id	question	evidence	analysis /compared with	note
RQ0	benchmark validity: noise, real-versus-real reference, which metrics discriminate, redundancy	all F (3 references), TSTR	rand vs rand_b (mid, max); real; Friedman	
RQ1	**ranking of every method in every slice**	all leaves: F × reference, U × split, cost	all methods	leaderboard
RQ2	assembly from a window pool vs generators of long series	F (val, OOT), TSTR all tasks	long gen. vs rand, assembly median	
RQ3	inside assembly: window selection × seam processing	F (val, OOT), TSTR all tasks	vs random path; seam-processing contrasts	
RQ4	does fidelity predict utility (which metrics, which reference)	F (3 references) ↔ transfer ratio	Spearman across methods	
RQ5	gap to real data; stability over time and model class; decision transfer	TSTR / TRTR, U6	real (TRTR); Kendall τ	(c) max only
RQ6	speed and cost: training, generation, memory; quality–time Pareto	time, memory	all methods	
RQ7	moderators: dataset properties and family differences	dataset descriptors	effects of RQ2, RQ3	explor.
RQ8	class balance: shares of every classification task, rare events	F8.1, F11; TSTR ECE	vs random path; long gen. vs assembly median	
RQ9	extremes: tails, outliers, block maxima / minima	F1 tails, F2.4, F7.4, F12; extreme error, rare PR AUC	vs random path; long gen. vs assembly median	
C1–C4	**confirmatory: methods differ; long vs assembly; selection; seam processing**	U2.1 AUC, U2.6 MASE (TSTR, linear, OOT)	Friedman; assembly median; contrasts	confirm.
CS1–CS15	case study of one method (default vae_zseries) vs its controls and alternatives	F (val), TSTR	vae_rand, A, B, C	secondary

Evidence colour: utility TSTR/TRTR (green), cost and descriptors (grey), fidelity and utility together (white). F = fidelity metric groups (Fig. 12); A, B, C = method groups of Fig. 8.
Confirmatory family: joint Holm over C1–C4 in both cells. Other questions: BH within a question; RQ7 exploratory. The case study is secondary.