Sketch2CT: Multimodal Diffusion for Structure-Aware 3D Medical Volume Generation

Supplementary Material

A. Structured Semantic Text Description

To obtain geometry-focused textual descriptions for multimodal conditioning, we provide GPT-4o-mini with (1) three canonical 3D renderings of each organ (axial, sagittal, and coronal views), and (2) a carefully designed prompt instructing the model to extract purely geometric information. Examples of the text descriptions are presented in Figure A1.

Prompt design. GPT-4o-mini is directed to serve as an expert in the geometric interpretation of anatomical 3D structures. The prompt emphasizes strict geometric reasoning and forbids any clinical or physiological interpretation. In summary, the model is tasked with:

- analyzing the organ's 3D geometry using the three provided views;
- describing the volumetric shape, dominant axes, convexity, and proportions;
- characterizing surface morphology, including curvature patterns, smoothness, ridges, protrusions, and indentations;
- assessing bilateral or rotational symmetry and quantifying asymmetry;
- identifying high-level topological traits such as cavities, branching, or major geometric landmarks; and
- providing additional structural features that could assist downstream geometry-aware diffusion models.

The prompt explicitly limits the model to geometric and morphological observations, instructing it to avoid using medical terminology, diagnoses, clinical relevance, or biological functions. The aim is to produce a text description that focuses solely on the structural characteristics of the input 3D shape.

Output structure. GPT-4o responds with a structured description organized into the following conceptual components:

- an identifier for the input organ instance;
- a summary of global volumetric shape and coarse structural proportions;
- a characterization of surface-level geometric features;
- an analysis of symmetry or asymmetry patterns;
- a summary of key topological features; and
- high-level geometric guidance relevant for shape-conditioned generative modeling.

Although internally expressed in a structured format, we use only the distilled textual content for multimodal conditioning in the Sketch2CT framework.

B. Sketch Granularity Analysis

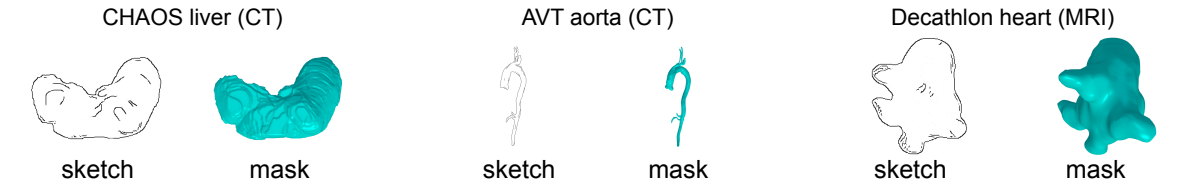
To investigate how sketch detail influences the quality of generated segmentation masks, we conduct a controlled analysis using sketches derived from 2D organ snapshots. As described in the main paper, sketch contours are generated using an edge-based extractor with a sensitivity parameter ranging from 0 to 10. Lower values produce sparse and coarse structural outlines, while higher values generate denser sketches with fine-grained details.

We select three representative parameter values across this range to create sketches with increasing levels of granularity. These sketches are then input into our segmentation generation module, while maintaining consistency with all other model components and conditions. The results are shown in Figure A2. Our results reveal a clear trend:

- **Coarse sketches** (low parameter settings) yield segmentation masks with smooth, global surfaces and minimal small-scale fluctuations. The generated structures faithfully capture the overall anatomical form, with stable topology and consistent volumetric shape.
- **Detailed sketches** (high parameter settings) introduce additional local variations, resulting in segmentation masks with slightly rougher or more irregular surface patterns. These high-detail sketches accurately capture fine contour variations and transfer them into the predicted mask.

Despite the surface-level differences, the global anatomical fidelity remains highly consistent across various levels of detail. The final reconstructed 3D medical volumes derived from these segmentation masks are visually similar, and quantitative metrics indicate minimal variation across different levels of sketch detail. In particular, the Dice score between coarse and medium sketches fluctuates within 0.02 ± 0.01. In contrast, the Dice difference between coarse and fine sketches remains similarly small at approximately 0.06 ± 0.02, confirming that increased sketch detail leads to only marginal changes in segmentation quality.

Considering both performance and computational efficiency, we recommend using sketches with low to moderate granularity. Coarse sketches offer sufficient structural guidance for the multimodal diffusion model while minimizing unnecessary local noise and reducing preprocessing overhead. This supports a practical use case where users can provide simple, clean sketches without compromising the quality of downstream generation.



global shape: The structure forms a broad, horizontally oriented mass composed of two major volumetric lobes connected by a shallower central region. Its footprint is wide and flattened, with the dominant axes extending laterally. The left portion appears thicker and more convex, while the right portion displays a more elongated and tapered contour.
surface characteristics: The surface exhibits a mix of smooth convex regions and numerous layered, terrace-like undulations that wrap across the volume.
symmetry: The shape is strongly asymmetric. The two large lobes differ in size, thickness, and curvature distribution. The right region has more pronounced layered striations, while the left region is smoother and more globular, reinforcing the asymmetric mass distribution.
topology: The topology is that of a single connected solid without holes or branching protrusions. It contains multiple shallow depressions, internal-looking concave surfaces, and a central notch-like indentation separating the two primary volumetric lobes.

global shape: An elongated, curved tubular structure with a dominant vertical axis and a pronounced C-shaped bend near the upper region. The main volume tapers gradually from a thicker proximal segment to a narrower distal end. Several smaller offshoots extend laterally from the main tube.
surface characteristics: Surface appears smooth and continuously curved with no sharp ridges. The upper curved segment shows broad, gentle convexity, while the descending portion exhibits a consistent cylindrical profile.
symmetry: The structure is strongly asymmetric. The primary curvature and the lateral branch orientations both deviate from bilateral symmetry. No axial or rotational symmetry is apparent due to the varied placement and orientation of side branches.
topology: Topologically a single connected tubular form with multiple smaller branches extending from the main path. No cavities or closed loops are visible. The topology is tree-like, with one dominant trunk and several secondary bifurcations.

global shape: The structure has an overall compact, bulbous mass with several elongated protrusions extending outward in different directions. The main body is rounded and slightly tapered toward its lower end, while three major extensions emerge asymmetrically from its mid-to-upper regions.
surface characteristics: Protrusions blend gradually into the main body with soft curvature transitions. The extended arms vary slightly in thickness and curvature, some bending gently while others remain nearly straight. The main mass exhibits gentle convexity with mild surface undulations but no prominent indentations.
symmetry: The structure is highly asymmetric. The protrusions differ in length, direction, and placement, resulting in no bilateral or rotational symmetry.
topology: It is a single connected solid with multiple outward extensions, but no holes, cavities, or branching bifurcations that split into sub-branches. The global topology is compact with several single-branch appendage-like forms.

Figure A1. Examples of textual geometry descriptions used as the text-based conditioning input in Sketch2CT. For each organ, we show the input sketch, the generated mask, and the corresponding structured text description that captures the global shape, surface characteristics, symmetry, and topology.

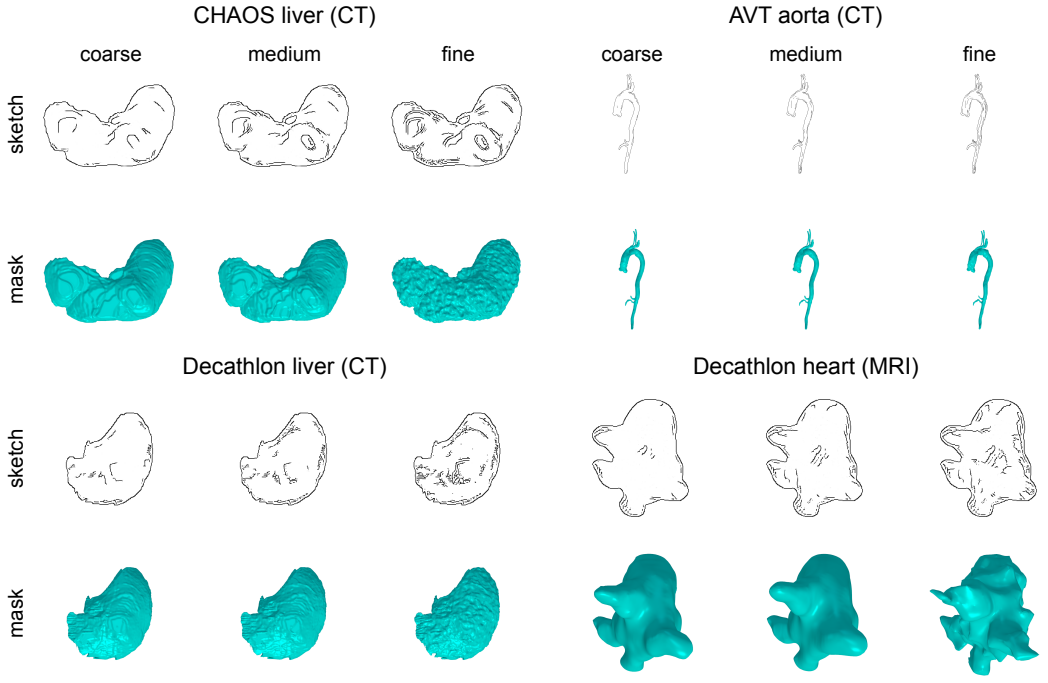


Figure A2. Effect of sketch granularity on segmentation mask generation. For each dataset, we vary the sketch detail in three levels, coarse, medium, and fine, and visualize the corresponding 3D masks produced by Sketch2CT. Increased sketch detail introduces more local structural variations, while the overall anatomical geometry remains consistent across granularity levels.

C. Ablation Study

To evaluate the contribution of the core components in Sketch2CT, we conduct an ablation study aligned with the modules defined in the main framework. Specifically, we examine the impact of removing: (1) TSFE, which refines sparse sketch embeddings via text-guided FiLM modulation; (2) CGFM, which performs global semantic alignment through hierarchical cross- and self-attention; and (3) the segmentation latent diffusion model, which reconstructs coherent 3D masks in latent space under multimodal conditioning. For each ablated variant, synthetic images are generated and used to train the same downstream segmentation network [18, 44] as in the main experiments. Table A1 reports Dice scores on real test sets. Removing any single module leads to a consistent decrease in performance across all datasets, demonstrating that all three components play complementary roles in ensuring accurate multimodal alignment and anatomically faithful 3D mask generation. The full model achieves the best performance across all benchmarks.

	CHAOS liver (CT)	AVT aorta (CT)	Decathlon liver (CT)	Decathlon heart (MRI)
full model	**0.893**	**0.889**	**0.904**	**0.711**
w/o TSFE	0.864	0.859	0.872	0.683
w/o CGFM	0.825	0.818	0.831	0.671
w/o Seg-LDM	0.642	0.629	0.633	0.545

Table A1. Ablation study of Sketch2CT. Removing TSFE, CGFM, or the segmentation latent diffusion model (Seg-LDM) reduces downstream segmentation performance, confirming that all components are essential for generating anatomically coherent and text-aligned 3D masks.

We further analyze the individual contributions of sketches and text by evaluating two additional variants: sketch-only and text-only conditioning. As shown in Figure A3, relying solely on sketches leads to incomplete or distorted 3D structures due to the inherent limitations of single-view contours, which lack depth information and volumetric context. Conversely, using text alone removes spatial constraints entirely, resulting in incorrect global shape, misplaced structures, and anatomically implausible geometries. These observations highlight that sketches and text provide complementary forms of guidance. Sketches anchor the spatial structure [21, 33, 64], while text provides semantic and morphological context, both of which are essential for accurate and stable 3D mask generation.

D. Qualitative Evaluation of Generative Diversity

To qualitatively demonstrate the diversity of the synthesized volumes generated by Sketch2CT, we repeat the generation process three times under the same sketch and text conditions. As shown in Figure A4, each run produces anatomically consistent 3D segmentation masks that follow the shared multimodal conditioning, while the synthesized CT volumes exhibit natural variability in texture, intensity distribution, and fine-scale appearance. This stochasticity reflects the inherent randomness of the diffusion process, enabling Sketch2CT to produce diverse yet structurally faithful volumetric data. The qualitative results are consistent with the quantitative findings reported in the main paper, indicating that Sketch2CT maintains stable geometry while supporting realistic appearance-level variation, which is valuable for creating diverse synthetic datasets.

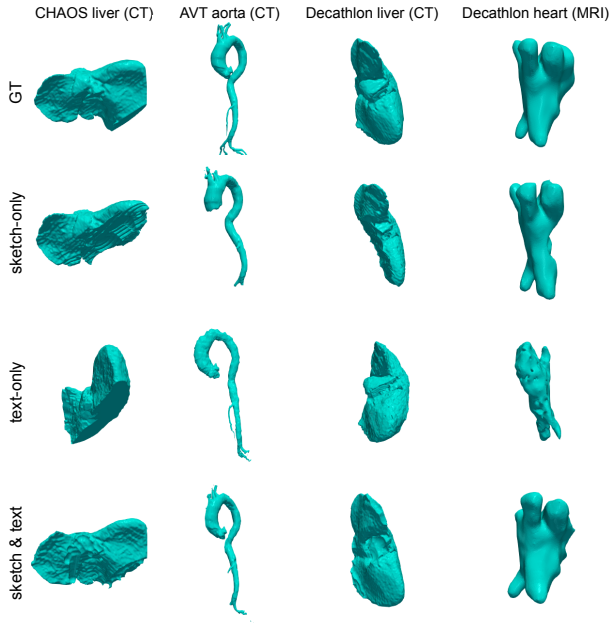


Figure A3. Comparison of segmentation masks generated using sketch-only and text-only conditions. Sketch-only guidance fails to recover full 3D geometry due to single-view ambiguity. In contrast, text-only guidance yields incorrect global shape and spatial placement, underscoring the need to combine sketches and text.

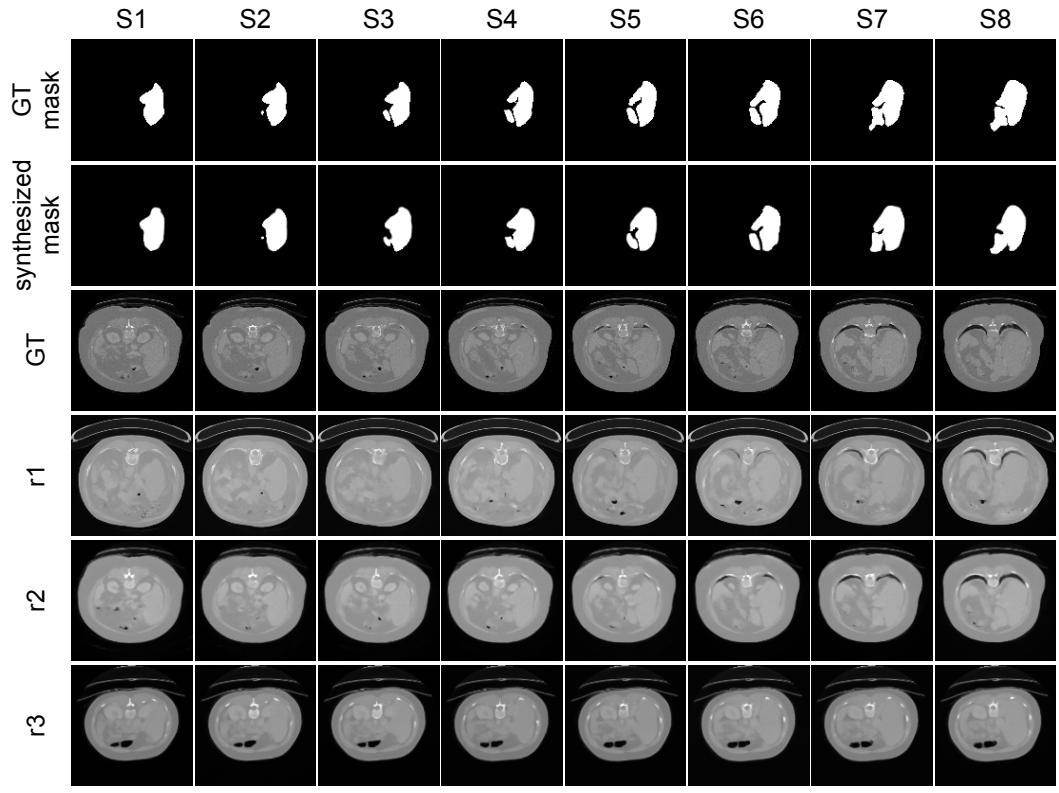


Figure A4. Qualitative diversity demonstration using a single Decathlon liver case. Under identical sketch and text conditions, three independent runs (r1-r3) generate anatomically consistent yet appearance-varying CT volumes, illustrating the stochastic diversity of Sketch2CT.