Knowledge-Guided 3D CT Generation: A Conditioning-Centric Taxonomy

Francesca Pia Panaccione , Eugenio Lomurno , Matteo Matteucci

Department of Electronics, Information, and Bioengineering
Politecnico di Milano

francescapia.panaccione@polimi.it, eugenio.lomurno@polimi.it, matteo.matteucci@polimi.it

Abstract

Controllable generation guided by external knowledge is a key requirement in modern generative deep learning applications, enabling the synthesis of samples with explicit constraints on semantic content, structural properties, and variability. In 3D Computed Tomography (CT), such control is essential for clinical applications, including data augmentation, privacy-preserving data sharing, and the simulation of specific anatomical or pathological scenarios. While research on conditional 3D CT generation has expanded rapidly, the diversity of existing approaches makes systematic comparison difficult and obscures fundamental design choices. In this survey, we propose a conditioning-centric taxonomy that organizes the literature along three orthogonal dimensions: the type of external knowledge (K), the knowledge integration paradigm (I), and the generative architecture (A). This factorization defines an explicit design space ($\mathcal{K} \times \mathcal{I} \times \mathcal{A}$) that provides a unified perspective on prior work. Using this framework, we systematize existing methods, identify dominant trends and recurring design patterns, and highlight underexplored regions of the design space that point toward promising directions for future research.

1 Introduction

Recent advances in deep generative modeling have enabled the synthesis of complex data across multiple domains. In medical imaging, synthetic data generation addresses limitations of real-world datasets, including high acquisition costs, extensive annotation requirements, and privacy constraints [Koetzier *et al.*, 2024; Lomurno and Matteucci, 2025]. Three dimensional (3D) medical imaging modalities represent volumetric data as spatially correlated slices, offering richer anatomical information than two-dimensional imaging [Wu *et al.*, 2025] while requiring generative models to maintain coherence across the entire volume [Friedrich *et al.*, 2024b]. Among these, 3D computed tomography (CT) scans are increasingly adopted in clinical practice for diagnosing a wide range of conditions through detailed volumetric representations.

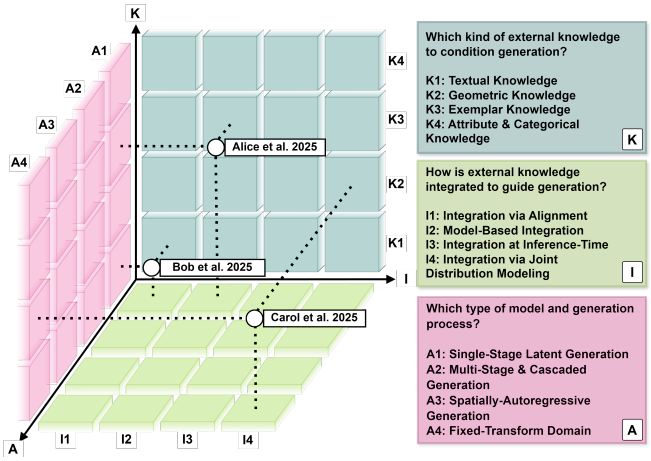

Figure 1: Conceptual representation of the proposed taxonomy as a three-dimensional design space. Each knowledge-guided 3D CT generation method can be positioned as a point in $\mathcal{K} \times \mathcal{I} \times \mathcal{A}$, defined by its choices along three orthogonal axes: external knowledge type ($\mathcal{K}$), integration paradigm ($\mathcal{I}$), and generative architecture ($\mathcal{A}$).

Early approaches to 3D CT generation have primarily focused on unconditional modeling, learning the data distribution directly from volumetric samples without external guidance. Unfortunately, this setting has often failed to adequately control artifacts or ensure semantic consistency [Friedrich *et al.*, 2024b]. Motivated by these limitations, and supported by the increasing availability of large-scale public CT datasets [Hamamci *et al.*, 2024b], recent work has shifted toward the development of more expressive, *knowledge-guided* generative frameworks. These approaches incorporate structured priors—semantic descriptors, anatomical constraints, or population attributes—to introduce inductive biases regulating generation toward coherent, anatomically consistent, and clinically meaningful volumetric patterns. [Dorjsembe *et al.*, 2024; Xu *et al.*, 2024; Guo *et al.*, 2025b; Yoon *et al.*, 2025b].

This rapid growth has revealed that guidance effectiveness depends critically on knowledge type and integration mechanism [Hamamci *et al.*, 2024a; Amirrajab *et al.*, 2025; Molino *et al.*, 2025]. Nevertheless, the conditioning process itself has not yet been systematically examined to identify patterns and relationships across methods. Existing surveys

organize this emerging field by architectural families or application domains [Zhou *et al.*, 2025; Friedrich *et al.*, 2024b; Chen and Ramsey, 2024; Liu *et al.*, 2024], focusing on traditional categorization patterns that result in a fragmented landscape where trends, failure modes, and principled future research directions remain obscured.

This survey addresses this gap by reframing the literature through a conditioning-centric perspective, examining how knowledge is represented, injected, and exploited throughout generation. To this end we present an innovative faceted taxonomy (Figure 1) that organizes methods along three orthogonal dimensions: (i) the type of external knowledge used for conditioning, (ii) how such knowledge is exploited within the conditioning paradigm, and (iii) the underlying generative process and model architecture. Together, these dimensions define a three-dimensional design space in which methods can be systematically positioned and compared. To summarize, the contributions of this survey are as follows:

- **A conditioning-centric taxonomy** that introduces a new perspective for analyzing 3D CT generation methods, enabling principled positioning and comparison of existing techniques while remaining extensible to include future work.
- **A unified design-space analysis** grounded in the interpretable $\mathcal{K} \times \mathcal{I} \times \mathcal{A}$ space, revealing dominant paradigms and underexplored regions.
- **Clear research directions** derived from explicit gaps in the design space, identifying actionable conditioning strategies beyond incremental refinements of current approaches.

To facilitate adoption, we provide an open-source repository with reference implementations and an interactive classification tool at https://github.com/eugeniolomurno/3D-CT-taxonomy.

2 Preliminaries

This section formalizes knowledge-guided 3D CT generation as a constrained stochastic sampling problem over high-dimensional volumetric data.

Knowledge-Guided 3D CT Generation. Let $x \in \mathcal{X} \subset \mathbb{R}^{H \times W \times D}$ denote a 3D CT volume represented as a voxel-based grid with inherent spatial coherence, and let $k \in \mathcal{K}$ denote external information available at generation time. Knowledge-guided CT generation is formalized as sampling from a conditional distribution $p_\theta(x \mid k)$, where θ parameterizes the generative model. The conditioning variable k is assumed exogenous with respect to the volumetric state. For fixed k, the conditional distribution retains stochastic support, admitting multiple plausible realizations that satisfy the imposed constraints. Knowledge-guided generation is therefore treated as constrained stochastic sampling, where modeling choices regulate how uncertainty is resolved under external constraints.

Probabilistic Conditioning. Conditioning can be instantiated through distinct probabilistic formulations. Most approaches directly parameterize the conditional distribution $p_\theta(x \mid k)$, treating k as a fixed input that modulates the generative process during training. Others model the joint distribution $p_\theta(x, k)$, where both volume and conditioning signal are co-generated, enforcing intrinsic alignment rather than treating k as an external constraint. A third paradigm modifies the sampling process at inference time: the model parameters remain fixed, but the sampling trajectory is dynamically adjusted to favor configurations consistent with k. These formulations correspond to different probabilistic objectives and encode different assumptions on how constraints interact with generative uncertainty.

Survey	Organization	Breadth	Quant. Analysis	Design Space
[Khader *et al.*, 2023]	Architecture	Med-3D	Performance	–
[Friedrich *et al.*, 2024b]	Modality	Med-3D	–	–
[Liu *et al.*, 2024]	Application	Brain/Heart	Performance	–
[Chen and Ramsey, 2024]	Architecture	Gen-3D	–	–
[Zhou *et al.*, 2025]	Application	Med-All	Performance	–
Ours	**Conditioning**	**3D CT**	**Distributional**	✓

Table 1: Comparison with related surveys. This work introduces the first conditioning-centric taxonomy with explicit method positioning a structured design space ($\mathcal{K} \times \mathcal{I} \times \mathcal{A}$).

Distribution Decomposition in Volumetric Generation. The dimensionality and spatial structure of $\mathcal{X}$ require decomposing the target distribution into tractable components. Abstractly, volumetric synthesis proceeds by selecting a decomposition strategy for $p(x \mid k)$: through global stochastic processes acting on the full volume, through explicit spatial factorizations $p(x \mid k) = \prod_n p(x_n \mid x_{<n}, k)$ as in spatial-autoregressive schemes, through hierarchical multi-scale decompositions generating coarse-to-fine structure, or through deterministic invertible transformations $x = T^{-1}(y)$ with $y \sim p(y \mid k)$ that redistribute spatial dependencies into structured representation spaces. In practice, these decompositions are often implemented in learned latent spaces via an encoder–decoder pair (E, D), yielding the generative objective $p_\theta(z \mid k)$ with $z = E(x)$ and $x \approx D(z)$ at reduced computational cost.

3 Taxonomy

This section introduces the proposed taxonomy for organizing 3D CT knowledge-guided approaches. In contrast to prior surveys, which predominantly organize the literature by architectural families, imaging modalities, or application domains (Table 1) [Khader *et al.*, 2023; Friedrich *et al.*, 2024b; Chen and Ramsey, 2024; Liu *et al.*, 2024; Zhou *et al.*, 2025], our formulation explicitly disentangles the conditional generative process in 3D CT synthesis. Rather than ranking methods or prescribing optimal designs, the taxonomy provides a descriptive conceptual framework for organizing and interpreting the existing literature.

3.1 Taxonomy Structure

The taxonomy is structured as a three-dimensional design space $\mathcal{K} \times \mathcal{I} \times \mathcal{A}$ that factorizes knowledge-guided 3D CT generation methods along three independent axes (Figure 1). Each axis captures a distinct structural aspect of model design: the type of external knowledge involved (Axis $\mathcal{K}$), the

Axis $\mathcal{K}$: External Knowledge	
Category	**Representative Instantiations**
K1 - Textual Knowledge	Radiology reports, free-text clinical descriptions, open-vocabulary prompts, unstructured findings.
K2 - Geometric Knowledge	Organ segmentation masks, anatomical layouts, bounding boxes, landmark coordinates, sparse structural maps.
K3 - Exemplar Knowledge	Reference volumes from complementary modalities (e.g., MRI, PET, CBCT), patient-specific priors, atlas templates.
K4 - Attribute & Categorical Knowledge	Patient demographics (e.g., age, sex, BMI), diagnostic class labels, acquisition metadata, structured clinical attributes.

Table 2: Taxonomy of Axis $\mathcal{K}$ (External Knowledge). Categories represent different types of information used to condition 3D CT generation, ranging from unstructured text to dense volumetric priors.

Axis $\mathcal{I}$: Knowledge Integration	
Category	**Representative Instantiations**
I1 - Integration via Alignment	Pre-generative embedding alignment (e.g., CLIP, BioViL), dual-encoder architectures, contrastive representation learning.
I2 - Model-Based Integration	Cross-attention layers, channel-wise concatenation, feature-wise modulation (FiLM, AdaGN), ControlNet adapters.
I3 - Integration at Inference-Time	Classifier-free guidance, energy-based Guidance, trajectory smoothing, compositional guidance.
I4 - Integration via Joint distribution modeling	joint probability factorization $p(x, k)$, unified diffusion over concatenated modalities (e.g., image + mask).

Table 3: Taxonomy of Axis $\mathcal{I}$ (Knowledge Integration). Categories distinguish the stage and mechanism through which external knowledge influences the generative process, from pre-generative alignment to joint distribution modeling.

mechanism by which such knowledge interacts with the generative process (Axis $\mathcal{I}$), and the strategy adopted for volumetric synthesis (Axis $\mathcal{A}$). Within this structure, each method is represented as a tuple (k, i, a), where $k \subseteq \mathcal{K}$, $i \subseteq \mathcal{I}$, and $a \subseteq \mathcal{A}$. All components are non-empty, and multiple categories per axis are permitted to account for hybrid approaches. This explicit factorization enables systematic comparison across heterogeneous methods, making recurrent patterns as well as sparsely explored regions of the design space directly observable. The categories defining each axis are detailed in the following subsections, with representative instantiations summarized in Tables 2, 3, and 4.

3.2 Axis $\mathcal{K}$: External Knowledge

This axis characterizes the form of external knowledge used to condition generation. Consistently with the problem formulation in Section 2, external knowledge $k \in \mathcal{K}$ is defined as information exogenous to the CT volume. Different instantiations of k vary in modality and in the degree of constraint they impose on generation. Based on these differences, we identify four broad categories, described below and illustrated with representative examples in Table 2:

K1: Textual Knowledge. K1 corresponds to unstructured semantic information encoded in linguistic form and mapped to latent representations. It provides high-level semantic conditioning without explicitly enforcing structural constraints, such that anatomical localization and organization are expected to emerge implicitly through the generative dynamics.

K2: Geometric Knowledge. K2 encodes constraints on anatomical structure and spatial organization, either through voxel-aligned representations or abstract geometric descriptors. While these priors enforce structural plausibility, they abstract away non-geometric variation, limiting the modeling of physiological patterns not directly captured by geometry.

K3: Exemplar Knowledge. K3 comprises dense volumetric priors from reference instances. By providing voxel-aligned structural and appearance information, exemplar knowledge anchors generation to specific anatomical realizations, constraining synthesis at the instance level.

K4: Attribute & Categorical Knowledge. K4 encodes non-spatial descriptive variables associated with the target volume, enabling global modulation of the data distribution. This conditioning captures population-level regularities linked to phenotypic or acquisition-related factors without prescribing local spatial structure.

3.3 Axis $\mathcal{I}$: Knowledge Integration

This axis characterizes how external knowledge influences the generative process, independently of its instantiation (Axis $\mathcal{K}$) or the architectural backbone (Axis $\mathcal{A}$). We identify four paradigms (Table 3). Throughout this section, the generative target is denoted by x, representing either volumetric data or its latent representation.

I1: Integration via Alignment. I1 paradigms enforce conditioning by establishing a correspondence between external knowledge k and x within a shared representation space prior to generation. Both modalities are mapped to compatible representations aligned under a common similarity or consistency criterion. Conditioning is thus imposed implicitly at the representation level, with knowledge influencing generation through alignment rather than direct intervention in the model's internal dynamics.

I2: Model-Based Integration. In I2 paradigm, external knowledge k is incorporated directly into the generator architecture to condition the generative dynamics, corresponding to learning a conditional predictor (e.g., $\epsilon_\theta(x_t, t \mid k)$) in which knowledge modulates the model's internal computations. The integration mechanism depends on the structural properties of k: non-spatial or abstract signals induce global modulation, whereas spatially aligned knowledge enable localized conditioning that preserves correspondence. Conditioning is embedded within the learned model and consistently guides generation.

I3: Integration at Inference-Time. I3 paradigms apply conditioning exclusively at inference time, leaving the learned generative distribution unchanged. In diffusion-based models, a canonical instantiation is classifier-free guidance,

Axis $\mathcal{A}$: Generative Architecture	
Category	**Representative Instantiations**
A1 - Single-Stage Latent Generation	Holistic generation in learned latent spaces, 3D-VQGAN backbones, one-shot synthesis.
A2 - Multi-Stage & Cascaded Generation	Coarse-to-fine synthesis, super-resolution cascades, hierarchical residual modeling.
A3 - Spatially-Autoregressive Generation	Slice-wise autoregression, video-like generation (z-axis evolution), temporal transformers, spatial decomposition.
A4 - Fixed-Transform Domain	Wavelet-domain diffusion, spectral diffusion, generation in deterministic invertible spaces.

Table 4: Taxonomy of Axis $\mathcal{A}$ (Generative Architecture). Categories characterize the structural strategy for volumetric synthesis, distinguishing methods by their procedural decomposition and representation space.

which linearly combines unconditional and conditional predictions at each diffusion step t as $\tilde{\epsilon}(x_t \mid k) = \epsilon_\theta(x_t, t) + s \cdot (\tilde{\epsilon}\theta(x_t, t \mid k) - \epsilon\theta(x_t, t))$, where s controls the fidelity–diversity trade-off during reverse diffusion.

I4: Integration via Joint Distribution Modeling. Unlike prior paradigms where k is fixed, I4 methods model the joint distribution $p_\theta(x, k)$, generating knowledge and target variables simultaneously. This enforces intrinsic coupling between x and k, enabling jointly emerging structure and appearance. While more computationally demanding, I4 approaches remove conditioning asymmetry and capture bidirectional correlations.

3.4 Axis $\mathcal{A}$: Generative Architecture

This axis characterizes the architectural strategy for volumetric synthesis, independently of knowledge representation ($\mathcal{K}$) or integration ($\mathcal{I}$). It describes how the high-dimensional generation problem is decomposed into tractable subproblems. We identify four strategies based on scope and representation space (Table 4).

A1: Single-Stage Latent Generation. A1 strategies model volumetric synthesis holistically by generating either the full volume x or a learned latent embedding from which x is decoded. Generation is performed in a single stage, without explicit hierarchical or spatial decomposition. This formulation favors computational efficiency and global coherence, relying on the expressive capacity of the latent representation to capture structural consistency.

A2: Multi-Stage & Cascaded Generation. A2 strategies decompose volumetric synthesis into a sequence of conditional stages, typically operating at increasing resolutions. Generation follows a coarse-to-fine hierarchy, with $x^{(1)} \sim p_{\theta_1}(x^{(1)})$ and $x^{(2)} \sim p_{\theta_2}(x^{(2)} \mid x^{(1)})$, where the initial stage captures low-resolution global structure and subsequent stages progressively refine local details.

A3: Spatially-Autoregressive Generation. A3 methods factorize the generation along a spatial axis (typically the axial depth D), treating the 3D volume as a sequence of 2D slices or sub-volumes. The process follows an ordered dependency, where the generation of the n-th slice is conditioned on the previously synthesized context ($x_{<n}$). This formulation models inter-slice dependencies through sequential factorization.

A4: Fixed-Transform Domain Generation. A4 strategies perform synthesis in a deterministic representation space defined by a fixed, invertible transform T. Generation is carried out over transformed coefficients $y \sim p_\theta(y \mid k)$, with reconstruction via $x = T^{-1}(y)$. These approaches reduce dimensionality while preserving exact invertibility, at the expense of flexibility in representation learning.

4 Literature Trends

This section applies the proposed taxonomy ($\mathcal{K} \times \mathcal{I} \times \mathcal{A}$) to organize and analyze the literature on knowledge-guided 3D CT generation. Table 5 provides a comprehensive overview of all reviewed methods, while Figure 2 quantifies the distribution of approaches across the three axes. The analysis is structured thematically, identifying dominant trends, recurring patterns, and underexplored regions in the proposed design space.

External Knowledge Trends. Figure 2 (top-left) shows the distribution of methods across knowledge categories. *Geometric knowledge* (K2) is the most prevalent, accounting for 44% of methods, instantiated primarily through organ segmentation masks. Representative approaches include MAISI, MAISI-v2, NodMAISI, LAND, and MedLoRD [Guo *et al.*, 2025b; Zhao *et al.*, 2025; Tushar *et al.*, 2025; Oliveras *et al.*, 2025; Seyfarth *et al.*, 2025].

Exemplar conditioning (K3) accounts for 28% of existing methods and is predominantly used in cross-modal translation settings (e.g., MRI→CT). Representative approaches such as 3DLDM, 3D-WLDM, and Med-LVDM leverage source latents as structural exemplars, enabling strong alignment between input and output. Hybrid K2+K3 strategies, including DiffTumor and Lung-DDPM, demonstrate lesion-aware or layout-guided synthesis with high structural fidelity [Mahdi *et al.*, 2025; Zheng *et al.*, 2025; Kui *et al.*, 2025; Chen *et al.*, 2024; Jiang *et al.*, 2025].

Text-based conditioning (K1) accounts for 22% of existing methods. Representative approaches include GenerateCT, Text-to-CT, Report2CT, and Text2CT, which adopt medical-specific encoders or task-adapted language models to obtain richer semantic representations [Hamamci *et al.*, 2024a; Molino *et al.*, 2025; Amirrajab *et al.*, 2025; Guo *et al.*, 2025a]. *Distributional knowledge* (K4) accounts for 6% of existing methods. Representative examples include Cascaded-3D and Surf2CT, as summarized in Table 5. Non-spatial descriptive variables such as demographics can be obtained without additional encoders [Yoon *et al.*, 2025b; Yoon *et al.*, 2025a].

Knowledge Integration Trends. The distribution across integration paradigms (Figure 2, top-center) reveals strong concentration. *Model-based integration* (I2) accounts for 74% of methods, incorporating external knowledge directly into the generator architecture. Representative approaches employ mechanisms such as cross-attention or feature concatenation to inject conditioning signals, enabling compat-

Method	K	I	A	Conditioning Source	Conditioning Mechanism	Generation Strategy	Dataset (Resolution)	Code
MedGen3D [Han *et al.*, 2023]	**K2**	**I4**	**A3**	Voxel-aligned Semantic Maps	Joint mask–image diffusion (MC-DPM)	Autoregressive slice-wise	SegTHOR (96×320^2)	–
GenerateCT [Hamamci *et al.*, 2024a]	**K1**	**I1, I2**	**A2**	Free-form Text Prompts	Cross-attention, CFG	Cascaded low-res→sup-res	CT-RATE ($512^2\times201$)	✓
MedSyn [Xu *et al.*, 2024]	**K1, K2**	**I4**	**A2**	Textual Description + Anatomical Semantic Layouts	Joint Diffusion	Multi-Staged low-res→sup-res	Private Lung Dataset (256^3)	✓
GEM-3D [Zhu *et al.*, 2024]	**K2, K3**	**I2**	**A3**	Anatomical Masks + Reference slice	Latent Concatenation	Sequential Window-based	AbdomenCT-1K ($512^2\times$Z, Z variable)	✓
CM3dLDM [Tapp *et al.*,	**K3**	**I2**	**A1**	Cross-modal Volume (MRI)	Frozen Encoder Injection	Single-stage Patch-based Diffusion	Private MR-CT Head Dataset; SynthRad (224^3)	✓
cWDM [Friedrich *et al.*, 2024a]	**K3**	**I2**	**A4**	Cross-modal Volume (MRI)	Wavelet-domain Feature Modulation	Fixed-transform Wavelet Diffusion	BraTS 2024 (128^3)	✓
DiffTumor [Chen *et al.*, 2024]	**K2, K3**	**I2**	**A1**	Masks + Healthy CT Volume	Latent Patch Concatenation	Single-stage Latent Diffusion (Inpainting)	LiTS; MSD; KiTS; AbdAtlas-8K; Hopkins (96^3)	✓
MC-IDDDPM [Pan *et al.*, 2024]	**K3**	**I2**	**A4**	Cross-modal Volume (MRI)	Feature Concatenation	Single-stage Diffusion	Private Brain and Prostate Datasets MRI-CT ($192^2\times96$)	–
LN-DDPM [Yu *et al.*, 2024]	**K2**	**I2**	**A1**	Lymph Node + Organ Masks	Spatial Concatenation	Single-stage patch-based diffusion	Private Colorectal Dataset; ABD-LN; (128^3)	–
Med-LVDM [Kui *et al.*, 2025]	**K3**	**I2**	**A1**	Cross-modal Volume (MRI)	Latent concatenation	Single-stage Latent Diffusion	Pelvis MR–CT ($256^2\times$Z)	✓
Cascaded-3D [Yoon *et al.*, 2025b]	**K4**	**I2**	**A2**	Demographics	AdaGN	Cascaded coarse→ sup-res	AutoPET ($224\times224\times384$)	–
Surf2CT [Yoon *et al.*, 2025a]	**K2,K4**	**I2**	**A2**	Skin Surface + Demographics	Conditional flow matching	Cascaded coarse→super-res	Private Torso CT Dataset; AutoPET ($224\times224\times352$)	–
CTFlow [Wang *et al.*, 2025a]	**K1**	**I2**	**A3**	Clinical Reports	Cross-attention	Slice-as-video Flow Matching	CT-RATE ($256^2\times$Z, Z variable)	–
Report2CT [Amirrajab *et al.*, 2025]	**K1**	**I2, I3**	**A1**	Clinical Reports	Multi-encoder Cross-Attention, CFG	Single-stage Latent Diffusion	CT-RATE ($480^2\times256$)	✓
Text-to-CT [Molino *et al.*, 2025]	**K1**	**I1, I2**	**A1**	Radiology text	3D CLIP pretraining, Cross-Attention	Single-stage Latent Diffusion	CT-RATE ($512^2\times128$)	✓
Text2CT [Guo *et al.*, 2025a]	**K1**	**I2, I3**	**A1**	Free-text descriptions	Cross-attention, CFG	Single-stage Latent Diffusion	CT-RATE; RadChestCT ($512^2\times192$)	–
LAND [Oliveras *et al.*, 2025]	**K2**	**I2**	**A1**	Lung + Nodule Segmentation Masks	Latent Concatenation, Cross Attention	Single-stage Latent Diffusion	LIDC-IDRI; NLST (256^3)	–
MedLoRD [Seyfarth *et al.*, 2025]	**K2**	**I2**	**A1**	Anatomical Segmentation Masks	ControlNet	Single-stage Latent Diffusion	Private Coronary Dataset; LUNA16 ($512^2\times256$)	✓
MAISI [Guo *et al.*, 2025b]	**K2**	**I2**	**A1**	Multi-organ Segmentation Maps	ControlNet	Single-stage Latent Diffusion	Curated multi-source, multi-organ CT dataset ($512^2\times768$)	✓
MAISI-v2 [Zhao *et al.*, 2025]	**K2**	**I2**	**A1**	Multi-organ Segmentation Maps	ControlNet, RCL	Single-stage latent	Curated multi-source multi-organ CT dataset ($512^2\times768$)	✓
NodMAISI [Tushar *et al.*, 2025]	**K2**	**I2**	**A1**	Nodule-specific Semantic Masks	ControlNet	Single-stage Latent Diffusion	Curated multi-source public lung CT datasets ($512^2\times768$)	✓
TRACE [Shao *et al.*, 2025]	**K1, K2**	**I2, I3**	**A3**	Text + Multi-modal Anatomical Masks	Multi-modal Mask Concatenation	Autoregressive Slice-pair (Video-like)	CT-RATE ($256^2\times$Z, Z variable)	✓
Lung-DDPM [Jiang *et al.*, 2025]	**K2, K3**	**I2**	**A1**	Semantic Layouts + Reference CT	Layout concatenation + AAS	Single-stage Layout-guided	Private Lung Dataset; LIDC-IDRI (128^3)	✓
3DLDM [Mahdi *et al.*, 2025]	**K3**	**I2**	**A1**	Cross-modal Volume (MRI)	Latent Feature Concatenation	Single-stage Latent Diffusion	SynthRAD2023 ($96^2\times256$)	–
3D-WLDM [Zheng *et al.*, 2025]	**K3**	**I2**	**A4**	Cross-modal Volume (MRI)	Latent Wavelet Concatenation	Fixed-transform Latent Diffusion	Private PET/MR–PET/CT Datasets (128^3)	–
LabelG [Wang *et al.*, 2025b]	**K2**	**I4**	**A1**	Semantic Segmentation Masks	Joint Latent-space Modeling	Single-stage Latent Diffusion	Private Abdominal CT dataset, AbdomenCT; SegTHOR; MSD10-Colon; ($256\times256\times128$)	–

Table 5: Comprehensive overview of knowledge-guided 3D CT generation methods organized under the proposed taxonomy. For each method, we report the classification along axes K, I, and A, the conditioning source, the conditioning mechanism, the generation strategy, the dataset used with volumetric resolution (as reported by authors when available), and code availability. Methods using multiple categories within a single axis are listed with all applicable labels. Note: CFG = Classifier-Free Guidance; AdaGN = Adaptive Group Normalization; RCL = Region-specific Contrastive Loss; AAS = Anatomically-Aware Sampling; MC-DPM = Multi-Condition Denoising Probabilistic Model. ✓ indicates publicly available code repository.

ibility with all knowledge types (K1–K4) and architectural strategies. This flexibility has made I2 the most prevalent strategy for 3D CT generation [Hamamci *et al.*, 2024a; Guo *et al.*, 2025a; Oliveras *et al.*, 2025; Mahdi *et al.*, 2025; Kui *et al.*, 2025]. *Inference-time integration* (I3) and *joint modeling* (I4) each comprise 10%. A strong coupling between K1 and I3 is observed (e.g., GenerateCT, Report2CT [Amirrajab *et al.*, 2025], Text2CT [Guo *et al.*, 2025a], TRACE [Shao *et al.*, 2025]), while geometric (K2) and exemplar (K3) approaches do not adopt this paradigm. *Pre-generative alignment* (I1) accounts for 6% of existing methods and is currently confined to text-conditioned generation (K1), utilizing transformer-based tokenization and contrastive objectives [Hamamci *et al.*, 2024a; Molino *et al.*, 2025].

Architectural Trends. Figure 2 (top-center) shows that *single-stage latent generation* (A1) accounts for 56% of existing methods. Representative approaches such as MAISI

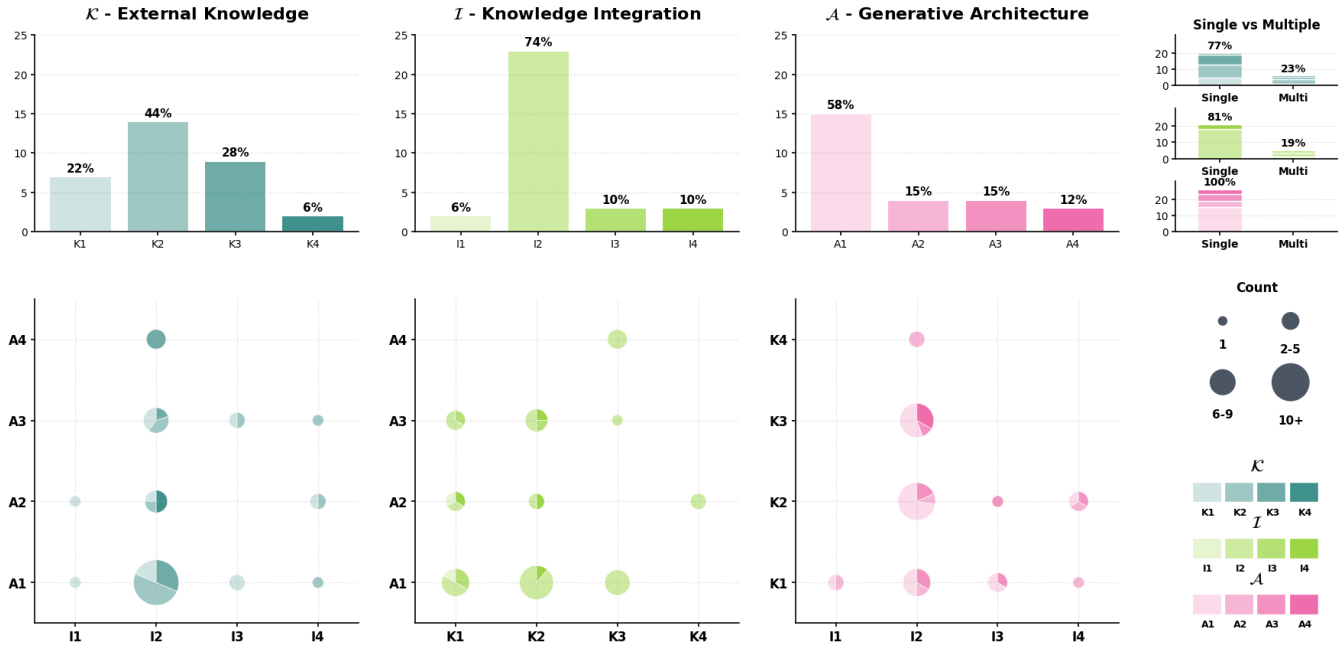

Figure 2: Statistical distribution of methods across the taxonomy. Top row: frequency of each category in axes $\mathcal{K}$, $\mathcal{I}$, and $\mathcal{A}$, plus single versus multiple category usage per axis. Bottom row: pairwise interactions between axes, where node size indicates method count and pie slices show the third axis's distribution (methods using multiple categories appear in multiple counts).

and MAISI-v2 leverage VQ-VAE latent spaces, while LAND, MedLoRD, and DiffTumor adopt similar strategies [Guo *et al.*, 2025b; Zhao *et al.*, 2025; Oliveras *et al.*, 2025; Seyfarth *et al.*, 2025; Chen *et al.*, 2024]. *Cascaded pipelines* (A2) decompose synthesis into coarse-to-fine stages and account for 15% of methods. Representative examples include GenerateCT, MedSyn, and Cascaded-3D [Hamamci *et al.*, 2024a; Xu *et al.*, 2024; Yoon *et al.*, 2025b]. *Autoregressive architectures* (A3) represent 15% of methods and factorize generation along the axial dimension. By treating volumes as ordered slice sequences, approaches such as TRACE and CTFlow model inter-slice dependencies [Shao *et al.*, 2025; Wang *et al.*, 2025a]. *Fixed-representation generation* (A4) accounts for 12% of methods. Representative approaches include cWDM and 3D-WLDM, which operate in wavelet space [Friedrich *et al.*, 2024a; Zheng *et al.*, 2025].

Cross-Dimensional Patterns. The scatter plots in Figure 2 (bottom row) reveal strong dependencies between knowledge type, integration strategy, and architectural design, highlighting both consolidated practices and systematic gaps in the design space. A clear asymmetry emerges along the knowledge–integration axis (Figure 2, K–I). Textual conditioning (K1) consistently co-occurs with inference-time guidance (I3), while geometric (K2), exemplar (K3), and distributional (K4) knowledge rely almost exclusively on direct architectural integration (I2). Architectural preferences further reveal distinct patterns (Figure 2, K–A and I–A). Geometric conditioning (K2) concentrates in single-stage latent generation (A1), forming the most frequent configuration. Exemplar-based methods (K3) distribute across latent (A1) and fixed-representation architectures (A4). Textual conditioning (K1) spans latent (A1) and autoregressive designs (A3), whereas distributional knowledge (K4) appears exclusively in cascaded pipelines (A2). Model-based integration (I2) within single-stage latent models (A1) dominates the integration–architecture space. Taken together, these trends converge on the triplet (K2, I2, A1)—geometric masks with in-process modulation in single-stage latent diffusion—as the prevailing paradigm, exemplified by MAISI, MAISI-v2, NodMAISI, LAND, and MedLoRD [Guo *et al.*, 2025b; Zhao *et al.*, 2025; Tushar *et al.*, 2025; Oliveras *et al.*, 2025; Seyfarth *et al.*, 2025]. Several regions of the design space remain comparatively underrepresented.

5 Design Space Interpretation

The concentration around geometric conditioning (K2) with in-process modulation (I2) in single-stage latent diffusion (A1) reflects pragmatic convergence rather than fundamental optimality. Automated segmentation tools made voxel-aligned priors widely available, while I2's architectural flexibility—enabling cross-attention, concatenation, or adaptive modulation across all knowledge types—made it universally applicable without structural constraints. Geometric masks resolve spatial uncertainty directly, enforcing anatomical correctness while constraining appearance variability. This established geometry as structural backbone, effective for targeted synthesis but limiting semantic diversity and population-aware generation. This convergence reveals three fundamental design patterns.

First, textual conditioning (K1) pairs exclusively with classifier-free guidance (I3) [Hamamci *et al.*, 2024a; Guo *et al.*, 2025a; Shao *et al.*, 2025] because linguistic descriptions exhibit irreducible semantic ambiguity: for instance, "enlarged liver" maps to diverse configurations requiring

inference-time balancing between fidelity and diversity. Geometric conditioning eliminates this need by resolving spatial constraints through voxel-aligned masks, explaining why text enables semantic specification without spatial grounding while geometry ensures precision without appearance flexibility.

Second, pre-generative alignment (I1) remains confined to text-image pairs due to established vision-language frameworks (CLIP, contrastive learning), leaving geometric-demographic (K2–K4) or cross-modal exemplar (K3–K3) alignment unexplored despite potential for population-conditioned scaffolds or registration-free correspondence.

Third, architectural decompositions encode distinct trade-offs: single-stage latent models (A1) depend critically on autoencoder fidelity, manifesting compression artifacts as slice discontinuities; cascaded pipelines (A2) propagate low-resolution errors through refinement stages; autoregressive methods (A3) accumulate sequential prediction drift; fixed-transform approaches (A4) avoid learned artifacts but cluster in MRI→CT translation where exemplar constraints (K3) stabilize wavelet-domain synthesis, is currently inapplicable to abstract conditioning (K1, K2) lacking inherent spatial structure.

Notably, some design space gaps reflect mechanistic constraints rather than oversight. Demographic attributes (K4) are structured and unambiguous (age: 65, sex: female), lacking the variability that motivates inference-time guidance (I3) for textual descriptions—direct feature modulation suffices, rendering K4-I3 combinations redundant. Similarly, fixed-transform generation (A4) with textual conditioning (K1) remains absent because linguistic semantics lack the geometric regularity required for deterministic transform domains.

6 Open Research Directions

Untapped Demographic Potential. Despite convergence on (K2, I2, A1), demographic conditioning (K4) accounts for only 6% of methods yet offers unique advantages: attributes are acquired without segmentation pipelines or learned encoders, enabling direct population-level modeling of age-related atrophy, sex-specific morphology, or pathology prevalence at minimal annotation cost. Combining K4 with geometric scaffolds (K2) through pre-generative alignment (I1) could establish population-conditioned anatomical priors, while K4 integration via adaptive normalization (I2) could modulate global appearance without spatial constraints.

Multimodal Fusion Strategies. Multimodal integration beyond geometry remains nascent. Text-exemplar combinations (K1+K3) could ground semantic descriptions in structural priors, mitigating spatial ambiguity through voxel-aligned constraints. Text-demographic pairing (K1+K4) could enable population-specific semantic generation, addressing limited spatial grounding in pure linguistic conditioning. More fundamentally, joint modeling (I4) warrants extension beyond paired image-mask generation: CT-MRI co-generation could enforce intrinsic structural alignment without explicit registration, while volume-text co-generation could prevent semantic drift inherent in conditional formulations. Combining I4 with hierarchical decompositions (A2) or fixed-transform domains (A4) could enable staged co-generation while controlling complexity.

Architectural Diversification. Architectural exploration reveals untapped potential. Fixed-transform strategies (A4) concentrate in exemplar-driven translation but could extend to coarse geometric conditioning (K2), leveraging wavelet or spectral representations to reduce memory requirements while preserving anatomical consistency without learned compression. Conversely, identifying minimal sufficient geometric priors—treating masks as structural scaffolds while secondary sources (K1, K3, K4) modulate appearance—could substantially reduce annotation overhead. Hybrid integration strategies combining alignment (I1) with model-based conditioning (I2), or I2 with inference-time guidance (I3), remain unexplored despite complementary strengths: I1 for representation coherence, I2 for spatial modulation, I3 for generation-time controllability.

7 Limitations and Conclusions

This survey introduces the first conditioning-centric taxonomy for knowledge-guided 3D CT generation, organizing methods along orthogonal axes—knowledge type (K), integration paradigm (I), and generative architecture (A)—to enable systematic positioning within an interpretable design space. The taxonomy provides a framework for comparing conditioning strategies across heterogeneous approaches, though reliable performance assessment remains constrained by fundamental evaluation limitations.

Indeed, methods rely on non-overlapping datasets with inconsistent protocols: many remain proprietary, while publicly available data undergo arbitrary resolution adjustments, confounding conditioning strategy contributions with dataset-specific effects. More critically, no standardized validation pipeline exists. These constraints prevent not only quantitative comparison within this survey but also reliable cross-work assessment in the literature itself—methods are rarely compared directly, obscuring which conditioning strategies, integration mechanisms, or architectural decompositions prove most effective and under what metrics. Even when numerical results are reported, metrics are applied non-uniformly: Fréchet inception distance uses different feature extractors (ImageNet versus RadImageNet), CLIP evaluations employ distinct encoders, while distributional and clinical validity assessments lack consensus on combination or sufficiency thresholds. Conditioning claims are often accepted without systematic ablation studies, and computational costs remain disconnected from quality improvements, preventing principled trade-off analysis.

The field's rapid evolution (2023–2025) further complicates assessment: overlapping approaches—text-based versus report-based conditioning, ControlNet versus cross-attention integration—proliferate without principled comparison, obscuring incremental progress and leaving fundamental design questions unresolved. Despite these limitations, organizing literature into K×I×A design space reveals actionable opportunities and enables identification of concrete research directions, positioning future work toward cumulative refinement beyond continued proliferation of incomparable variants.

Ethical Statement

There are no ethical issues.

Acknowledgments

This paper is supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence).

References

[Amirrajab *et al.*, 2025] Sina Amirrajab, Zohaib Salahuddin, Sheng Kuang, Henry C Woodruff, and Philippe Lambin. Radiology report conditional 3d ct generation with multi encoder latent diffusion model. *arXiv preprint arXiv:2509.14780*, 2025.

[Chen and Ramsey, 2024] Kaiqi Chen and Libby Ramsey. Deep generative models for 3d content creation: A comprehensive survey of architectures, challenges, and emerging trends. 2024.

[Chen *et al.*, 2024] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11147–11158, 2024.

[Dorjsembe *et al.*, 2024] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024.

[Friedrich *et al.*, 2024a] Paul Friedrich, Alicia Durrer, Julia Wolleb, and Philippe C Cattin. cwdm: Conditional wavelet diffusion models for cross-modality 3d medical image synthesis. *arXiv preprint arXiv:2411.17203*, 2024.

[Friedrich *et al.*, 2024b] Paul Friedrich, Yannik Frisch, and Philippe C Cattin. Deep generative models for 3d medical image synthesis. In *Generative Machine Learning Models in Medical Image Computing*, pages 255–278. Springer, 2024.

[Guo *et al.*, 2025a] Pengfei Guo, Can Zhao, Dong Yang, Yufan He, Vishwesh Nath, Ziyue Xu, Pedro RAS Bassi, Zongwei Zhou, Benjamin D Simon, Stephanie Anne Harmon, et al. Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model. *arXiv preprint arXiv:2505.04522*, 2025.

[Guo *et al.*, 2025b] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE, 2025.

[Hamamci *et al.*, 2024a] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024.

[Hamamci *et al.*, 2024b] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024.

[Han *et al.*, 2023] Kun Han, Yifeng Xiong, Chenyu You, Pooya Khosravi, Shanlin Sun, Xiangyi Yan, James S Duncan, and Xiaohui Xie. Medgen3d: A deep generative framework for paired 3d image and mask generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 759–769. Springer, 2023.

[Jiang *et al.*, 2025] Yifan Jiang, Yannick Lemaréchal, Josée Bafaro, Jessica Abi-Rjeile, Philippe Joubert, Philippe Després, and Venkata Manem. Lung-ddpm: Semantic layout-guided diffusion models for thoracic ct image synthesis. *arXiv preprint arXiv:2502.15204*, 2025.

[Khader *et al.*, 2023] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.

[Koetzier *et al.*, 2024] Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, 2024.

[Kui *et al.*, 2025] Xiaoyan Kui, Bo Liu, Zanbo Sun, Qinsong Li, Min Zhang, Wei Liang, and Beiji Zou. Med-lvdm: Medical latent variational diffusion model for medical image translation. *Biomedical Signal Processing and Control*, 106:107735, 2025.

[Liu *et al.*, 2024] Yanbin Liu, Girish Dwivedi, Farid Boussaid, and Mohammed Bennamoun. 3d brain and heart volume generative models: a survey. *ACM Computing Surveys*, 56(6):1–37, 2024.

[Lomurno and Matteucci, 2025] Eugenio Lomurno and Matteo Matteucci. Federated knowledge recycling: Privacy-preserving synthetic data sharing. *Pattern Recognition Letters*, 191:124–130, 2025.

[Mahdi *et al.*, 2025] Mohammed A Mahdi, Mohammed Al-Shalabi, Ehab T Alnfrawy, Reda Elbarougy, Muhammad Usman Hadi, and Rao Faizan Ali. 3d latent diffusion model for mr-only radiotherapy: Accurate and consistent synthetic ct generation. *Diagnostics*, 15(23):3010, 2025.

[Molino *et al.*, 2025] Daniele Molino, Camillo Maria Caruso, Filippo Ruffini, Paolo Soda, and Valerio Guarrasi. Text-to-ct generation via 3d latent diffusion model with contrastive vision-language pretraining. *arXiv preprint arXiv:2506.00633*, 2025.

[Oliveras *et al.*, 2025] Anna Oliveras, Roger Marí, Rafael Redondo, Oriol Guardià, Ana Tost, Bhalaji Nagarajan, Carolina Migliorelli, Vicent Ribas, and Petia Radeva. Land: Lung and nodule diffusion for 3d chest ct synthesis with anatomical guidance. *arXiv preprint arXiv:2510.18446*, 2025.

[Pan *et al.*, 2024] Shaoyan Pan, Elham Abouei, Jacob Wynne, Chih-Wei Chang, Tonghe Wang, Richard LJ Qiu, Yuheng Li, Junbo Peng, Justin Roper, Pretesh Patel, et al. Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. *Medical Physics*, 51(4):2538–2548, 2024.

[Seyfarth *et al.*, 2025] Marvin Seyfarth, Salman Ul Hassan Dar, Isabelle Ayx, Matthias Alexander Fink, Stefan O Schoenberg, Hans-Ulrich Kauczor, and Sandy Engelhardt. Medlord: A medical low-resource diffusion model for high-resolution 3d ct image synthesis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 1–12. Springer, 2025.

[Shao *et al.*, 2025] Minye Shao, Xingyu Miao, Haoran Duan, Zeyu Wang, Jingkun Chen, Yawen Huang, Xian Wu, Jingjing Deng, Yang Long, and Yefeng Zheng. Trace: Temporally reliable anatomically-conditioned 3d ct generation with enhanced efficiency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 627–637. Springer, 2025.

[Tapp *et al.*, 2024] Austin Tapp, Abhijeet Parida, Can Zhao, Van Lam, Natasha Lepore, Syed Muhammad Anwar, and Marius George Linguraru. Mr to ct synthesis using 3d latent diffusion. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[Tushar *et al.*, 2025] Fakrul Islam Tushar, Ehsan Samei, Cynthia Rudin, and Joseph Y Lo. Nodmaisi: Nodule-oriented medical ai for synthetic imaging. *arXiv preprint arXiv:2512.18038*, 2025.

[Wang *et al.*, 2025a] Jiayi Wang, Hadrien Reynaud, Franciskus Xaverius Erick, and Bernhard Kainz. Ctflow: Video-inspired latent flow matching for 3d ct synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6750–6758, 2025.

[Wang *et al.*, 2025b] Lu-Yan Wang, Tzung-Dau Wang, and Shang-Hong Lai. Labelg : Consistent pairwise 3d CT image and segmentation mask generation via medical foundation model. In *Submitted to Medical Imaging with Deep Learning*, 2025. under review.

[Wu *et al.*, 2025] Jing Wu, Yuli Wang, Zhusi Zhong, Weihua Liao, Natalia Trayanova, Zhicheng Jiao, and Harrison X Bai. Vision-language foundation model for 3d medical imaging. *npj Artificial Intelligence*, 1(1):17, 2025.

[Xu *et al.*, 2024] Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afrooz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich. Medsyn: text-guided anatomy-aware synthesis of high-fidelity 3-d ct images. *IEEE Transactions on Medical Imaging*, 43(10):3648–3660, 2024.

[Yoon *et al.*, 2025a] Siyeop Yoon, Yujin Oh, Pengfei Jin, Sifan Song, Matthew Tivnan, Dufan Wu, Xiang Li, and Quanzheng Li. Surf2ct: Cascaded 3d flow matching models for torso 3d ct synthesis from skin surface. *arXiv preprint arXiv:2505.22511*, 2025.

[Yoon *et al.*, 2025b] Siyeop Yoon, Sifan Song, Pengfei Jin, Matthew Tivnan, Yujin Oh, Sekeun Kim, Dufan Wu, Xiang Li, and Quanzheng Li. Cascaded 3d diffusion models for whole-body 3d 18-f fdg pet/ct synthesis from demographics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–109. Springer, 2025.

[Yu *et al.*, 2024] Yongrui Yu, Hanyu Chen, Zitian Zhang, Qiong Xiao, Wenhui Lei, Linrui Dai, Yu Fu, Hui Tan, Guan Wang, Peng Gao, et al. Ct synthesis with conditional diffusion models for abdominal lymph node segmentation. *arXiv preprint arXiv:2403.17770*, 2024.

[Zhao *et al.*, 2025] Can Zhao, Pengfei Guo, Dong Yang, Yucheng Tang, Yufan He, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. *arXiv preprint arXiv:2508.05772*, 2025.

[Zheng *et al.*, 2025] Jiaxu Zheng, Meiman He, Xuhui Tang, Xiong Wang, Tuoyu Cao, Tianyi Zeng, Lichi Zhang, and Chenyu You. 3d wavelet latent diffusion model for whole-body mr-to-ct modality translation. *arXiv preprint arXiv:2507.11557*, 2025.

[Zhou *et al.*, 2025] Xuanru Zhou, Cheng Li, Shuqiang Wang, Ye Li, Tao Tan, Hairong Zheng, and Shanshan Wang. Generative artificial intelligence in medical imaging: Foundations, progress, and clinical translation. *arXiv preprint arXiv:2508.09177*, 2025.

[Zhu *et al.*, 2024] Lingting Zhu, Noel Codella, Dongdong Chen, Zhenchao Jin, Lu Yuan, and Lequan Yu. Generative enhancement for 3d medical images. *arXiv preprint arXiv:2403.12852*, 2024.