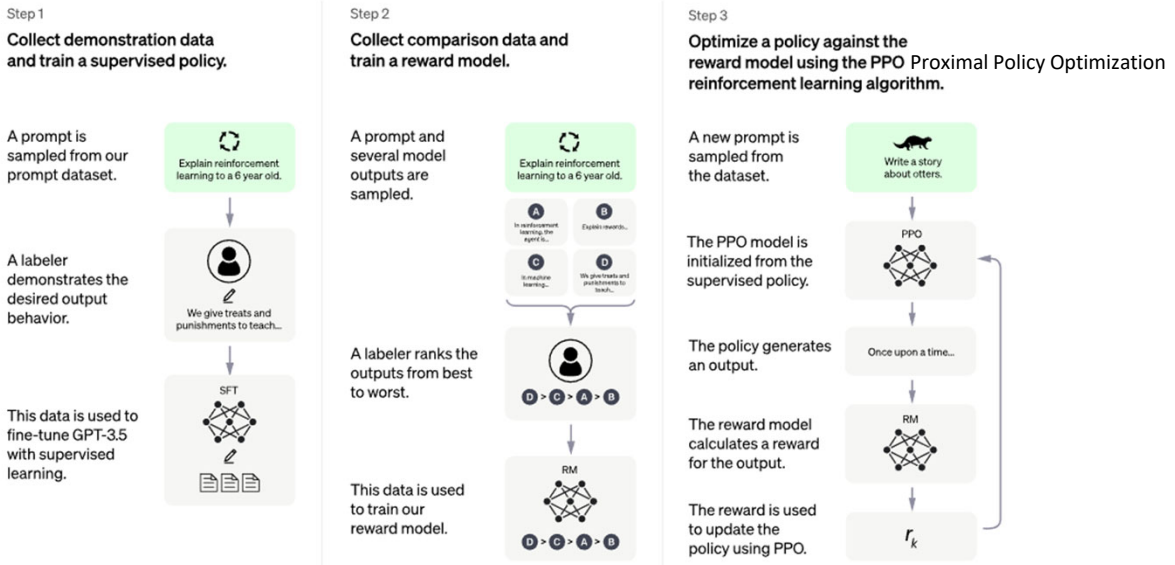
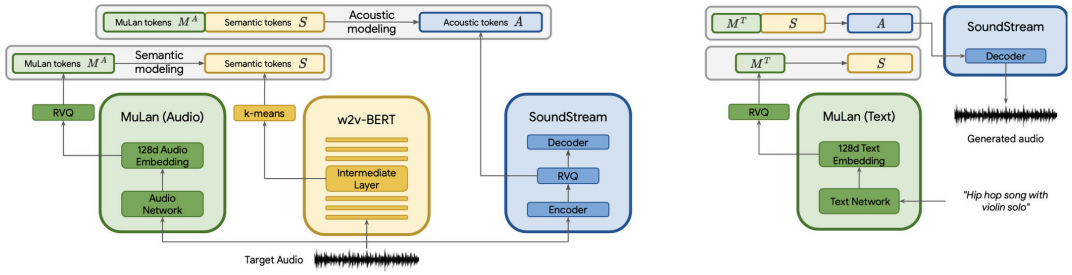


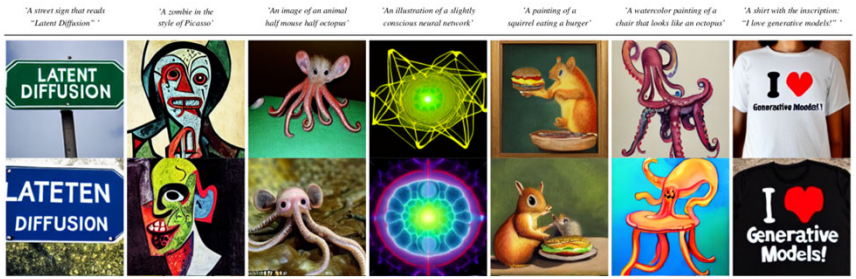
Fig. 8: The fine-tuning process of ChatGPT

https://openai.com/blog/chatgpt

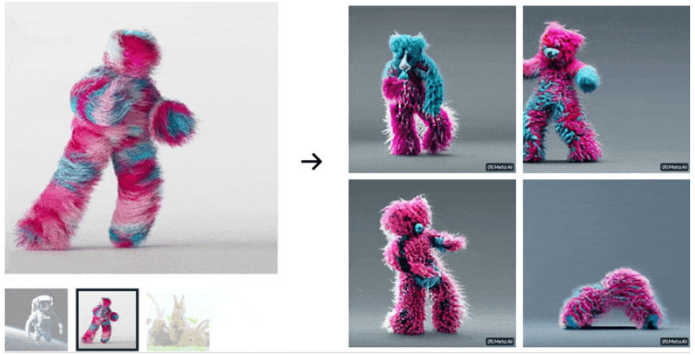
AI-generated Texts, Images, Videos, 3D



AudioLM



Text-to-image



make-a-video

https://openai.com/blog/chatgpt

Applications of Mobile AIGC Networks

Application	Models	Network Architectures	Datasets	Evaluation Metrics
Text Generation	GPT-3 [67], GPT-4, BERT [68], LaMDA [69], ChatGPT [10]	Transformer [70]	WebText, BookCorpus [71], Common Crawl	BLEU [72], ROUGE [73], Perplexity
Image Generation	StyleGAN [74], BigGANs [75], StyleGANXL [76], DVD-GAN [77], DALLE [6], DALLE2 [7], CLIP [5], VisualGPT [78], VAE [79], Energy-based GAN [62], Flow-based models [64], Imagen [80], diffusion probabilistic models [81], DDPM [82], DDIM [83]	Quantised-Variational GAN [84], VQ-VAE [85], Transformer [70]	ImageNet [86], CelebA [87], COCO [88]	FID [89], IS [90], LPIPS [91]
Music Generation	MuseNet [92], Jukedeck, WaveNet [93], AudioLM [94]	Transformer, RNN, CNN	MIDI Dataset, MAESTRO [95]	ABC-notation, Music IS
Video Generation	Diffusion models beat GANs [96], Video Diffusion Models [97], Dreamfusion [98]	DDPM, DDIM	Kinetics [99]	PSNR, SSIM
3D Generation	NeRF [100]	MLP	Synthetic and real-world scenes	PSNR, SSIM, LPIPS

TABLE I: Summary of State-of-the-art AIGC models.

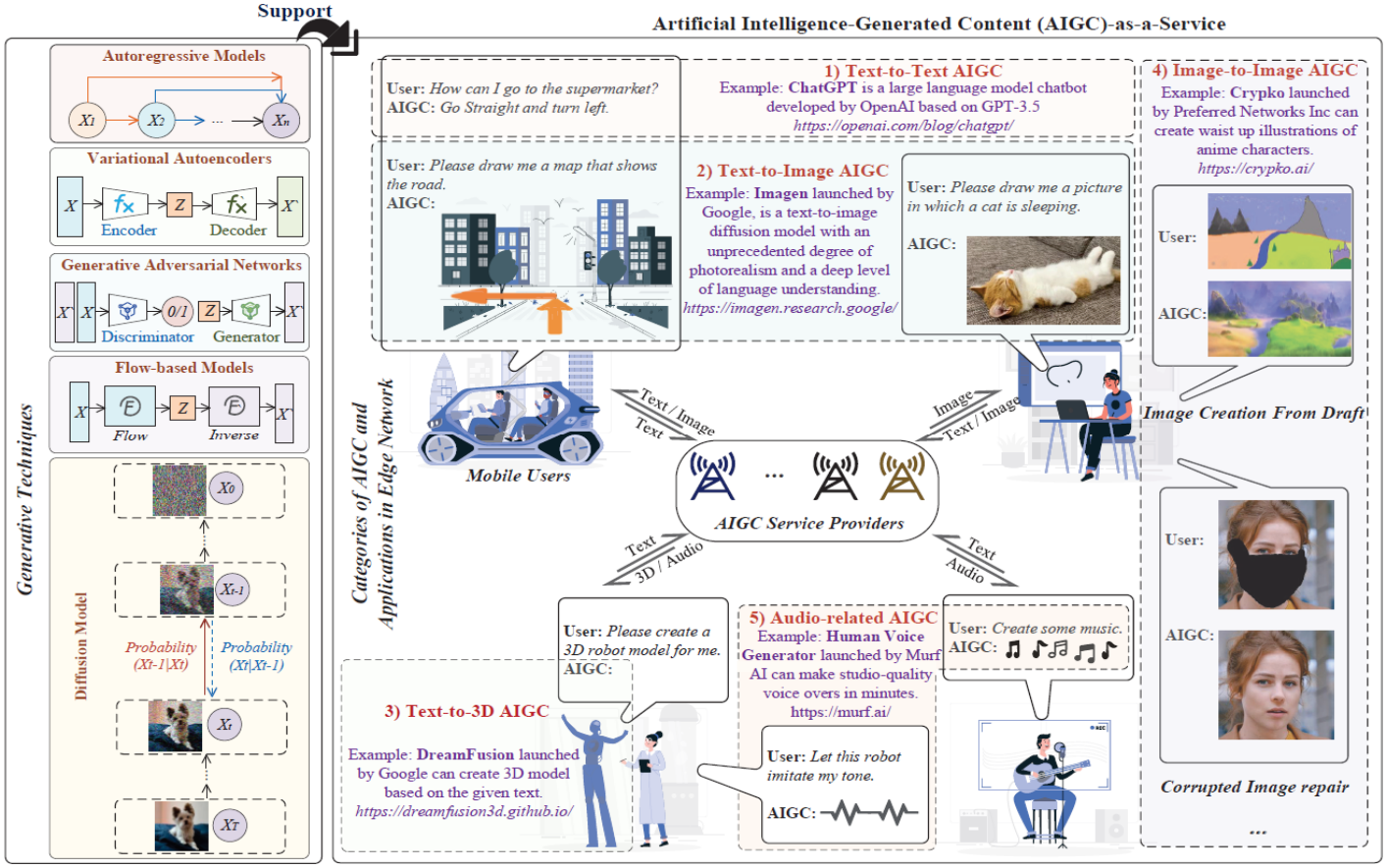
Generative-AI technologies

- **Data-based generative AI** employs learning algorithms to create novel content from existing data
- **Model-based generative AI** leverages predefined models and simulations to generate new outputs by manipulating model parameters

Data-based Generative-AI technologies		
Technologies	**Advantages**	**Disadvantages**
Neural style transfer [5]	-Ability to combine style and content from different images -Real-time style transfer with optimized algorithms -Applicable to various forms of media	-Dependence on high-quality style and content images -Difficulty in preserving fine details -Difficulty in controlling the level of stylization
Pixel CNN/Pixel RNN [6]	Autoregressive modeling of image pixels -Explicitly captures local dependencies in data -Effective in modeling discrete data	-Computationally expensive -Difficulty in capturing long-range dependencies -Sequential nature limits generation speed
Transformer-based model [7]	-Effective modeling of long-range dependencies -Scalability with parallelization techniques -Superior performance on various tasks	-Memory consumption in large-scale applications -Complexity increases with sequence length -Dependence on large training datasets
Example-based synthesis [8]	-Leveraging existing data for generation -Ability to produce high-quality results -Flexibility in incorporating domain-specific constraints	-Dependence on the quality and variety of input examples -Difficulty in generalizing to unseen data -Computational complexity
Model-based Generative-AI Technologies		
Technologies	**Advantages**	**Disadvantages**
Generative adversarial network [2]	-Exceptional quality of generated images -Can incorporate conditions to control generated output -Ability to address mode collapse with advanced techniques	-Training instability -Difficulty in evaluating model quality -Susceptibility to mode collapse
Variational Autoencoder [9]	-Effective representation of data in a probabilistic latent space -Achieving disentangled representations -Unified framework for both inference and generation	-Blurry image generation -Limited expressive power of the prior distribution -Difficulty in choosing likelihood functions
Autoregressive model [10]	-Direct modeling of temporal dependencies -Scalability achieved through parallelization techniques -Compatibility with both discrete and continuous data	-Sequential nature limits speed -Complexity increases with sequence length -Difficulty in modeling long-range dependencies
Diffusion-based model [4]	-Generative process based on denoising score matching -Flexible noise schedule selection -Resistance to mode collapse and overfitting	-Slower generation process -Complexity in training and selecting hyperparameters

Ruichen Zhang, Ke Xiong, Hongyang Du, Dusit Niyato, et al "Generative AI-enabled Vehicular Networks: Fundamental, Framework, and Case Study."

The Role of Wireless Network



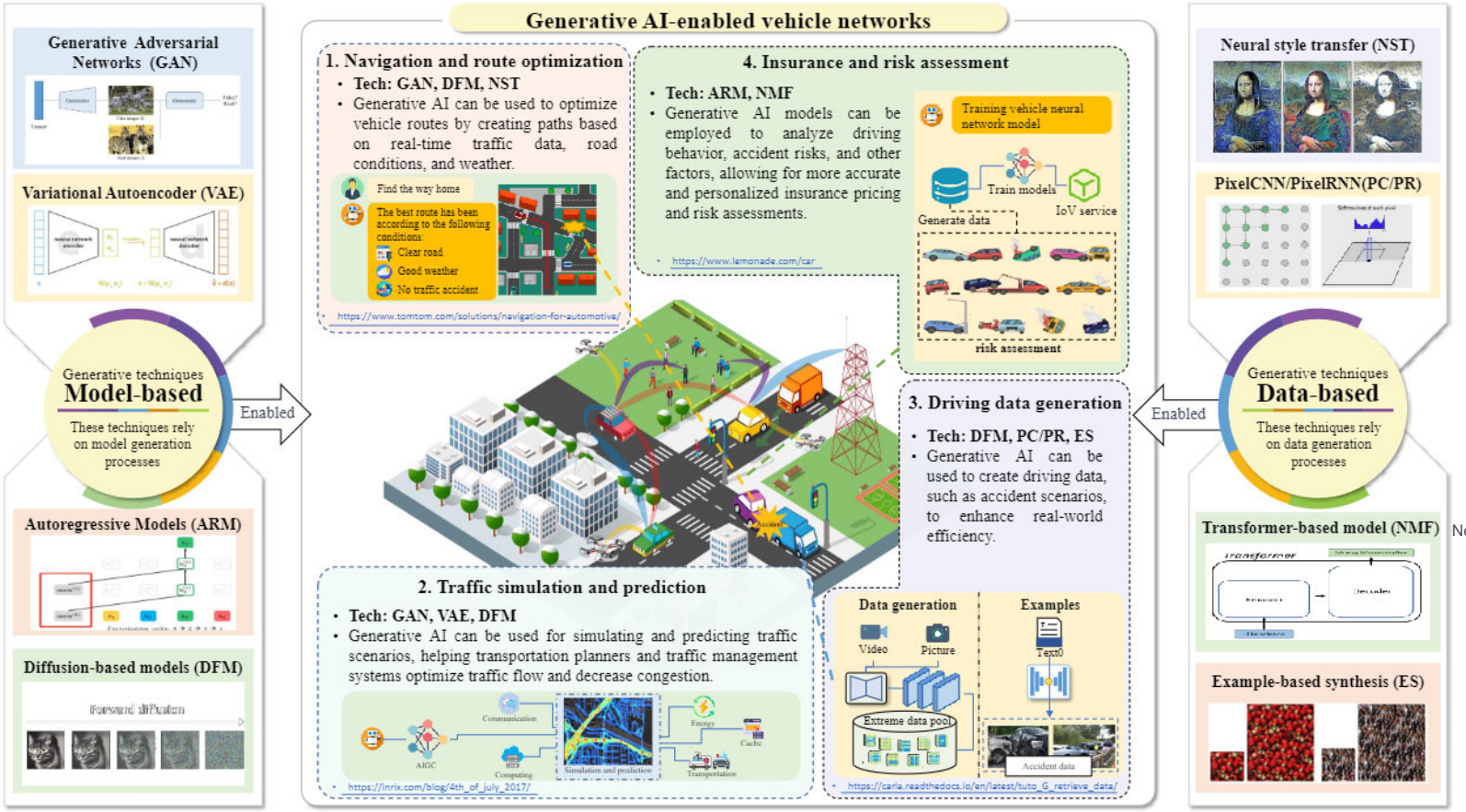
Generative techniques in AIGC [9], categories of AIGC, and applications in wireless edge networks. We list several online available AIGC services as examples, e.g., ChatGPT for text-to-text AIGC (https://openai.com/blog/chatgpt/), Imagen for text-to-image AIGC (https://imagen.research.google/), DreamFusion for text-to-3D AIGC (https://dreamfusion3d.github.io/), Crypko for image-to-image AIGC (https://crypko.ai/), and Human Voice Generator for audio-related AIGC (https://murf.ai/).

Motivation of Mobile AIGC Networks

- Low-latency:
  - users can access low-latency services in mobile AIGC networks
- Localization and Mobility:
  - AIGC servers at the network's edge can fine-tune pre-trained models by localizing service requests
  - user mobility can be integrated into the AIGC service provisioning process
- Customization and Personalization:
  - request personalized services based on their preferences
  - provide customized services according to local service environments
- Privacy and Security:
  - AIGC users only need to submit service requests to edge servers
  - privacy and security of AIGC users can be preserved during AIGC service provisioning

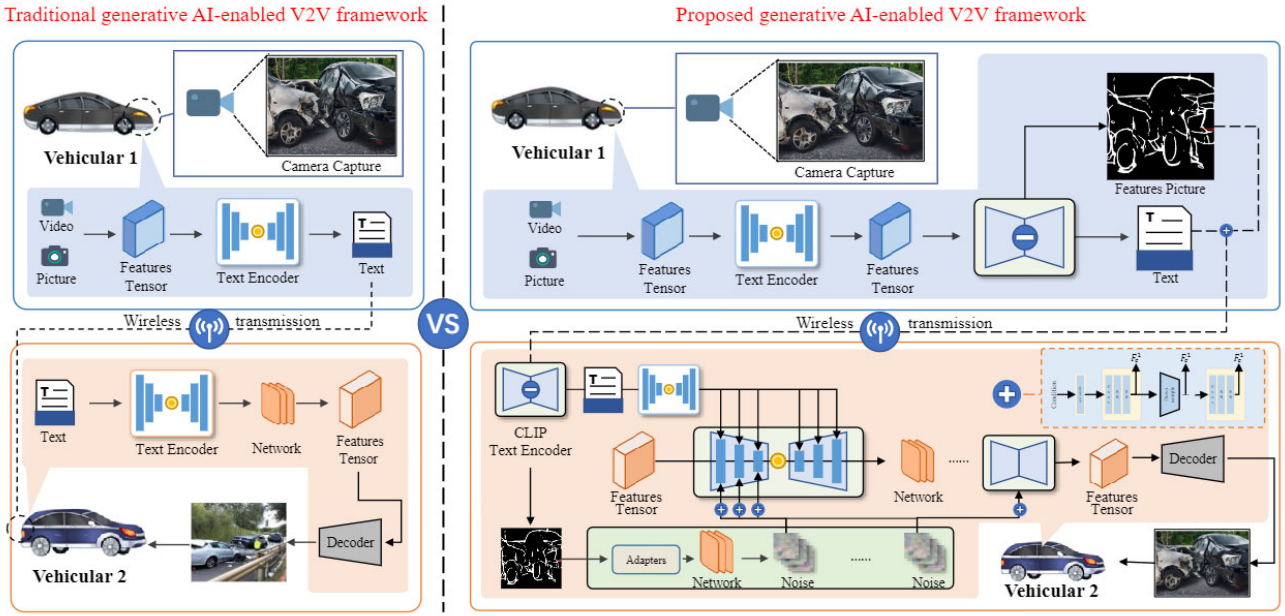
Generative AI enabled Vehicular Networks

Generative AI-enabled vehicular network is capable of analyzing traffic patterns and providing drivers with real-time suggestions for alternative routes to circumvent congestion or accidents



Ruichen Zhang, Ke Xiong, Hongyang Du, Dusit Niyato, Jiawen Kang, Xuemin Shen, H. Vincent Poor, "Generative AI-enabled Vehicular Networks: Fundamental, Framework, and Case Study," https://arxiv.org/abs/2304.11098

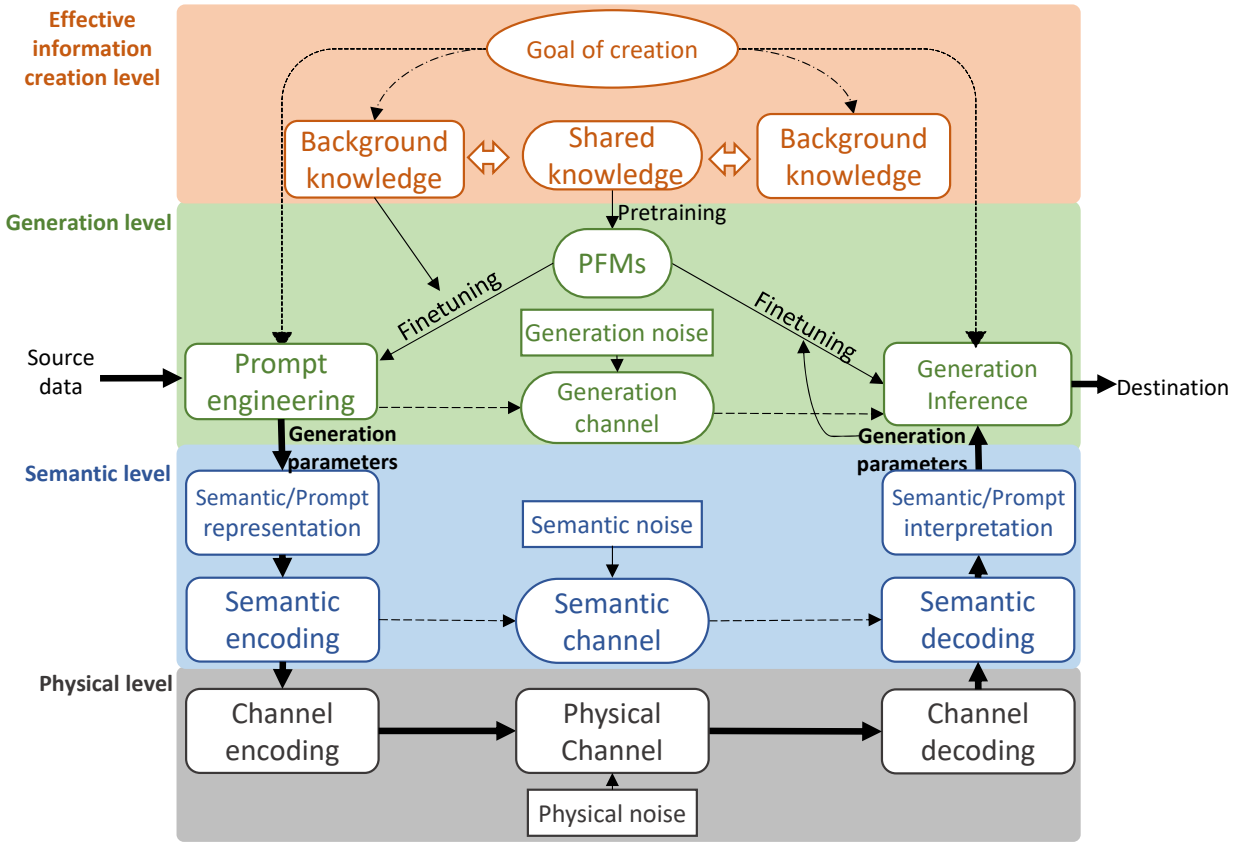
Semantic-aware generative AI-enabled Vehicular Network



- Traditional generative AI-enabled vehicles can use semantic communication technology to extract information from images at the transmitter end and restore it using generative AI technology at the receiver end
- This approach reduces transmitted data, decreasing communication delays and improving V2V communication efficiency in vehicular networks

Ruichen Zhang, Ke Xiong, Hongyang Du, Dusit Niyato, Jiawen Kang, Xuemin Shen, H. Vincent Poor, "Generative AI-enabled Vehicular Networks: Fundamental, Framework, and Case Study," https://arxiv.org/abs/2304.11098

Semantic Communications Enabled AIGC



Physical level encapsulates transmission operations within the physical layer, including modules responsible for channel encoding and decoding, this level necessitates symbol transmission over a physical channel susceptible to noise

Semantic level is composed of semantic representation and encoding modules at the transmitter end. These components work together to extract and denote semantic information. Furthermore, semantic interpretation and decoding modules situated at the receiver end are for reconstructing and interpreting the received semantic information.

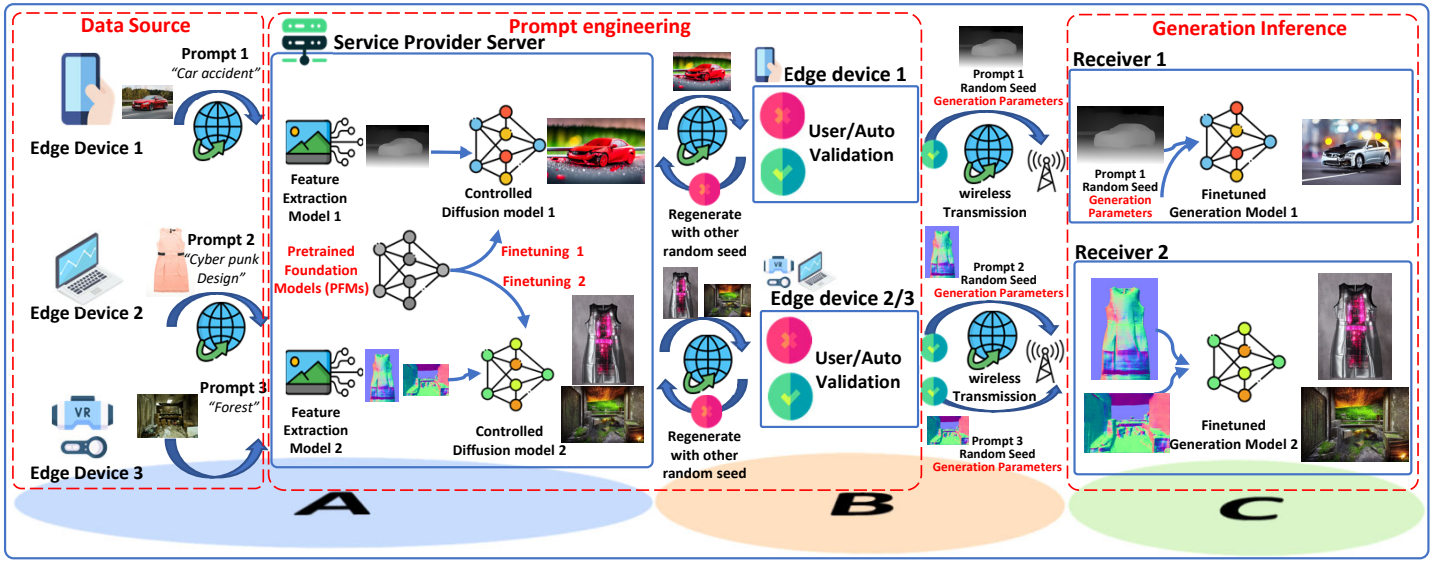
Effective information creation level transforms raw user data into semantic representations, which are then used to guide the generation process. It involves feature extraction and the creation of model fine-tuning parameters based on context-specific requirements.

Generation level provides an engine of content creation that utilizes semantically-encoded information and the generation parameters to run controlled diffusion models, generating the final content.

Together, these levels augment the traditional SemCom model by adding control and personalization, thereby making AIGC a more adaptable and effective tool for content generation and transmission.

Guangyuan Liu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Dong In Kim, and Xuemin (Sherman) Shen, "Semantic Communications for Artificial Intelligence Generated Content (AIGC) Toward Effective Content Creation"

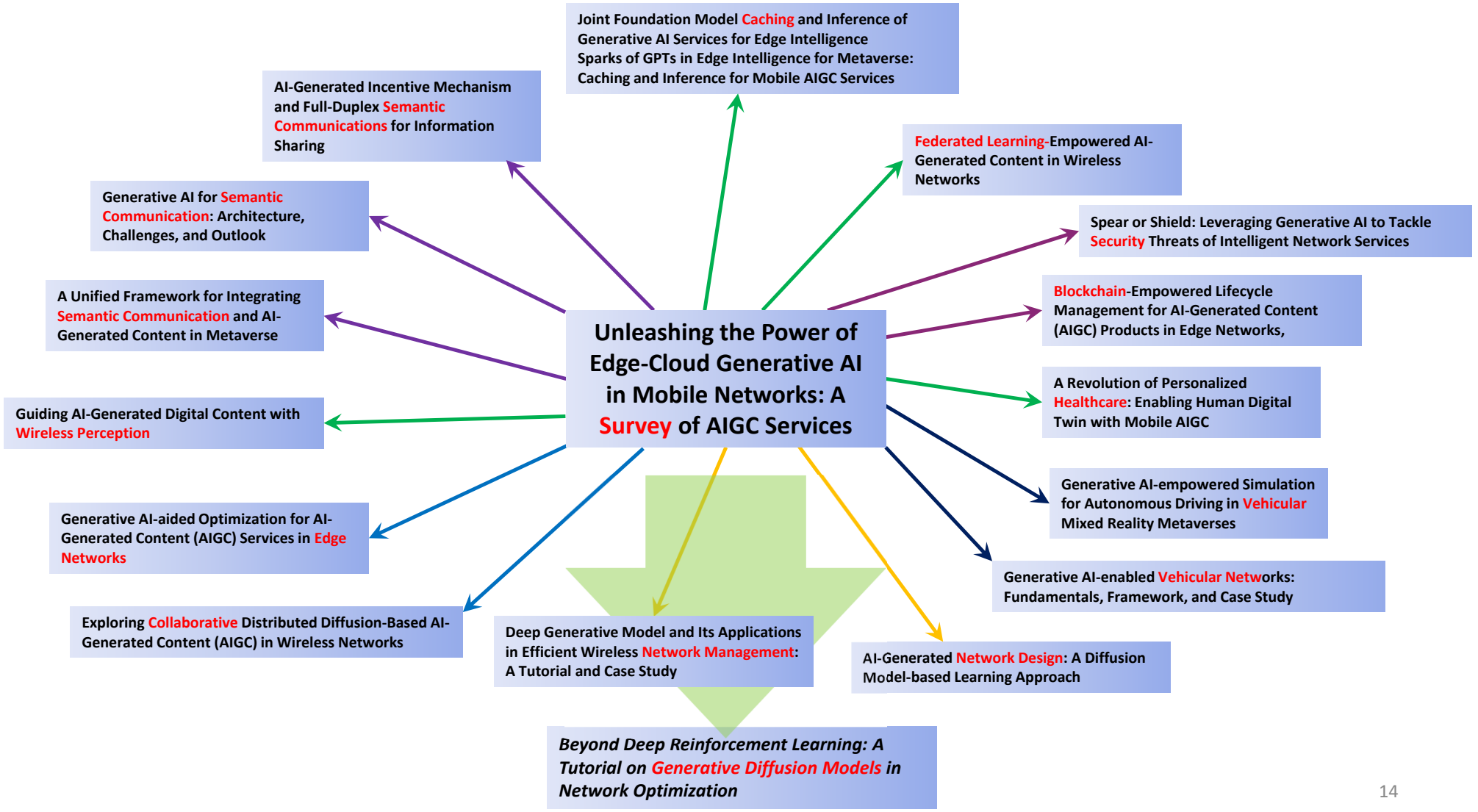
Semantic Communications Enabled AIGC



AIGC model as generation encoder and decoder

- Flow of information in this framework starts from the transmitter, where a prompt is extracted and optimized through prompt engineering
- Semantic/prompt representation plays a crucial role in this process
- Optimized prompt, along with the generation parameters, is used to generate content through generative AI models
- At the receiver end, the received semantic information/prompt representation is interpreted and used for generation inference to generate the desired content as intended by the transmitter

Guangyuan Liu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Dong In Kim, and Xuemin (Sherman) Shen, "Semantic Communications for Artificial Intelligence Generated Content (AIGC) Toward Effective Content Creation"


Joint Foundation Model Caching and Inference of Generative AI Services for Edge Intelligence
Sparks of GPTs in Edge Intelligence for Metaverse: Caching and Inference for Mobile AIGC Services
AI-Generated Incentive Mechanism and Full-Duplex Semantic Communications for Information Sharing
Federated Learning-Empowered AI-Generated Content in Wireless Networks
Generative AI for Semantic Communication: Architecture, Challenges, and Outlook
Spear or Shield: Leveraging Generative AI to Tackle Security Threats of Intelligent Network Services
A Unified Framework for Integrating Semantic Communication and AI-Generated Content in Metaverse
Blockchain-Empowered Lifecycle Management for AI-Generated Content (AIGC) Products in Edge Networks,
Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services
A Revolution of Personalized Healthcare: Enabling Human Digital Twin with Mobile AIGC
Guiding AI-Generated Digital Content with Wireless Perception
Generative AI-empowered Simulation for Autonomous Driving in Vehicular Mixed Reality Metaverses
Generative AI-aided Optimization for AI-Generated Content (AIGC) Services in Edge Networks
Generative AI-enabled Vehicular Networks: Fundamentals, Framework, and Case Study
Exploring Collaborative Distributed Diffusion-Based AI-Generated Content (AIGC) in Wireless Networks
Deep Generative Model and Its Applications in Efficient Wireless Network Management: A Tutorial and Case Study
AI-Generated Network Design: A Diffusion Model-based Learning Approach
Beyond Deep Reinforcement Learning: A Tutorial on Generative Diffusion Models in Network Optimization