photonics

MDPI

Article

Reproducible Simulation Benchmark of Hybrid Interferometric Profilometry with Coincidence Proxy Priors on Measured Rough Surfaces

Dawid Kucharski

Division of Metrology and Measurement Systems, Institute of Mechanical Technology, Faculty of Mechanical Engineering, Poznan University of Technology, 60-965 Poznan, Poland; dawid.kucharski@put.poznan.pl

Abstract

This paper presents a reproducible simulation benchmark for rough surface interferometric profilometry. The benchmark compares three complete reconstruction pipelines under matched detected count assumptions: classical four-step phase-shifting interferometry (PSI), direct coincidence proxy reconstruction, and hybrid coarse-to-fine reconstruction in which a coincidence-derived observable supplies the coarse fringe-order prior. Fifty-nine focus variation (FV) topographies exported as Mountains/DigitalSurf `.sur` files (Digital Surf, Besancon, France) provide a shared FV prior for simulated optical observations. The coincidence channel is a simulation proxy rather than a validated quantum hardware implementation. The main result is architectural role separation. On the measured surface benchmark, the hybrid branch gives the lowest median detrended height RMSE (314.0 nm) and wins on 32 of 59 surfaces. The same ordering is retained in a rate-based coincidence control, with median hybrid RMSE of 290.9 nm under ideal matched-count rates and 376.3 nm under detector non-idealities. Roughness endpoints define the boundary of this result: hybrid gives the lowest matched bandwidth S_a and S_q errors, whereas direct coincidence proxy reconstruction is selectively strongest for S_z and remains process-dependent. Classical two-colour and classical frontier controls show that following the broad long-wavelength envelope is not sufficient evidence for overall architecture-level superiority within this simulation benchmark. The benchmark identifies coincidence-derived information as most useful when used as a coarse prior inside a hybrid estimator, while final fine texture remains anchored by short-wavelength PSI.

Keywords: coincidence proxy interferometry; simulation benchmark; surface profilometry; hybrid reconstruction; coarse-to-fine reconstruction; phase-shifting interferometry; synthetic wavelength; roughness metrology; focus variation microscopy; areal surface texture

Received: 8 May 2026
Revised: 25 May 2026
Accepted: 26 May 2026
Published: 28 May 2026

1. Introduction

Surface topography and texture must be quantified across a wide range of manufacturing and functional applications, which is why optical methods continue to attract interest for on-line and in-process measurement [1,2]. Practical non-contact routes include laser-scatter surface finish measurement [3] and interferometric surface metrology deployed in industrially relevant settings [4]. Recent structured-light Doppler metrology provides a related example of the same photonic measurement design pressure; tailored optical fields can improve the measured observable or operating range, but endpoint-level performance must still be tested under the relevant propagation and noise conditions [5,6]. In areal

surface metrology, the reported descriptors, filtering operations, instrument classes, and uncertainty statements are governed by established geometrical product specification and uncertainty frameworks [7–10].

Phase-shifting interferometry (PSI) remains one of the standard high-precision optical routes for surface measurement, but practical deployment is limited by environmental sensitivity and non-ideal stepping in real optical testing [11]. Several strategies have therefore been developed to stabilise or regularise PSI, including spatiotemporal phase-shifting methods [12], full-field least-squares phase retrieval [13], extended multi-frame algorithms that trade speed for robustness [14], and calibration-oriented treatments for highly reflective surfaces [15].

Signal conditioning for phase-stepped interferometric processing has likewise been studied through filtering and image enhancement [16]. More recent work has extended that effort into data-driven virtual phase shifting for surface profiling [17], robust interferometric thickness sensing [18], broader surveys of digital interferometry for surface relief [19], and two-colour phase-modulating interferometry for distance measurement with refractive-index self-correction [20]. In the ultra-low-light regime, PSI has also been analysed from the viewpoint of photon statistical precision limits and phase–transmittance trade-offs [21].

Quantum and quantum-inspired interferometric protocols have been proposed to extend sensitivity or operating range by exploiting nonclassical light [22,23]. Related precision-sensing perspectives include weak-value amplification [24], fourth-order interference with photon-pair coincidences [25], and coincidence-based entangled photon interferometry [26]. Nonclassical enhancement has also been explored in angular displacement metrology [27], lossy interferometry [28], and displacement measurement with position-entangled photon pairs [29]. On the implementation side, integrated sources for entangled qubit/qudit states [30], thin-film lithium niobate quantum photonics [31], and detector calibration with correlated photons [32] all point toward a credible hardware pathway.

For rough surface profilometry, however, the unresolved question is not whether coincidence-derived observables can alter phase sensitivity in idealised settings. Rather, the practical metrology question is whether such observables improve the reconstructed surface or the areal texture parameters reported in an engineering measurement. End-to-end comparisons between classical PSI and coincidence-based channels are still rarely reported at the level of reconstructed surface accuracy under explicitly matched photon/count budgets, particularly for realistic engineering textures rather than only idealised synthetic objects [19,20]. Consequently, the literature does not yet provide a clear endpoint-level map indicating when a long-effective-wavelength channel should carry the reconstruction and when it should serve only to resolve ambiguity for a short-wavelength PSI branch. Robustness requirements in interferometric optical testing therefore motivate comparisons that go beyond idealised phase sensitivity alone [11]. Phase-step deviations and time-varying disturbances motivate evaluation under non-ideal stepping and drift [12], while photon statistical precision limits in PSI require explicit accounting of the photon/count budget when comparing architectures and operating conditions [21]. The missing piece, therefore, is a benchmark that separates architectural roles and identifies the conditions under which a coincidence-derived channel merits experimental follow-up. FV-seeded simulation can make that next experiment better targeted by identifying which branch, endpoint, and limiting case should be tested first.

1.1. Contribution

A reproducible simulation benchmark is presented for rough surface profilometry in which a classical Michelson-type four-step PSI channel and selected coincidence proxy channels are evaluated under matched resource assumptions, matched non-idealities, and

a shared FV prior. The novelty is the controlled role separation of the same coincidence-derived observable when it is used directly versus when it is used as a coarse prior inside a hybrid estimator. The study makes four contributions relative to the surrounding interferometric literature. First, it compares complete reconstruction pipelines under explicitly matched detected count assumptions rather than relying on isolated sensitivity arguments alone. Second, it evaluates those pipelines on $n = 59$ FV topographies spanning six material groups and ten treatment classes, with area-averaged benchmark grid reduction used consistently for the forward model and primary height and roughness endpoints. Third, it separates direct coincidence proxy reconstruction from hybrid coarse-to-fine use of the same observable, which is the practical architectural question for photonic instrumentation: whether the coincidence-derived channel should carry the final map or provide a coarse ambiguity-resolving prior to a classical fine-texture branch. Fourth, it ties the conclusion to controls that matter operationally, including matched bandwidth roughness bias, roughness filter sensitivity, material and treatment holdout stability, limiting-case incidence, paired effects, spectral texture fidelity, classical two-colour baselines, a classical frontier oracle, and a rate-based coincidence model with detector non-idealities.

Together, these elements define a controlled architecture screening benchmark. The direct coincidence proxy channel is evaluated, where it improves the target endpoint, while the hybrid architecture is assessed where the proxy observable provides coarse or absolute-height information and the short-wavelength PSI branch preserves fine texture. Because the FV topographies provide a shared FV prior for simulated optical observations, the measured surface branch serves as an architecture screen. The reported evidence includes detrended height RMSE, roughness parameter biases, bootstrap uncertainty bounds, tolerance exceedance summaries, holdout stability checks, per-surface dominance counts, Holm-adjusted paired tests, paired-effect summaries, spectral fidelity comparisons, and a structured failure taxonomy. The benchmark output therefore defines an experimental prioritisation map with a reduced subset.

1.2. Benchmark Claim

The resulting claim is a reproducible architecture screening result. Within a fixed simulation hierarchy seeded by measured FV topographies, coincidence-derived information plays its strongest role as a coarse, ambiguity-resolving prior in a hybrid estimator, while the short-wavelength PSI branch carries the final fine texture. The manuscript therefore provides a photonic instrumentation architecture screen and experimental priority map for hardware follow-up.

2. Materials and Methods

2.1. Measured Surfaces (Focus Variation Microscopy)

Real rough surfaces are provided as measured height maps exported from focus variation (FV) microscopy in Mountains/DigitalSurf `.sur` format (Digital Surf, Besancon, France) [9]. The measured benchmark comprises $n = 59$ unique surfaces spanning six material groups (stainless steel 1.4301, aluminium 7075, steel C45, brass, titanium, and graphite) and ten treatment classes (six surfaces per treatment except burnishing, $n = 5$). The `.sur` grids can be very large; therefore, two representations are used in the benchmark. First, a fixed, downsampled benchmark grid (default 256×256), obtained by area-averaged decimation from the same measurement, is used as the common input surface for the forward interferometer simulations and as the primary manuscript-facing reference for height RMSE and matched bandwidth roughness. The 256×256 default was chosen as a computationally tractable full-dataset grid that still retains enough lateral samples for pointwise RMSE, areal roughness, and PSD diagnostics; the dependence on this choice

is tested explicitly at 128 × 128, 256 × 256, 384 × 384, and 512 × 512. Second, the native-resolution FV grid is retained as a diagnostic roughness stress test after explicit invalid masking and best-fit-plane removal. Area averaging is therefore a deterministic low-pass box-filtering step: it suppresses aliasing more effectively than raw index subsampling while keeping the simulation workload tractable, and it changes the benchmark grid PSD and roughness relative to the native FV export. The matched bandwidth comparison makes that benchmark grid–native grid distinction explicit. Because both the simulator input surface and the main comparison surface are derived from the same FV export, the measured surface branch acts as an internal consistency benchmark under a shared FV prior for architecture screening. The native lateral grids span 6941–13,884 by 6944–13,884 pixels over 0.38–0.76 mm fields of view, corresponding to the native lateral sampling of 0.03–0.11 µm in both axes. These sampling intervals define the stored grid spacing and should not be read as the optical lateral resolution of the FV instrument. The encoded height unit is consistently nm, while storage precision varies between 2 and 4 bytes per point across the exported files. Full SUR-header metadata and group counts are provided in the Supplementary Materials, Tables S1 and S2; the instrument settings that were not encoded in the available SUR headers are not inferred.

The FV export is used as an operational benchmark surface with a single reproducible shared prior across all compared reconstructions. Within the paper workflow, no additional ISO-style S/L filtering or cross-instrument calibration is imposed beyond explicit invalid sentinel masking and best-fit-plane removal on the native export. This design keeps the comparison internally consistent and makes absolute roughness values, especially S_z, interpretable in the context of FV bandwidth and outlier sensitivity. For that reason, the paper treats the matched bandwidth comparison referenced to the same downsampled benchmark grid used by the forward model as the primary roughness ranking, while the native grid descriptor comparison is retained as a reference sensitivity stress test.

The optical principle of FV microscopy used here is illustrated in Figure 1. Illumination is directed through the objective onto the rough surface, while the returned light is relayed to the camera. During the axial scan, a stack of images is acquired as the objective moves along z, relative to the sample, so that different surface heights pass through the focal plane. A local sharpness or contrast measure is then evaluated as a function of axial position, and the height assigned to each lateral pixel is taken from the z position of the maximum focus response. The exported `.sur` map therefore encodes the reconstructed topography from that FV scan and is used here as the reference surface.

Data collection was performed on machined sample coupons mounted in a dedicated holder; an example photograph is provided in the Supplementary Materials, Figure S1. The photographed specimens are nominally 40 × 40 × 20 mm blocks. During acquisition, each exposed surface was positioned under the FV microscope and scanned to produce a topography map, and was then exported in Mountains/DigitalSurf `.sur` format for the later reconstruction benchmark.

A representative measured benchmark surface is saved as both a 2D height map and a 3D rendering (Figure 2). In other words, the FV measurement is not compared to another independent optical measurement here; rather, it provides the shared benchmark surface from which simulated classical and coincidence proxy interferometric observations are generated. For this reason, the reduced experimental validation subset is framed as a follow-up subset that must be measured against an independent reference chain, such as calibrated coherence-scanning interferometry or a suitable stylus/AFM/areal reference route for the selected surface class.

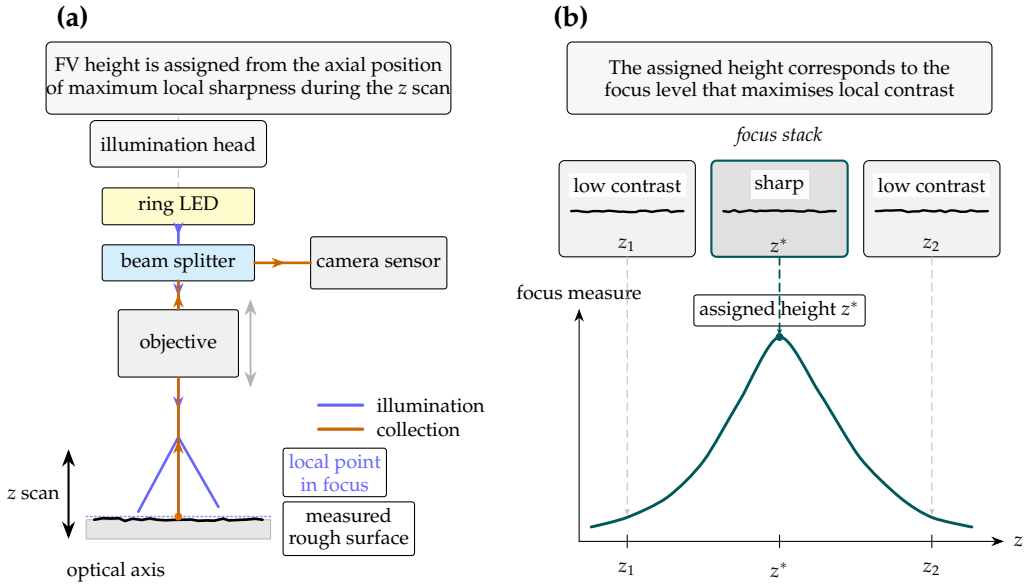

Figure 1. Schematic principle of focus variation (FV) microscopy used for the measured surface reference data. (**a**) Simplified optical head: illumination is sent through the objective to the rough sample, the returned light is separated and relayed to the camera, and the objective is scanned along z relative to the sample. (**b**) Height reconstruction principle: an image stack is acquired at different axial positions (z_1, z^*, z_2), a local sharpness measure is evaluated for each surface location, and the assigned height corresponds to the axial position z^*, at which the sharpness reaches its maximum.

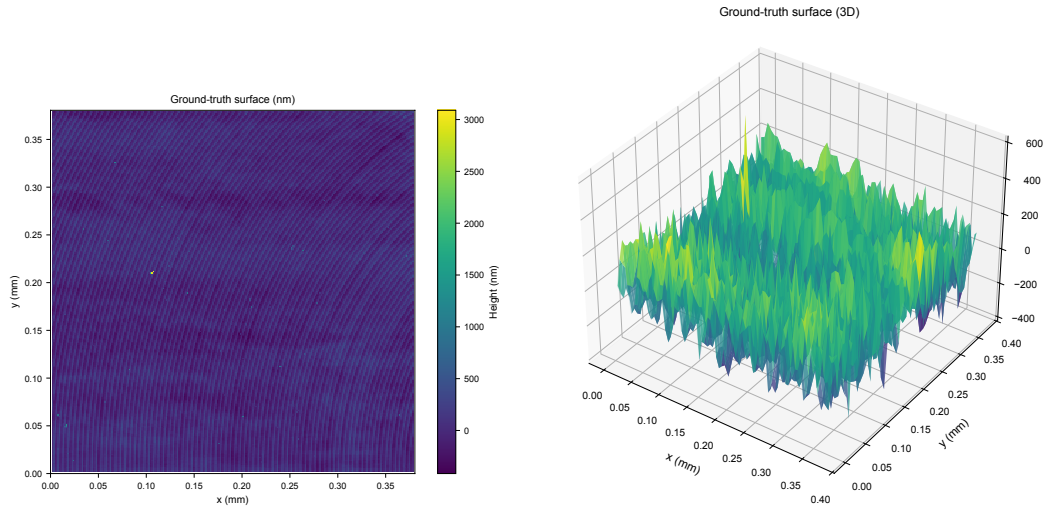

Figure 2. Representative *measured* benchmark surface derived from focus variation (FV) microscopy. The sample shown is titanium, Ti-6Al-4V, after turning (finishing). The displayed map is the area-averaged downsampled interferometric input surface used for the simulated classical and coincidence proxy measurements (2D height map and 3D rendering). In the main manuscript, the height RMSE, matched bandwidth roughness, and PSD comparisons are referenced to this same benchmark grid surface, whereas native grid FV descriptors are retained only as harsher diagnostic stress tests.

2.2. Classical Interferometer: Michelson-Type 4-Step PSI

A Michelson-type reflective interferometer operated in standard four-step PSI is modelled. The per-pixel intensity at phase step d_k is modelled as

$$I_k = B + A(1 + V\cos(\phi + d_k)), \tag{1}$$

where B is background, A is amplitude, V is fringe visibility, and the reflection phase is $\phi = \frac{4\pi}{\lambda}h(x, y)$. Shot noise is modelled by Poisson counting with a nominal photon budget

(photons per pixel per frame). Unless stated otherwise, the tables use $\lambda = 532\,\text{nm}$, baseline visibility $V = 0.85$, and a nominal detected budget of 8×10^4 photons/pixel/frame before any sample-dependent scaling; the phase-step error and drift knobs are set to zero ($\sigma_d = 0$, no background/amplitude drift).

2.3. Coincidence Proxy Forward Model

A coincidence proxy observable is simulated in a four-step phase-stepped protocol, such that the same phase estimator as in classical PSI can be used. The coincidence mean is modelled analogously, but with an effective wavelength determined by the reported coincidence proxy (difference synthetic wavelength or NOON-like half-wavelength), following the general synthetic wavelength and multi-photon phase logic used in classical two-colour and quantum interferometric sensing [19,20,22,27,29].

In this study, the coincidence channel is used as an architecture-level proxy rather than as a fully calibrated fourth-order interference transfer model. The comparison therefore asks how a coincidence-derived long-effective-wavelength observable changes reconstruction performance under matched detected count assumptions, especially when used directly versus as a coarse prior in a hybrid pipeline. No experimentally validated quantum surface profilometry implementation is claimed.

Throughout the Sections 3 and 4, the direct long-Λ branch is referred to as the coincidence proxy branch. Internal code labels such as "quantum-like" denote the same proxy observable and do not imply a validated claim of entanglement-enhanced surface metrology.

Two detector models are used in this work: (i) a simple Poisson model around a cosine mean and (ii) a rate-based coincidence model with accidentals parametrised by a coincidence window and gate time. The latter is used only in the targeted detector-side sensitivity check and the full-dataset rate model control, where efficiency imbalance and a simple nonparalysable deadtime proxy are also added. Table 1 audits the proxy, count, non-ideality, and roughness filter settings used by the paper-facing controls. The proxy-based conclusions are therefore limited to the audited count, visibility, phase-step, drift, accidental, efficiency imbalance, and deadtime regimes. They should not be extrapolated to dark-count-dominated, poor mode overlap, poor-visibility closure, or severe low-light conditions without a measured fourth-order transfer function and detector-specific calibration. Unless stated otherwise, the tables use $\lambda_1 = 810\,\text{nm}$ and $\lambda_2 = 809\,\text{nm}$, baseline visibility $V = 0.6$, a nominal detected budget of 3×10^4 pairs/pixel/frame before any sample-dependent scaling, and the simple (Poisson) coincidence model.

Mapping a laboratory coincidence instrument onto this hierarchy would require calibration steps that are deliberately outside the present simulation benchmark: the measured fourth-order transfer function, mode overlap and visibility closure, detector timing jitter, dark-count and accidental coincidence rates, channel efficiencies, deadtime response, and alignment stability would all need to be measured rather than assumed. The present proxy therefore supports architecture screening and experimental design, while laboratory validation belongs to the hardware follow-up stage.

Within the present cosine–Poisson proxy, the local height–noise scale obeys the same qualitative dependence as standard PSI, namely $\sigma_h \propto \lambda_{\text{eff}}/(4\pi V\sqrt{N})$ for detected count budget N. The "matched budget" comparison in this paper should therefore be read as a matched detected count comparison under fixed transfer models, not as a claim of universal Fisher-equivalent normalisation across all interferometer architectures. The classical two-colour and wavelength spacing controls are included to determine whether the main regime split is robust to that resource-accounting choice. Accordingly, a targeted budget-normalisation sensitivity control later repeats the representative NOON-like step-

free case under three conventions: the paper-default budget split, equal nominal quanta per pixel, and matched detected counts in the coincidence proxy.

Table 1. Parameter audit for the proxy, detector, non-ideality, and roughness filter controls. The simple measured surface benchmark uses a cosine–Poisson coincidence proxy under matched detected count assumptions; rate model rows identify where accidentals, efficiency imbalance, and nonparalysable deadtime enter. These settings define the simulation hierarchy and should not be read as a calibrated hardware transfer function.

Control/Location	Observable Model	Count and Wavelength Settings	Non-Idealities or Filters	Repetitions
Measured surface benchmark	Classical PSI plus direct and hybrid coincidence proxy branches	$\lambda_{class} = 532$ nm; $\lambda_1/\lambda_2 = 810/809$ nm; difference branch $\Lambda = 655.29$ µm; classical budget 8×10^4 photons/pixel; proxy budget 3×10^4 pairs/pixel	Simple cosine–Poisson proxy; phase steps $0, \pi/2, \pi, 3\pi/2$; no imposed stepping jitter or frame drift; hybrid smoothing $\sigma = 1.5$ px	4 per surface/method
Full-dataset rate model control, ideal	Rate-based coincidence model with matched mean coincidence counts	Gate time 1 ms; target mean coincidence counts 3×10^4 per pixel; same wavelength settings as the measured benchmark	Accidentals disabled ($\tau_c = 0$); $\eta_1 = \eta_2 = 1$; dark rates 0; deadtime disabled	4 per surface/method
Full-dataset rate model control, non-ideal	Rate-based coincidence model with matched mean coincidence counts	Gate time 1 ms; target mean coincidence counts 3×10^4 per pixel; same wavelength settings as the measured benchmark	Accidentals window $\tau_c = 20$ ns; $\eta_1 = 0.70$, $\eta_2 = 0.50$; dark rates 0; nonparalysable deadtime 5 ns in both arms	4 per surface/method
Measured surface phase/drift control	Same simple proxy as the measured benchmark	Same count and wavelength settings as the measured benchmark	Phase-jitter branch uses Gaussian phase-step perturbation $\sigma_\delta = 3°$; drift branches use 5% frame-to-frame background and amplitude drift; drift+norm additionally uses least-squares phase recovery with frame mean normalisation	4 per surface/method
Step-free detector sensitivity	Rate model control in the representative $S_q \approx 80$ nm step-free regime	128×128 grid; target mean coincidence counts 3×10^4 per pixel; gate time 1 ms	Accidentals window 20 ns; imbalance $\eta_1 = 0.70$, $\eta_2 = 0.50$; nonparalysable deadtime 5 ns; zero-step case isolates direct-branch sensitivity	10 per scenario
Approximate roughness filter control	Same benchmark grid reconstructions, roughness descriptors recomputed after Gaussian nesting index surrogate	Same measured surface benchmark settings	Approximate Gaussian S/L control on the benchmark grid with $\lambda_s = 2.5$ µm and $\lambda_c = 80.0$ µm	4 per surface/method

2.4. Optical Layouts

Schematic optical layouts for the classical Michelson PSI and the coincidence proxy interferometer are shown in Figure 3. The classical branch uses a standard reflective Michelson geometry with phase stepping in one arm, whereas the coincidence branch replaces the single intensity readout with two detector channels and coincidence logic. The purpose of Figure 3 is therefore to make clear which parts of the benchmark differ only in the detection architecture and effective wavelength model, and which parts remain common between the compared pipelines.

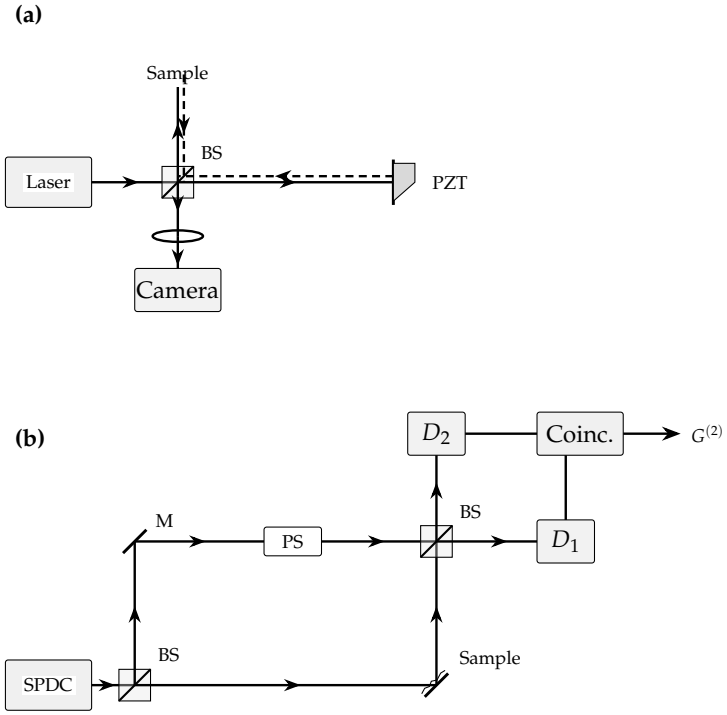

Figure 3. Schematic optical layouts used for method description. (**a**) Classical Michelson-type 4-step PSI. (**b**) Coincidence proxy interferometer with two detection channels and coincidence logic. Legend: BS/BS1/BS2, beam splitter; M, mirror; Sample, reflective test surface; PS, phase shifter; D1 and D2, single-photon detectors; "Coincidence", correlator producing $G^{(2)}$. Straight lines show beam paths; arrows indicate nominal propagation direction; dashed segments denote the return path in the reflective geometry.

2.5. Reconstruction and Metrics

Wrapped phase is reconstructed by a standard 4-step PSI estimator (or least-squares variant), unwrapped in the main benchmark by a simple separable 2D scheme (rows then columns), and converted to height via $h = \frac{\lambda}{4\pi}\phi$. Because direct branches could in principle be sensitive to that unwrap path, a targeted control later reruns the measured surface benchmark with a global least-squares Poisson unwrap; that control is reported in the Section 3 rather than being used to redefine the main paper workflow.

For the hybrid reconstruction, the final height map is not obtained by averaging the classical and coincidence-based reconstructions. Instead, a coarse-to-fine unwrapping strategy is used. First, a coarse absolute-height estimate $h_{\text{coarse}}(x, y)$ is reconstructed from the coincidence-based channel using its effective wavelength. Second, the classical PSI branch provides the wrapped short-wavelength phase $\phi_{\text{short},w}(x, y)$, from which the principal-value short-wavelength height

$$h_0(x, y) = \frac{\lambda_{\text{short}}}{4\pi}\,\phi_{\text{short},w}(x, y) \tag{2}$$

is computed. Because adding 2π to phase in reflective geometry corresponds to adding $\lambda_{\text{short}}/2$ in height, the hybrid estimator chooses the integer fringe order $k(x, y)$ that makes the short-wavelength reconstruction closest to the coarse prior,

$$k(x, y) = \text{round}\left(\frac{h_{\text{coarse}}(x, y) - h_0(x, y)}{\lambda_{\text{short}}/2}\right), \tag{3}$$

and the final hybrid height is then

$$h_{\mathrm{hyb}}(x,y) = h_0(x,y) + k(x,y)\frac{\lambda_{\mathrm{short}}}{2}. \tag{4}$$

Thus, the coincidence-based channel supplies only a coarse absolute-height prior for fringe-order selection, while the fine texture in the final hybrid map is still carried by the short-wavelength classical PSI branch. When the paper refers to H-diff, H-noon2, or related labels, the suffix indicates which coincidence proxy channel supplied h_{coarse}.

Figure 4 summarises this pipeline visually. The important architectural point is that the coincidence proxy branch is not the final texture carrier in the hybrid estimator. It enters only through the coarse height prior used to choose the short-wavelength fringe order.

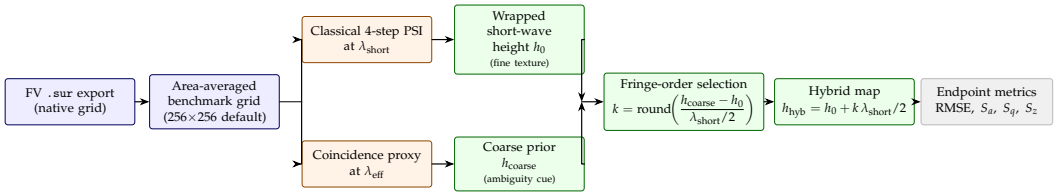

Figure 4. Hybrid coarse-to-fine reconstruction pipeline used in the benchmark. The FV export is first reduced to the common benchmark grid. The short-wavelength classical PSI branch supplies the fine wrapped height h_0, while the coincidence proxy branch supplies only the coarse prior h_{coarse}. The final hybrid map is obtained by choosing the integer fringe order that makes the short-wavelength reconstruction consistent with that coarse prior.

An auxiliary classical two-colour control is also used later in the Section 3. In that control, the coincidence-derived coarse prior is replaced by a classical synthetic wavelength prior reconstructed from two classical PSI stacks at 810 and 809 nm, allowing generic two-wavelength behaviour to be separated from coincidence model effects.

The following are reported:

- Height RMSE after plane detrending (texture-focused);
- Roughness biases for Sa, Sq, and Sz, where the primary manuscript benchmark uses matched bandwidth reference values computed on the same area-averaged benchmark grid used by the forward model, while a secondary diagnostic stress test recomputes the descriptors against the native FV `.sur` height map after invalid-pixel masking and plane removal.

These areal descriptors are chosen because both the FV reference and the interferometric reconstructions are full-field topography maps rather than extracted line profiles. The notation follows the ISO areal surface texture convention [7]. S_a tracks mean texture amplitude, S_q is the RMS height descriptor and therefore the natural bridge to the synthetic sweep roughness parameter, and S_z tracks extreme envelope excursions. Profile quantities such as R_a or R_q are not used as primary endpoints here because they would require an additional line-selection convention that is orthogonal to the architectural comparison. S_z is nevertheless retained, because long-wavelength channels may follow broad peak-to-valley structure; however, it is interpreted as the most outlier- and bandwidth-sensitive endpoint and is never treated in isolation as proof of overall metrology superiority. The approximate Gaussian S/L control is reported only as a standards-inspired nesting index sensitivity check, because the primary benchmark ranking is defined on the common-area-averaged grid rather than through a fully specified ISO filtering workflow [8].

2.6. Uncertainty and Traceability Model

The benchmark treats uncertainty at the level required for architecture screening. The FV maps define a shared FV prior rather than traceable absolute reference standards, so the

reported height and roughness errors are internal benchmark errors under that common prior. The uncertainty contributors represented in the simulation hierarchy are shot noise, visibility, wavelength selection, detected count budget, phase-step perturbation, frame drift, detector accidentals, efficiency imbalance, and deadtime. Additional benchmark-side contributors are the deterministic area-averaging step, invalid-pixel masking, plane removal, unwrap path, roughness filter choice, and the bandwidth difference between the native FV grid and the benchmark grid. These terms describe the sensitivity structure of the comparison, while a full GUM-style measurement uncertainty budget remains outside the present simulation benchmark [10]. The sensitivity and uncertainty-contributor propagation layer is summarised in the Section 3 and maps these contributors to observed changes in endpoint medians or within-surface dispersion using the already-regenerated benchmark controls.

The statistical summaries therefore quantify ranking stability within the fixed measured surface dataset rather than calibrated measurement uncertainty. Bootstrap intervals describe uncertainty in the surface-level medians; paired Wilcoxon tests and rank-biserial effects describe paired method separation on the same 59 surfaces; material and treatment holdouts test whether the ranking is portable across grouped surface families. In metrological terms, these controls support a reproducible architecture decision under common assumptions. Calibration, traceability, and deployment-facing accuracy require an independent reference chain, measured FV instrument performance, measured interferometer transfer functions, and a laboratory uncertainty budget.

2.7. Reproducibility Record

The analysis is organised around fixed regeneration scripts. The measured surface workflow regenerates the FV benchmark, manuscript figures, and tables derived from benchmark outputs and SUR headers. Additional workflows regenerate the resolution, unwrap, classical two-colour, measured rate model, optical property, step/ambiguity, non-ideal stress, and detector-sensitivity controls. The manuscript tables are written from the corresponding CSV outputs, and the paper is compiled from the LaTeX source in the manuscript directory.

The public release includes the manuscript sources, simulation code, generated tables and figures, derived 256×256 benchmark grid surfaces, public CSV summaries, regeneration commands, and SHA-256 checksums. The executable workflows are stored under `scripts/`, reusable package code under `src/qiprof/`, measured surface CSV and figure artefacts under `outputs/paper_alicona_benchmark/`, and synthetic, non-ideality, and detector-control artefacts under the corresponding `outputs/paper_optprops_*` directories. This derived data record preserves the benchmark inputs used by the simulator while avoiding distribution of the full native `.sur` files in the Git repository.

3. Results

The Section 3 first exposes the endpoint-level claim and then reports the controls that bound it. Table 2 separates the strongest fixed-workflow height fidelity result from descriptor-specific roughness behaviour; Table 3 translates the same evidence into an architecture screening map for follow-up experiments; Table 4 condenses the sensitivity and uncertainty-contributor propagation layer. This ordering is intentional: detrended height RMSE is treated as the primary architecture endpoint, whereas S_a, S_q, and S_z are treated as descriptor-specific portability tests.

Table 2. Endpoint-level evidence summary for the submission-facing claim. The table separates the fixed-workflow architecture result from endpoint-specific roughness behaviour and lists the strongest control or comparator that bounds each claim.

Endpoint	Supported Fixed-Workflow Result	Strongest Bound or Control	Submission-Facing Interpretation
Detrended height RMSE	Hybrid: 314.0 nm median; 32/59 per-surface wins	Classical frontier oracle remains higher at 559.1 nm; non-ideal rate model preserves hybrid ordering at 376.3 nm	Strongest claim: coincidence-derived information is useful as a coarse prior for height fidelity reconstruction.
Matched bandwidth $\|\Delta S_a\|$	Hybrid: 86.3 nm median absolute error	Treatment holdout agreement is 70.0%; filtered roughness control preserves the endpoint split	Useful texture-amplitude result within the benchmark grid, but not a portable roughness rule.
Matched bandwidth $\|\Delta S_q\|$	Hybrid: 152.4 nm median absolute error	Treatment holdout agreement is 60.0%; filtered material holdout weakens to 66.7%	Supports hybrid as the best primary-branch RMS texture estimator on the benchmark grid, with process dependence.
Matched bandwidth $\|\Delta S_z\|$	Coincidence proxy: 2653.1 nm median absolute error	Descriptor-scaled tolerance is still exceeded on 42.4% of surfaces; classical frontier reaches 1208.2 nm	Envelope-following benefit is selective and endpoint-specific, not evidence for overall metrology superiority.
Direct long-Λ reconstruction	Retained only for ambiguity or envelope tasks	Classical two-colour control reproduces much of the direct long-wavelength behaviour	Direct branch is a comparator and coarse observable, not the recommended final fine-texture workflow.

Table 3. Architecture regime map distilled from the measured surface benchmark and the synthetic controls. The table identifies which role each optical branch plays in the present benchmark and which cases should be prioritised in hardware follow-up.

Metrology Need	Preferred Architecture Role	Follow-Up Priority	Evidence in the Benchmark
Benchmark grid height fidelity under ambiguity	Coincidence-derived channel supplies coarse fringe order; short-wavelength PSI carries the final texture	Validate hybrid coarse-to-fine first	Hybrid gives the lowest median height RMSE (314.0 nm), wins on 32/59 surfaces, and remains best under the rate model control.
Fine-texture recovery without height ambiguity	Short-wavelength classical PSI remains the baseline; H-noon2 is effectively tied when no fringe-order decision is needed	Validate classical PSI and H-noon2 as the zero-ambiguity baseline	In the step-free sweep, hybrid advantage appears only when the coarse prior changes an ambiguity decision.
Coarse step or discontinuity estimation	Long-effective-wavelength difference branch used as a coarse observable	Test direct and hybrid difference branches on stepped surfaces	In the step/ambiguity control, long-Λ branches become the best coarse step estimators once the discontinuity enters the ambiguous regime.
Matched bandwidth S_a and S_q reporting	Hybrid branch within the primary three workflows	Include roughness reporting after matched bandwidth and filtering checks	Hybrid gives the lowest matched bandwidth S_a and S_q errors, but holdout stability remains process-dependent.
Envelope-dominated S_z diagnostics	Direct coincidence proxy branch retained as a selective comparator	Test only on targeted smoother or envelope-following cases	Direct coincidence proxy reconstruction gives the lowest matched bandwidth S_z median but exceeds descriptor-scaled tolerance on 42.4% of surfaces.

Table 4. Numerical sensitivity and uncertainty-contributor propagation for the measured surface benchmark. Entries report how endpoint medians or within-surface dispersion change when one contributor or modelling choice is perturbed. The table is a benchmark-level sensitivity summary, not a calibrated GUM uncertainty budget for a realised instrument.

Source/Control	Numerical Perturbation	Propagated Height Response	Roughness or Descriptor Response	Interpretation
Between-surface benchmark population	Percentile bootstrap over 59 surface-level medians	Hybrid height RMSE 314.0 [259.6, 435.6]	Hybrid $\lvert\Delta S_z\rvert$ 5349.5 [2204.4, 15,599.6]	Surface-family composition dominates the uncertainty envelope, especially for S_z.
Shot-noise repeatability	$n_{\mathrm{rep}} = 4$ stochastic realisations per surface and method	Median within-surface IQR: Classical 2.6 nm; Q-like 1.7 nm; Hybrid 0.9 nm	Hybrid $\lvert\Delta S_q\rvert$ IQR 2.4 nm; Q-like $\lvert\Delta S_z\rvert$ IQR 409.5 nm	Monte Carlo noise is small relative to surface-family and detector/model effects.
Phase-step jitter	Gaussian step perturbation, $\sigma_\delta = 3^\circ$	Height shift vs. baseline: Classical −10.8 nm; Q-like +0.4 nm; Hybrid −0.4 nm	Largest matched bandwidth shift: Q-like $\lvert\Delta S_z\rvert$ +513.5 nm	Moderate phase jitter weakly perturbs the primary height ranking.
Frame drift	5% frame-to-frame background and amplitude drift	Height shift vs. baseline: Classical −5.4 nm; Q-like +8.6 nm; Hybrid +0.9 nm	Largest matched bandwidth shift: Q-like $\lvert\Delta S_z\rvert$ +1038.6 nm	Drift affects the direct branch more than the hybrid primary endpoint.
Detector count formation	Rate model with $\tau_c = 20$ ns, $\eta_1 = 0.70$, $\eta_2 = 0.50$, $\tau_d = 5$ ns	Non-ideal rate shift vs. simple proxy: Classical −7.2 nm; Q-like +598.7 nm; Hybrid +62.3 nm	Rate model descriptor medians are reported in the measured surface rate model control below.	Detector non-idealities erode direct coincidence-only reconstruction before they erase the hybrid height advantage.
Unwrap algorithm	Global least-squares Poisson unwrap vs. separable row/column unwrap	Height shift vs. simple unwrap: Classical −506.5 nm; Q-like +0.6 nm; Hybrid −0.4 nm	Classical $\lvert\Delta S_z\rvert$ +3946.8 nm; Hybrid $\lvert\Delta S_z\rvert$ +341.5 nm	Classical height is unwrap-path sensitive; hybrid is nearly unchanged because the coarse prior selects fringe order.
Benchmark grid reduction	Full-dataset reruns at 128^2, 256^2, and 384^2 grids	Median height range: Classical 733.8–1203.3 nm; Q-like 709.6–710.2 nm; Hybrid 272.6–314.0 nm	Hybrid $\lvert\Delta S_z\rvert$ range 5349.5–8300.1 nm	Grid choice changes absolute values but preserves the hybrid height fidelity ordering.
Gaussian S/L roughness filter	Approximate Gaussian nesting index control, $\lambda_s = 2.5$ µm and $\lambda_c = 80$ µm	Height endpoint not recomputed; this is descriptor post-processing.	Hybrid shifts: $\lvert\Delta S_a\rvert$ +7.6 nm; $\lvert\Delta S_q\rvert$ −7.1 nm; $\lvert\Delta S_z\rvert$ +887.4 nm	Filtering changes descriptor magnitudes but preserves the endpoint winner split.

3.1. Benchmark on Focus Variation Measured Surfaces

To align the evaluation with realistic textures, the classical, coincidence proxy, and hybrid pipelines are benchmarked on $n = 59$ measured surfaces acquired using focus variation (FV) microscopy and exported as Mountains/DigitalSurf `.sur` height maps. The measured surface branch is intentionally framed as an architecture screening benchmark under a shared FV geometric prior: the same export supplies the simulator input surface and the comparison geometry, so this section determines whether a reconstruction architecture remains self-consistent after stronger controls are imposed, rather than claiming external metrological calibration. For each sample, the native FV `.sur` surface is reduced to the common 256×256 interferometric benchmark grid by block area averaging, and the optical forward model is then evaluated on that common-area-averaged surface. This benchmark grid/native grid distinction is used throughout the Section 3: Height RMSE is always computed on the benchmark grid, while roughness descriptors are reported in two domains. The primary roughness endpoints use the same area-averaged benchmark grid reference as the forward model, whereas native grid FV roughness is retained as a reference sensitivity stress test. The canonical paper benchmark uses repeated stochastic realisations per surface and method, which are collapsed to surface-level medians before between-surface statistics are reported. Filenames encode material and surface-treatment codes which are decoded into consistent English labels for reporting; detailed material and treatment summaries are provided in the Supplementary Materials, Tables S3 and S4.

The overall measured surface height-error behaviour is summarised in Figures 5 and 6. The distribution view condenses the detrended RMSE values across all measured samples,

while the roughness trend view shows how those RMSE values evolve with the roughness level of the benchmark grid reference surfaces. Together, these views show the same high-level regime split: the hybrid branch remains the strongest overall height fidelity option, while the direct coincidence proxy branch and the classical baseline trade wins only in narrower subsets of the dataset.

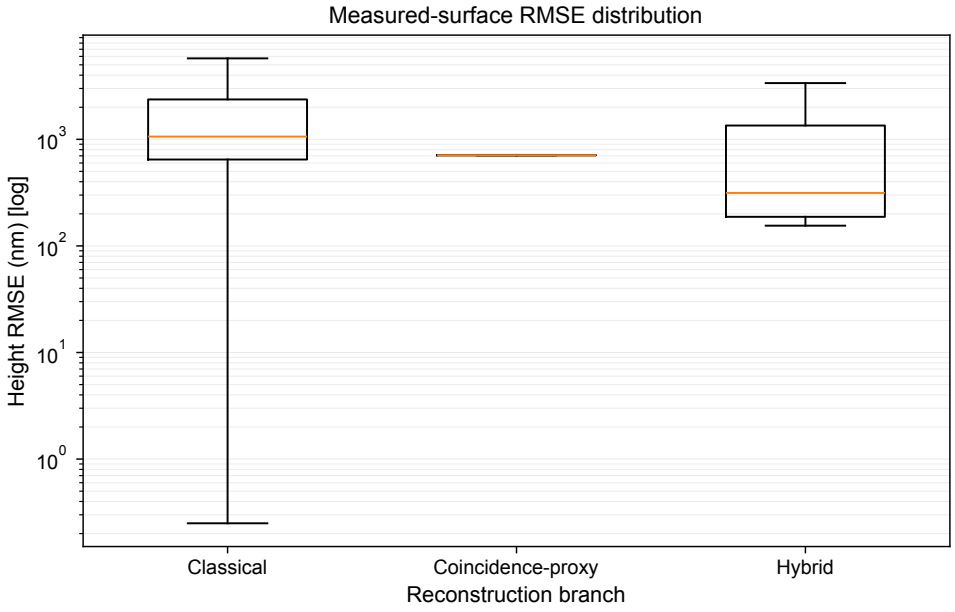

Figure 5. Measured surface height RMSE distributions across all FV `.sur` samples. All values are computed against the common benchmark grid reference after plane detrending, so the figure evaluates architecture-level height fidelity under the shared measured surface prior.

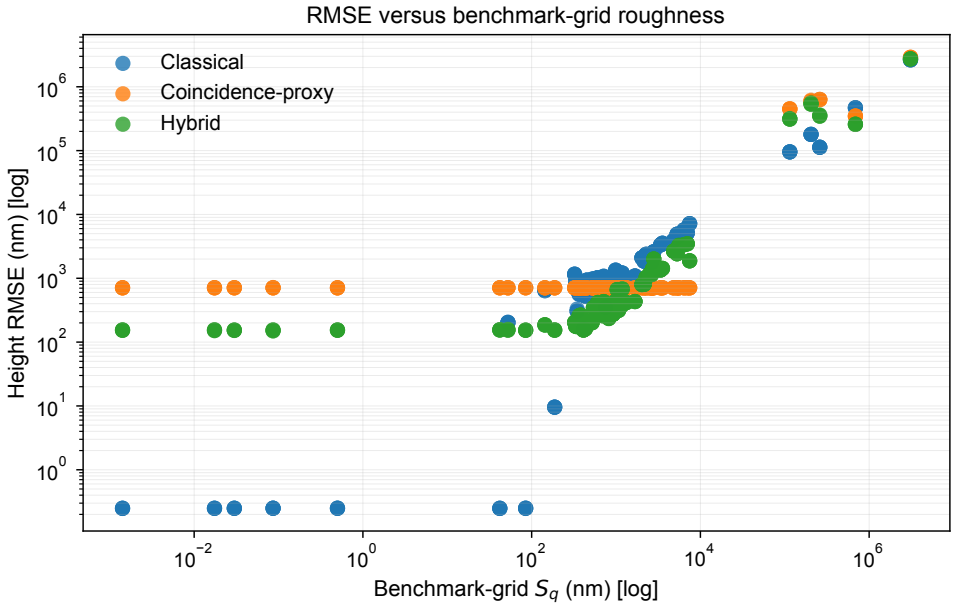

Figure 6. Measured surface height RMSE as a function of benchmark grid roughness S_q. Each point corresponds to one measured FV `.sur` surface after area-averaged reduction to the benchmark grid and plane detrending.

Those robust grouped summaries sharpen the regime split. Hybrid gives the lowest median material-wise height RMSE for aluminium 7075, stainless steel 1.4301, steel C45, and titanium, whereas the direct coincidence proxy branch is the grouped median winner only for brass and graphite. Across treatments, hybrid gives the lowest median height

RMSE in eight of ten classes; the only grouped exceptions are turning (roughing) and wire-EDM (roughing), where the direct coincidence proxy branch is still preferred. Classical PSI is therefore an essential short-wavelength reference and the donor branch inside the hybrid estimator, but it is not the robust median winner in any grouped measured surface slice.

Because the metrology target is not only the recovered height map but also the derived areal texture descriptors, the same benchmark was also compared directly in terms of the reconstructed roughness parameters. Figures 7–9 promote the matched bandwidth comparison to the primary roughness view: estimated S_a, S_q, and S_z are compared against the corresponding values computed on the same area-averaged benchmark grid used by the interferometric forward model. Under this bandwidth-consistent reference, hybrid is the most texture-faithful branch for S_a and S_q, whereas the coincidence proxy branch remains selectively favoured only for S_z. The matched bandwidth medians in the Supplementary Materials, Table S6 make the same split explicit: hybrid reduces median absolute S_a and S_q error to 86.3 and 152.4 nm, while the coincidence proxy branch retains the lowest median absolute S_z error at 2653.1 nm.

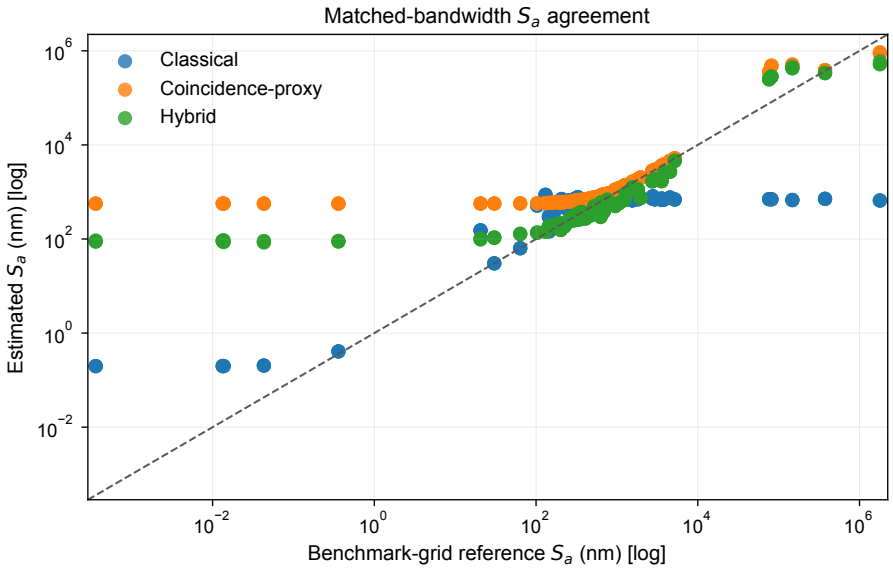

Figure 7. Primary matched bandwidth S_a comparison for the measured surface benchmark. Each point is one measured `.sur` surface; the dashed line marks perfect agreement between the reconstruction and the S_a value computed on the same area-averaged benchmark grid used by the interferometric forward model.

The native grid FV roughness comparison is retained only as a diagnostic stress test rather than as the main workflow recommendation. In the Supplementary Materials, Table S5 reports that stricter native grid view, where the direct coincidence proxy branch can appear more favourable for S_q and S_z because the native FV maps retain more local variance and more extreme peak-to-valley excursions than the downsampled benchmark surface used by the forward model. That diagnostic is still useful because it exposes reference sensitivity, but it should not be treated as the primary ranking. The native grid S_q preference does not survive the matched bandwidth control reported in the Supplementary Materials, Table S6, which shows that most of the apparent direct-branch S_q gain is explained by the native grid versus benchmark grid bandwidth mismatch.

The residual/Bland–Altman-style view in the Supplementary Materials, Figure S2, makes the matched bandwidth bias structure explicit. For S_a and S_q, the direct coincidence proxy branch is generally shifted upward relative to the benchmark grid reference, whereas the hybrid branch stays closer to zero across the central mass of surfaces. For S_z, all three branches remain negatively biased, but the central tendency is least negative for the

coincidence proxy branch. The broad spread across all three panels confirms that the bias structure is heteroscedastic rather than a single global offset.

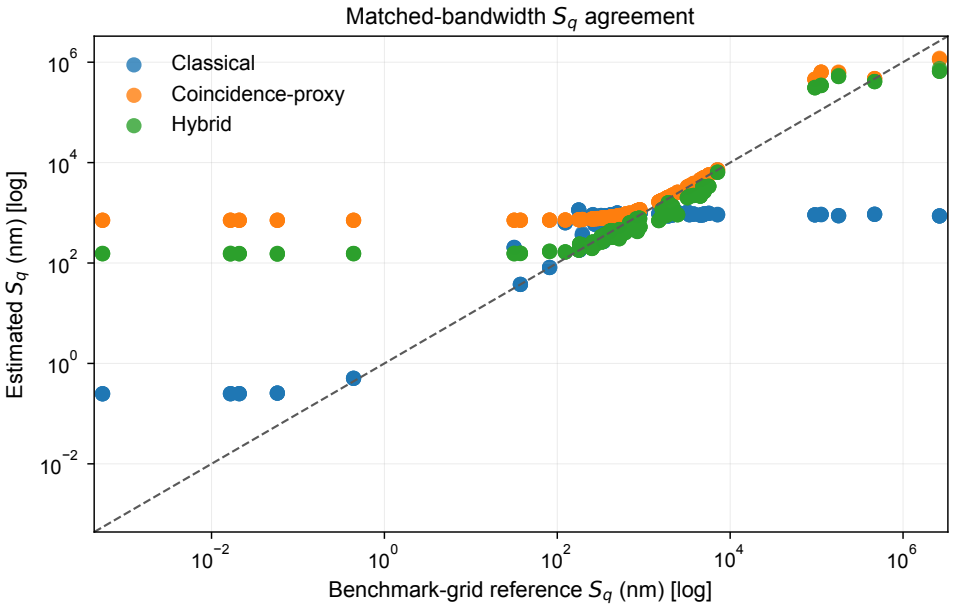

Figure 8. Primary matched bandwidth S_q comparison for the measured surface benchmark. Each point is one measured `.sur` surface; the dashed line marks perfect agreement between the reconstruction and the S_q value computed on the same area-averaged benchmark grid used by the interferometric forward model.

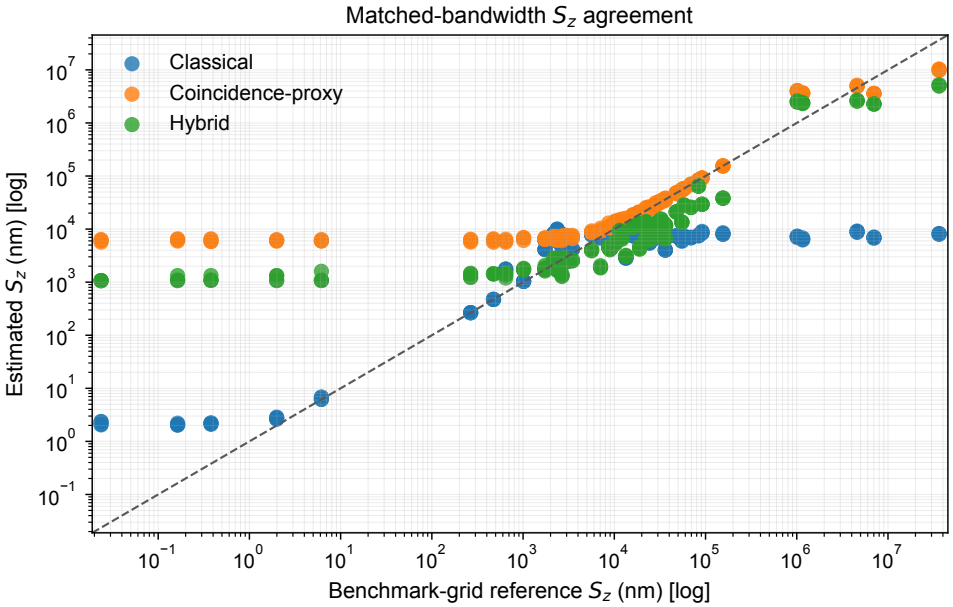

Figure 9. Primary matched bandwidth S_z comparison for the measured surface benchmark. Each point is one measured `.sur` surface; the dashed line marks perfect agreement between the reconstruction and the S_z value computed on the same area-averaged benchmark grid used by the interferometric forward model. The coincidence proxy branch remains selectively favourable for this envelope-sensitive descriptor.

The treatment-resolved view in the Supplementary Materials, Figure S4, shows that this is not a uniform material-independent effect. Hybrid reconstruction keeps the smallest median matched bandwidth $|\Delta S_a|$ and $|\Delta S_q|$ in most finishing-style classes, whereas the coincidence proxy branch often produces the smallest matched bandwidth $|\Delta S_z|$ on smoother treatments such as honed, burnished, and turned-finish surfaces. On rougher

classes, especially turning (roughing), milling (finishing), and wire-EDM textures, all three branches degrade sharply, confirming that the measured surface advantage is process-dependent rather than universal. The treatment holdout analysis below sharpens that caution: matched bandwidth treatment holdout agreement falls to 70.0% for $|\Delta S_a|$, 60.0% for $|\Delta S_q|$, and 60.0% for $|\Delta S_z|$.

To make that regime split explicit at the per-surface level, Table S7 in the Supplementary Materials reports winner counts for each endpoint. The dominance view reinforces the same message as the grouped medians for height fidelity: hybrid reconstruction remains strongest overall for height RMSE (32 wins). Under the matched bandwidth roughness reference, hybrid also leads the per-surface winner counts for $|\Delta S_a|$ (23 wins) and $|\Delta S_q|$ (25 wins), whereas the coincidence proxy branch dominates $|\Delta S_z|$ (30 wins). The low-roughness column still shows that the hybrid advantage is concentrated most strongly in the smoother regime for height fidelity, while the rougher column remains selectively favourable to the coincidence proxy branch for envelope-dominated descriptors. These winner counts are descriptive only and should be read alongside the holdout and tolerance summaries rather than as deployable ranking rules.

Bootstrap uncertainty and endpoint-referenced tolerance summaries sharpen that caution further. The detailed bootstrap, repeatability, holdout, tolerance, filtering, paired-effect, and Wilcoxon screens are reported in the Supplementary Materials, Tables S8–S15. In those tables, the hybrid branch gives the smallest surface-level median height RMSE, matched bandwidth $|\Delta S_a|$, and matched bandwidth $|\Delta S_q|$, while the coincidence proxy branch remains the most favourable direct branch only for matched bandwidth $|\Delta S_z|$. The within-surface repeat-dispersion view shows that the stochastic Monte Carlo spread is substantially smaller than the between-surface regime split that drives the main ranking. For height RMSE, hybrid remains the training set winner in all six material and all ten treatment splits, with held-out agreement on 4/6 and 8/10 splits, respectively. The roughness endpoints are still process-dependent across treatments, but no longer collapse as severely once the reference bandwidth is matched: treatment holdout agreement is 70.0% for $|\Delta S_a|$, 60.0% for $|\Delta S_q|$, and 60.0% for $|\Delta S_z|$. Even the best matched bandwidth $|\Delta S_z|$ branch still exceeds the descriptor-scaled tolerance band on 42.4% of surfaces, compared with 74.6% for classical and 62.7% for hybrid.

The paired statistical screen is defined at the surface level. Monte Carlo repetitions are first collapsed to one median value per surface, method, and endpoint; paired Wilcoxon signed-rank tests then compare methods on the same 59 surfaces. The Holm correction is applied within the endpoint/method-pair family reported in the Supplementary Materials, Table S15, and rank-biserial effects retain their sign so negative values favour the second-listed lower-error method. These tests provide dataset-level evidence for the benchmark ranking; metrological uncertainty treatment remains part of the experimental follow-up.

The paired-effect summary and the Holm-adjusted paired screen make the same trade-off more explicit. The hybrid approach shows the strongest paired height–RMSE separation from classical (median difference −625.1 nm, 79.7% paired wins, rank-biserial effect $r_{\mathrm{rb}} = -0.66$, Holm-adjusted $p = 1.1 \times 10^{-4}$), while the hybrid-versus-coincidence proxy separation is directionally favourable but weaker after multiplicity correction. For roughness, the inferential screen supports the endpoint map rather than a universal ranking: hybrid is clearly favoured over classical for $|\Delta S_a|$ and $|\Delta S_q|$, whereas the coincidence proxy branch retains its strongest paired evidence only for $|\Delta S_z|$ against classical.

One remaining question is whether a standards-aligned roughness filtering surrogate materially changes the endpoint regime map before any cross-instrument study is available. Table S13 in the Supplementary Materials answers this question narrowly for the present dataset: applying an approximate Gaussian S/L nesting index control on the same bench-

mark grid changes the absolute roughness medians but preserves the same endpoint split. The filtered view still favours hybrid for $|\Delta S_a|$ and $|\Delta S_q|$ and the coincidence proxy branch only for $|\Delta S_z|$, while treatment holdout agreement remains at 60.0% for all three filtered descriptors and the filtered $|\Delta S_q|$ material holdout agreement drops to 66.7%. The control therefore strengthens the boundary map around the main result: roughness portability is descriptor- and process-dependent, whereas the hybrid height fidelity result is the more stable architecture-level endpoint in this benchmark.

Figures 10–12 show the same ranking without group aggregation. Most surfaces lie below the identity line in both hybrid-versus-classical and hybrid-versus-coincidence proxy comparisons, consistent with the paired-effect table. The predictor view shows that the hybrid gain is strongest across the low-to-mid benchmark grid S_q regime and remains negative on most surfaces, while the roughest textures generate the largest remaining outliers. This is consistent with the treatment-level medians, where rough turning and rough wire-EDM remain the main grouped exceptions to the hybrid advantage.

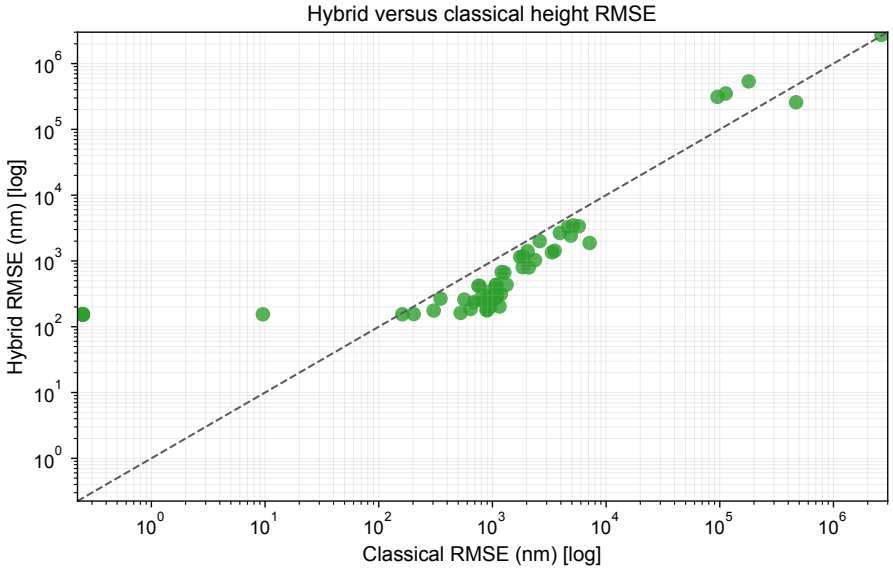

Figure 10. Per-surface paired comparison of hybrid and classical height RMSE on the same measured surfaces. Points below the dashed identity line favour the hybrid branch.

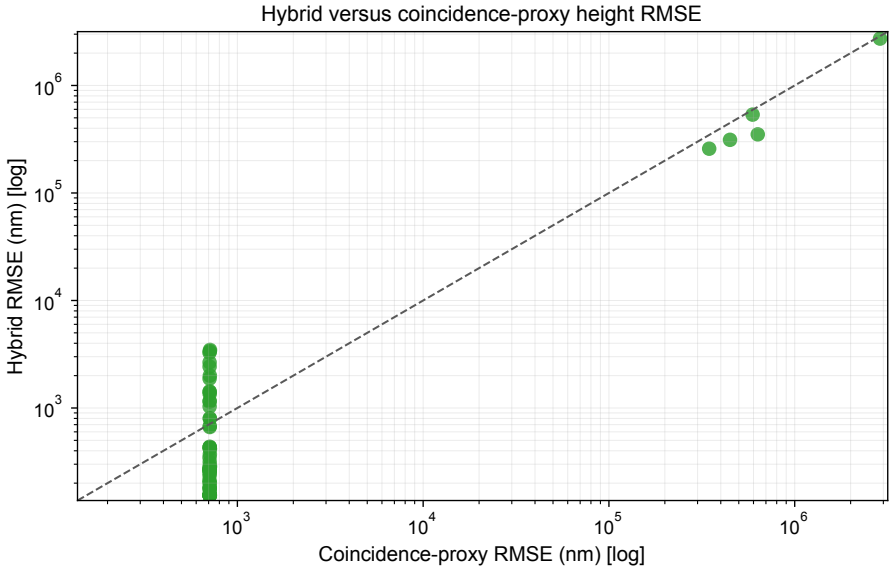

Figure 11. Per-surface paired comparison of hybrid and direct coincidence proxy height RMSE on the same measured surfaces. Points below the dashed identity line favour the hybrid branch.

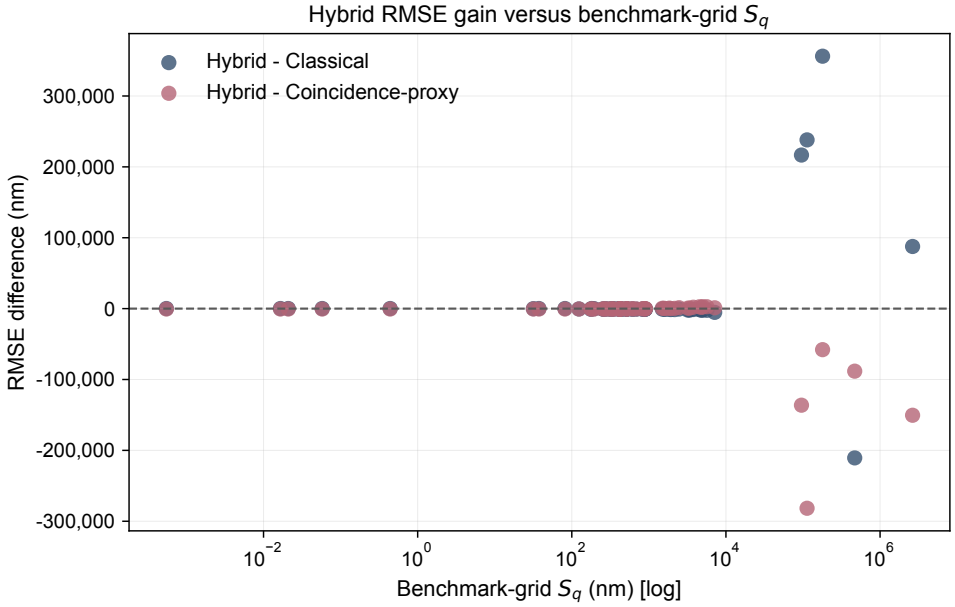

Figure 12. Hybrid height–RMSE gain as a function of benchmark grid reference roughness S_q. Negative values favour hybrid reconstruction over the comparator branch.

A second architectural control separates coincidence model effects from generic synthetic wavelength behaviour. Table S16 in the Supplementary Materials compares the direct Q-diff branch and H-diff against a classical two-colour synthetic wavelength baseline constructed from two classical PSI stacks at 810 and 809 nm. The direct long-Λ behaviour is partly reproduced by the classical baseline: the classical two-colour workflow gives roughness medians comparable to or better than Q-diff on the matched bandwidth grid, especially for $|\Delta S_z|$ (2346.7 versus 2879.5 nm), but it remains poor in pointwise height fidelity (1210.3 versus 710.0 nm). The key architectural result therefore survives at the hybrid level rather than at the direct long-Λ branch itself: H-diff retains the lowest height RMSE (313.6 nm), whereas H-2-colour becomes competitive mainly for matched bandwidth Sa and Sq bias. This control supports treating the coincidence-style channel as a useful coarse prior while preventing the direct envelope-descriptor effect from being overinterpreted.

That strengthened comparator can be pushed further. Table 5 forms an optimistic classical frontier oracle by allowing each surface and endpoint to choose the best available classical workflow among the default PSI, least-squares unwrap, LSQ plus frame normalisation, and classical two-colour controls. Even under that oracle, the classical family still does not close the main height fidelity gap: the best classical frontier remains at 559.1 nm versus 314.0 nm for hybrid. The roughness picture is less flattering to the hybrid claim: the oracle classical frontier reaches lower median absolute biases for Sa, Sq, and Sz (74.2, 86.4, and 1208.2 nm) by switching workflows across surfaces. The defensible claim is therefore narrower than the universal hybrid dominance over every endpoint. What survives the strengthened comparator is that hybrid remains the strongest fixed-workflow height fidelity architecture, while roughness leadership across the broader classical family remains endpoint- and workflow-dependent.

The same split is visible in the spatial frequency domain. Figure S3 in the Supplementary Materials shows the radial power spectral density (PSD) on two representative measured surfaces reconstructed by the three pipelines on the common benchmark grid: a smoother ground stainless-steel surface and a rougher turned Ti-6Al-4V surface. Because the reference PSD is the same area-averaged benchmark surface used to drive the forward model, this figure should be read as an internal spectral consistency diagnostic rather than as independent spectral validation. Across both panels, classical PSI and the hybrid branch

remain closer to the reference spectrum into the higher spatial frequency region, whereas the direct coincidence proxy reconstruction follows the lower-frequency envelope more strongly than the fine-texture tail. This benchmark grid frequency-domain view is consistent with the metric-level result that the coincidence proxy channel can remain useful for broader envelope descriptors while still losing fine-scale fidelity.

Table 5. Classical frontier control over $n = 59$ measured surfaces. The table compares the default classical paper workflow, a classical least-squares unwrap, a stronger classical least-squares plus frame-normalised workflow, and a classical two-colour synthetic wavelength baseline. The 'Best classical frontier (oracle)' column is an optimistic upper bound formed by taking the per-surface minimum for each endpoint across these classical baselines and then reporting the resulting dataset median; it is therefore not a single deployable workflow. The final column shows the main hybrid branch for reference.

Endpoint	Classical Default	Classical LS Unwrap	Classical LSQ + Norm	Classical 2-Colour	Best Classical Frontier (Oracle)	Hybrid
Height RMSE (nm)	1065.7	559.2	559.1	1209.9	559.1	314.0
$\lvert\Delta S_a\rvert$ (nm)	301.8	366.3	366.8	265.4	74.2	129.6
$\lvert\Delta S_q\rvert$ (nm)	423.4	729.4	728.4	316.8	84.7	419.1
$\lvert\Delta S_z\rvert$ (nm)	37,163.2	43,202.2	43,198.0	2278.5	1368.5	37,457.4

Figure 13 complements those aggregate diagnostics with spatial residual maps for three automatically selected measured surface cases: the surface with the lowest hybrid RMSE in the benchmark, the surface whose hybrid RMSE is closest to the dataset median, and the surface with the strongest direct coincidence proxy failure relative to the other branches. The same regime split remains visible in the image domain. In the low-error case, classical PSI and the hybrid branch keep the residual field locally structured and small, whereas the direct coincidence proxy reconstruction already shows broader envelope-scale distortion. In the mid-regime case, the hybrid branch visibly reduces the large-scale offset pattern relative to the direct coincidence proxy branch while remaining closer to the fine texture than the long-Λ inversion. In the direct-Q-failure case, the residual field confirms that the difference wavelength branch can fail catastrophically even when the classical and hybrid solutions remain well-behaved.

The full treatment-resolved roughness plot is provided in the Supplementary Materials, Figure S4.

Because the forward model operates on a downsampled benchmark grid, a full-dataset grid-sensitivity control was added at 128×128, 256×256, 384×384, and 512×512 (Supplementary Materials, Figure S5). The absolute values shift with grid choice, as expected. Across the full dataset, hybrid remains the lowest-median branch for pointwise height RMSE at every tested grid (272.6, 314.0, 303.5, and 326.1 nm), whereas the coincidence proxy branch remains nearly flat around 709–710 nm and the classical branch rises from 733.8 to 1399.2 nm as the grid is refined. For matched bandwidth roughness bias, hybrid remains the lowest-median branch for $\lvert\Delta S_a\rvert$ and $\lvert\Delta S_q\rvert$ at every tested grid, while the direct coincidence proxy $\lvert\Delta S_z\rvert$ advantage persists across the tested grids. This control does not remove the benchmark grid dependency, and it does not test non-uniform decimation, but it does show that the central hybrid height fidelity advantage is not created solely by the default 256×256 simulation grid.

Table S17 in the Supplementary Materials adds a second dataset-level control for the direct branches: the full measured surface benchmark is rerun with a global least-squares Poisson unwrap instead of the default separable row/column unwrap. The hybrid branch is unchanged by construction because its fringe-order assignment is driven directly by the coarse coincidence prior rather than by a full-map unwrap of the short-wavelength phase. The control does not leave all direct-branch medians unchanged: the global least-squares unwrap lowers classical height RMSE from 1065.7 to 559.2 nm and improves the classical

matched bandwidth $|\Delta S_a|$ and $|\Delta S_q|$ medians to 270.7 and 392.1 nm, but it worsens classical $|\Delta S_z|$ to 12,412.7 nm and leaves the direct coincidence proxy branch essentially unchanged. Hybrid therefore remains the lowest-median height fidelity branch overall, and the control does not overturn the architectural reading that the coincidence channel is more defensible as a coarse prior than as a direct final estimator.

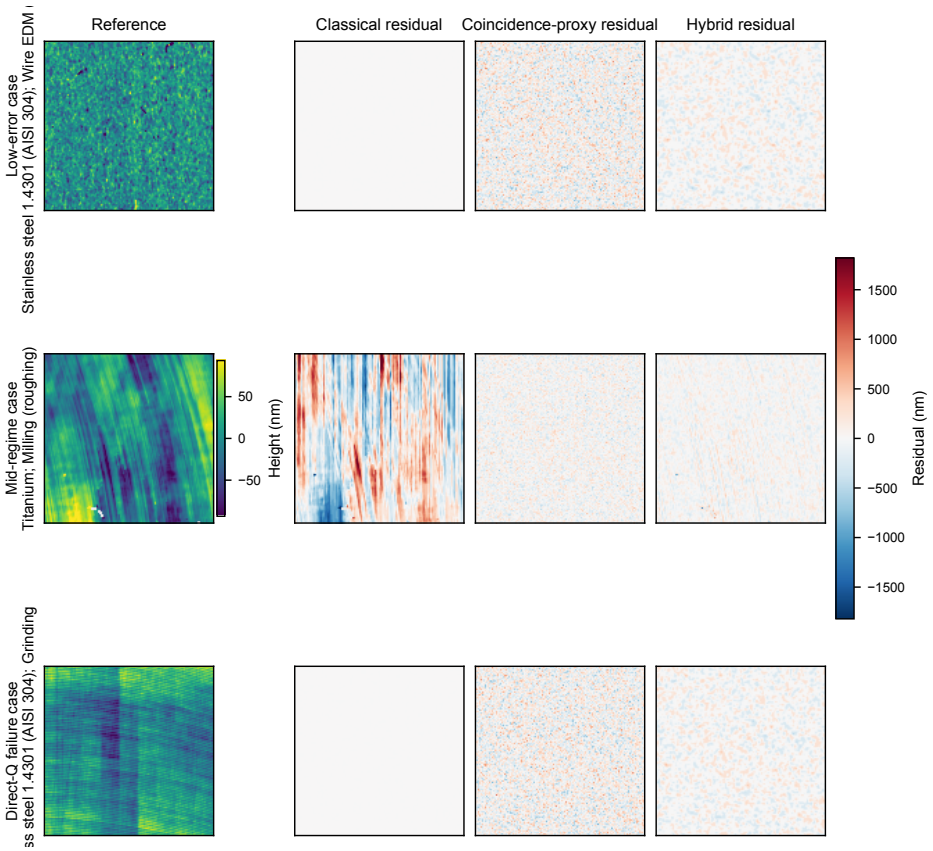

Figure 13. Representative measured surface residual maps for three automatically selected benchmark cases. Rows correspond to the lowest-hybrid-error case, the surface whose hybrid RMSE is closest to the dataset median, and the surface with the largest direct coincidence proxy failure ratio relative to the classical/hybrid baselines. Columns show the downsampled benchmark surface used for the forward simulation and the residual maps for the classical, coincidence proxy, and hybrid reconstructions. The figure makes the image-domain consequence of the regime split explicit: the direct long-effective-wavelength branch develops broader residual structures well before the hybrid branch loses the short-wavelength texture.

Table 6 then promotes the stronger coincidence forward model into the full measured surface branch. The ideal matched-count rate model is useful as an internal check: it lowers the direct coincidence proxy median height RMSE from 709.6 to 355.1 nm and the hybrid median from 314.0 to 290.9 nm, so the simple proxy is not artificially optimistic for the coincidence channel on the main height fidelity endpoint. For practical interpretation, however, the non-ideal rate setting is the more conservative benchmark. Once accidentals, efficiency imbalance, and deadtime are added in the same rate framework, the direct coincidence proxy branch becomes the most fragile path again, with median height RMSE rising to 1308.3 nm and $|\Delta S_a|$ to 741.9 nm, whereas the hybrid branch degrades more modestly to 376.3 nm and 184.7 nm. The rate model control therefore exposes a real deployment risk for direct coincidence-only reconstruction while preserving the narrower architectural reading: coincidence information can still be useful, yet it is more defensible as a coarse prior than as the sole final estimator on rough measured surfaces.

Table 6. Measured surface rate model control over the full benchmark dataset. The canonical simple coincidence proxy is compared against a matched-count rate model and a non-ideal rate model with accidentals, detector efficiency imbalance, and nonparalysable deadtime. The rate model controls use a gate time of 1.000 ms; the non-ideal setting additionally uses $\tau_c = 20.0$ ns, $\eta_1 = 0.70$, $\eta_2 = 0.50$, and deadtime $\tau_d = 5.0$ ns in each arm. Entries report surface-level medians after collapsing repeated runs within each surface and method.

Endpoint	Classical	Q-Like (Simple)	Q-Like (Rates)	Q-Like (Non-Ideal Rates)	Hybrid (Simple)	Hybrid (Rates)	Hybrid (Non-Ideal Rates)
Height RMSE (nm)	1065.7	709.6	355.1	1308.3	314.0	290.9	376.3
$\lvert\Delta S_a\rvert$ (nm)	301.8	308.1	142.3	741.9	129.6	138.6	184.7
$\lvert\Delta S_q\rvert$ (nm)	423.4	395.2	262.2	805.2	419.1	435.8	362.3
$\lvert\Delta S_z\rvert$ (nm)	37,163.2	25,745.5	26,473.1	23,823.9	37,457.4	37,642.6	37,190.8

3.2. Step-Free Optical Property Sweep

To isolate texture-fidelity effects from step-unwrapping effects, a second benchmark was carried out on step-free synthetic rough surfaces while varying the surface RMS, sample reflectivity R, and visibility scaling factor V_{scale}. The full height–RMSE table is provided in the Supplementary Materials, Table S18. Two representative regimes are promoted into the main figure set in Figure 14: a low-return case ($R = 0.20$, $V_{\text{scale}} = 0.70$) and a high-return case ($R = 1.00$, $V_{\text{scale}} = 1.00$).

Across both cases, the difference wavelength coincidence proxy (Q-diff) and the associated hybrid (H-diff) are clearly suboptimal for texture reconstruction: as the effective wavelength becomes very long, nm-scale roughness produces only weak phase modulation, and the resulting height RMSE becomes orders of magnitude larger than for the classical baseline. By contrast, the NOON-like channel stays close to the classical solution at low roughness and becomes lower RMSE in the rougher part of the high-return case. In the low-return scenario, classical PSI and H-noon2 remain the most stable fine-texture baselines. In the high-return case, Q-noon2 gives the lowest RMSE at $S_q = 80$ nm, while H-noon2 becomes the best-performing branch at $S_q = 120$ nm. Classical PSI is therefore not the universal winner across the synthetic sweep; the more defensible statement is that classical PSI and H-noon2 define the default fine-texture baseline at low-to-moderate roughness, whereas the NOON-like coincidence branches become competitive or superior only in the rougher step-free regime.

That reading is not a knife-edge consequence of choosing exactly 810/809 nm for the difference wavelength proxy. Supplementary Materials, Table S19 perturbs the spacing around that design point and repeats the representative step-free $S_q \approx 80$ nm case at two visibility levels. Shortening the synthetic wavelength improves both Q-diff and H-diff, but even the best H-diff case remains far above classical PSI. The direct and hybrid difference wavelength branches are therefore structurally poor for fine-texture recovery in this proxy model, not merely mistuned at the nominal 810/809 operating point.

The fairness question behind the matched-count comparison was also tested more broadly. Table S20 in the Supplementary Materials repeats the NOON-like step-free control in three representative regimes: a low-return/low-visibility case ($S_q \approx 49.4$ nm), the high-return transition case ($S_q \approx 79.0$ nm), and a rougher high-return case ($S_q \approx 118.5$ nm). Across all three, moving to matched detected counts consistently narrows the direct Q-noon2 gap to classical PSI without producing a hidden hybrid advantage. In the low-return case, the direct-Q RMSE falls from 1.40 to 0.86 nm; in the transition high-return case, it falls from 0.44 to 0.27 nm, and the same narrowing persists in the rougher high-return control. Classical PSI and H-noon2 remain numerically tied in all three zero-ambiguity scenarios. The result is therefore narrower than a claim of universal fair normalisation: budget convention matters quantitatively, and most strongly for the direct coincidence

branch, but it does not change the architectural conclusion that hybrid advantage appears only when ambiguity handling is actually needed.

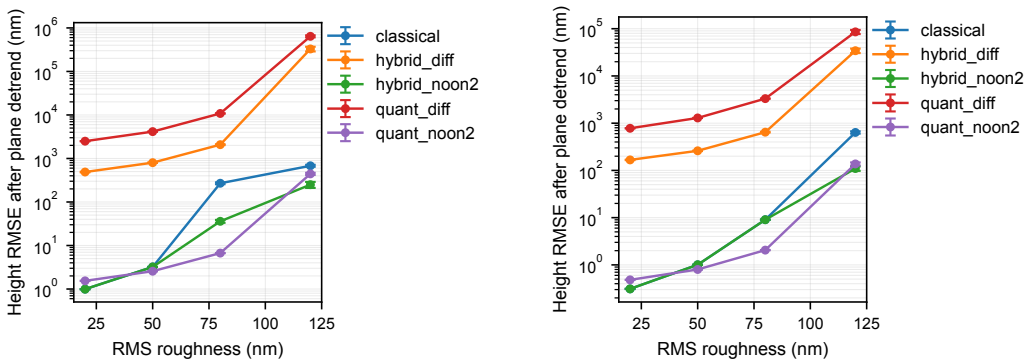

Figure 14. Promoted optical property sweep figures for the step-free case: height RMSE versus true surface roughness S_q for two bounding sample conditions. (**Left**): low reflectivity/low visibility ($R = 0.20$, $V_{\text{scale}} = 0.70$). (**Right**): high reflectivity/high visibility ($R = 1.00$, $V_{\text{scale}} = 1.00$). These two panels are retained in the main paper because they bracket the optical property range reported in the Supplementary Materials, Table S18, and make the method ordering visually explicit: classical PSI and H-noon2 define the low-to-moderate roughness baseline, the NOON-like branches become lower RMSE in the rougher high-return regime, and Q-diff/H-diff sacrifice fine-texture fidelity in exchange for a large effective wavelength.

3.3. Targeted Step/Ambiguity Regime Control

The step-free sweep above is intentionally blind to fringe-order ambiguity, so a complementary control was added on rough stepped surfaces with background roughness $S_q \approx 80$ nm and imposed vertical steps from 0 to 800 nm. Table S21 in the Supplementary Materials and Figure 15 report the absolute step-height error for low-return and high-return optical conditions. This control isolates the regime in which a long effective wavelength is valuable for metrology even though it is poor for fine-texture recovery.

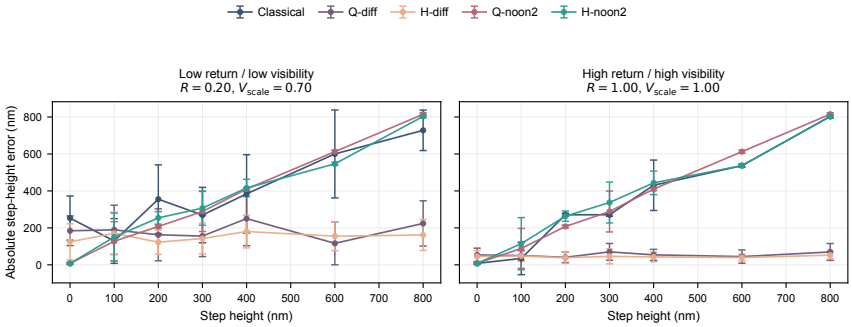

Figure 15. Targeted step/ambiguity control for rough stepped surfaces with background roughness $S_q \approx 80$ nm. The panels report absolute step-height error versus imposed step height for low-return and high-return conditions. The figure makes the architectural split explicit: long-effective-wavelength difference branches become the best coarse step estimators once ambiguity enters, whereas the NOON-like and short-wavelength branches remain limited by fringe-order ambiguity.

In the high-return case, once the imposed step reaches 200 nm the long-Λ branches clearly become the most accurate step estimators: the absolute step-height error drops from 270.8 nm for classical PSI to 40.7 nm for Q-diff and 38.4 nm for H-diff at 200 nm, and from 430.6 nm to 54.0 nm and 42.8 nm at 400 nm. The same pattern is noisier but still present in the low-return case, where at 600–800 nm Q-diff/H-diff stay within 116.4–224.1 nm and 155.1–161.6 nm, while classical PSI rises to 600.2–728.0 nm. By contrast, the NOON-like branches remain tied to the short-wavelength ambiguity burden and do not become the

preferred step estimators once the discontinuity grows. The missing use-case from the step-free sweep is therefore explicit here: long-Λ coincidence information is most useful when it is asked to resolve coarse discontinuity or fringe-order ambiguity, not when it is asked to carry the final fine-texture estimate.

3.4. Targeted Robustness Check Under Phase-Step Jitter and Frame Drift

Because coincidence-inspired channels are often discussed under idealised stepping conditions, two complementary non-ideality checks were added. Table S22 in the Supplementary Materials first promotes moderate phase-step jitter and frame drift into the full measured surface benchmark itself. At the level of full-dataset surface medians, neither perturbation materially changes the main height fidelity ordering: height RMSE remains at 1065.7 $\rightarrow$ 1054.9 $\rightarrow$ 1060.3 nm for classical PSI, 709.6 $\rightarrow$ 710.0 $\rightarrow$ 718.2 nm for the coincidence proxy branch, and 314.0 $\rightarrow$ 313.6 $\rightarrow$ 314.9 nm for the hybrid branch when moving from the baseline to $+3^\circ$ jitter and then to $+5\%$ drift. The roughness medians shift more for the direct coincidence branch, especially matched bandwidth $|\Delta S_z|$ (2509.1 $\rightarrow$ 3022.6 $\rightarrow$ 3547.7 nm), but the same architectural reading survives. The attempted mitigation gives the opposite message at benchmark scale: under the same 5% drift, least-squares phase recovery with frame mean normalisation remains benign for classical PSI (1049.3 nm) but catastrophically destabilises the coincidence proxy and hybrid branches (9.93×10^5 and 4.85×10^5 nm, respectively). The full measured surface benchmark therefore does not support the global transfer of that mitigation to coincidence-derived branches.

To understand why that mitigation still looked promising in the synthetic sweep, a second, intentionally local stress test was retained around the transition regime of Figure 14. Rather than running a second full factorial campaign, one representative step-free case was selected at $R = 1.00$, $V_{\text{scale}} = 1.00$, and $S_q = 80$ nm, i.e., the point where the ordering between classical PSI and the NOON-like variants first becomes practically important. Three controls already supported by the simulator were then applied: Gaussian phase-step jitter with $\sigma_\delta = 3^\circ$, frame-to-frame background and amplitude drift of 5%, and a mitigation control using least-squares reconstruction with frame mean normalisation under the same drift.

Table S23 and Figure S6 in the Supplementary Materials show that moderate phase-step jitter leaves the ranking essentially unchanged in the representative zero-step regime: the RMSE increase is about 1% for classical PSI and H-noon2, and about 3% for Q-noon2. By contrast, frame drift is the dominant non-ideality there, inflating the RMSE to 133.4 nm for classical PSI, 71.8 nm for H-noon2, and 47.3 nm for Q-noon2, with a marked rise in run-to-run variability. In the same moderate-roughness, high-return case, a simple LSQ plus frame mean normalisation step largely restores near-baseline performance (10.6 nm, 2.28 nm, and 10.6 nm, respectively). That local result is informative but not general: Table S22 in the Supplementary Materials shows that, once the full measured surface benchmark and ambiguity-bearing branches are reinstated, the same mitigation is not a safe benchmark-wide replacement workflow for the coincidence proxy or hybrid paths. The narrower lesson is therefore operational. Moderate phase-step jitter is not the main robustness threat in the present proxy hierarchy; illumination/background drift matters more, but any compensation step must be validated branch by branch rather than transferred automatically from the classical zero-step case.

3.5. Detector Model Sensitivity Under Matched Mean Coincidence Counts

Because the coincidence branch in the main benchmark is still represented by a simple phenomenological proxy, a second targeted check was added to the same step-free transition regime using the rate-based coincidence model under matched mean coincidence counts.

The scenarios isolate a baseline rate model, a finite coincidence window, accidentals plus efficiency imbalance, and accidentals plus a nonparalysable deadtime proxy; the gate time was fixed at 1 ms and the coincidence mean counts were matched to the simple proxy.

Table S24 and Figure S7 in the Supplementary Materials show that detector-side assumptions mainly perturb the direct Q-noon2 branch in this representative case. Relative to the matched-count rate baseline (1.03 nm), Q-noon2 increases to 2.19 nm with accidentals, 3.27 nm with additional efficiency imbalance, and 2.46 nm with deadtime. Classical PSI and H-noon2 remain numerically unchanged here because the selected step-free case contains no fringe-order ambiguity, so the hybrid estimator collapses to the classical short-wavelength branch. The detector-side study describes how moderate detector non-idealities erode the direct coincidence benefit before they materially alter the zero-step direct branch.

4. Discussion

The measured surface benchmark and the controlled optical property sweep establish an architectural role separation rather than a single universal winner. Hybrid reconstruction is favoured when the task is benchmark grid height fidelity under ambiguity, because the coincidence-derived observable selects the coarse fringe order while the short-wavelength PSI branch carries the final fine texture. Direct long-effective-wavelength reconstruction is useful only in narrower ambiguity or envelope-dominated regimes, where broad height excursions matter more than nm-scale texture fidelity.

That interpretation is supported by the two synthetic controls. In the step-free sweep, classical PSI and H-noon2 define the low-to-moderate-roughness fine-texture baseline, while the NOON-like coincidence branches become competitive in the rougher high-return regime. In the step/ambiguity control, the long-Λ difference branches become the best coarse step estimators once the imposed discontinuity enters the ambiguous regime, especially under high return. Together, these controls show why the coincidence-derived observable is best treated as an auxiliary coarse cue for rough surface reconstruction rather than as the default final height-map estimator.

The robustness checks add two practical caveats. Moderate 3° phase-step jitter and 5% frame drift do not materially alter the full-dataset median height ranking, but the direct coincidence proxy branch shows the clearest matched bandwidth roughness deterioration, especially in $|\Delta S_z|$. In addition, least-squares phase recovery with frame mean normalisation works in the representative zero-step transition regime but is not a safe global mitigation once ambiguous measured surfaces are reintroduced. For experimental follow-up, illumination/background stabilisation and branch-specific validation of any drift-compensation step are therefore more important than assuming that a classical mitigation transfers unchanged to coincidence-derived workflows [11–13,18].

The detector-side sensitivity checks bound the proxy result rather than validate hardware performance. In the representative rate model sweep, accidentals, efficiency imbalance, and deadtime mainly perturb the direct Q-noon2 branch. At the full measured surface level, the ideal matched-count rate model improves the direct coincidence proxy median height RMSE to 355.1 nm and the hybrid to 290.9 nm, but adding accidentals, efficiency imbalance, and deadtime degrades the direct branch to 1308.3 nm while the hybrid rises to 376.3 nm. Thus, a more physical count-formation model can shift quantitative boundaries, but within the audited regimes detector non-idealities erode direct coincidence-only gains faster than they erase the hybrid advantage.

The FV benchmark narrows the stronger coincidence claim further. Under the matched bandwidth roughness reference, the direct coincidence proxy channel no longer carries S_q; only the S_z advantage survives once the roughness comparison is referenced to the same area-averaged benchmark grid as the forward model. The native grid diagnostic

still matters because it shows how strongly the direct branch follows broad excursions when the reference bandwidth is widened, but it should be read as a stress test rather than as the primary roughness ranking. The auxiliary classical two-colour control goes one step further: much of the direct long-Λ envelope-following behaviour can be reproduced by a purely classical synthetic wavelength baseline. The added classical frontier oracle raises the comparator bar further still. Even when the best available classical workflow is selected per surface and endpoint, the hybrid branch retains the best median height RMSE, while the classical family recovers lower roughness-bias medians by switching between default, unwrap-stabilised, and two-colour baselines. The added wavelength spacing control reinforces the same point from the synthetic side: changing $\Delta\lambda$ shifts the severity of the Q-diff/H-diff penalty, but it does not recover a fine-texture regime in which the long-Λ difference channel rivals the short-wavelength classical baseline. The budget-sensitivity control adds a fairness caveat of a different kind: changing the resource-normalisation convention can move the direct NOON-like branch closer to or farther from the classical error floor, but it does not reverse the architectural reading that coincidence information is most defensible as a coarse prior rather than as a stand-alone fine-texture estimator. What remains distinctive in the present proxy benchmark is therefore not a unique direct long-Λ superiority, and not universal hybrid superiority over every roughness endpoint; rather, it is the fact that the coincidence-driven hybrid still yields the lowest fixed-workflow height RMSE while the roughness endpoints remain workflow- and process-dependent. The main roughness, paired-comparison, and classical frontier results, together with Tables S5, S6, S16, S19 and S20 and Figure S4 in the Supplementary Materials support that narrower reading.

Accordingly, the workflow message of this paper is benchmark-scoped but operationally clear. If one fixed workflow is forced within the present proxy simulator and the main endpoint is detrended height RMSE on the benchmark grid, the hybrid branch is the strongest of the tested architectures. If the endpoint is roughness bias, the result becomes an endpoint-specific regime map: matched bandwidth S_a and S_q favour hybrid within the three primary branches, while the remaining matched bandwidth S_z advantage of the coincidence proxy branch is treatment dependent and still breaches descriptor-scaled tolerance bands on 42.4% of surfaces. In that sense the roughness endpoints define the portability boundary of the architectural result.

The added image-domain, implementation, and uncertainty-propagation controls narrow that interpretation further. The residual maps in Figure 13 show that the failure of the direct long-Λ branch is not only a summary-statistic effect; it appears as a visible broad-structure distortion in the reconstructed height field. The full-dataset grid-sensitivity and unwrap controls in Figure S5 and Table S17 in the Supplementary Materials further indicate that the main hybrid height fidelity advantage is not generated solely by the particular 256×256 benchmark grid or by the default separable unwrap. Table 4 gives the corresponding numerical propagation view: the hybrid height–RMSE within-surface IQR is 0.9 nm, moderate phase jitter and frame drift shift the hybrid height median by less than 1 nm, and the non-ideal rate model shifts the hybrid height median by 62.3 nm while increasing the direct branch by 598.7 nm. The largest residual uncertainty in the benchmark therefore comes from surface-family composition, detector-side modelling, grid-dependent roughness behaviour, and S_z sensitivity rather than from stochastic repeatability alone.

Table 7 compresses the benchmark into the limiting modes that matter instrumentally. Long-Λ branches suppress texture, short-wavelength direct branches are ambiguity-limited, budget convention mainly shifts the apparent strength of the direct coincidence branch, frame drift dominates moderate phase-step jitter, detector non-idealities erode direct coincidence gains earlier than hybrid gains, and native grid S_z wins remain primarily

envelope-following events. Framed this way, the paper is both a ranking study and a screening map of which operating constraint is acceptable for a given metrology task.

Table 7. Failure-mode taxonomy distilled from the measured surface benchmark and the targeted control experiments. The table is intended as a screening-oriented synthesis: it identifies the dominant observable signature, the branch or benchmark component most exposed to the failure mode, the part of the paper where that mode is evidenced, and the corresponding follow-up implication.

Failure Mode	Primary Signature	Most Exposed Branch	Evidence in This Paper	Follow-Up Implication
Shared-prior self-reference	The same FV export defines both the simulator input and the measured surface scoring reference, so no external truth closure is achieved	All measured surface branches	Figures 2 and 7–9; Supplementary Materials, Figure S3	Treat the measured surface results as architecture screening evidence only and require independent reference measurement on every core validation case before any stronger claim.
Reference bandwidth mismatch	Native grid and benchmark grid roughness or PSD rankings diverge because area-averaged decimation is a low-pass filtering step	Roughness and PSD endpoints across all branches	Figures 7–9; Supplementary Materials, Figure S3; Supplementary Materials, Tables S5 and S6	Use matched bandwidth benchmark grid descriptors as the only main-text roughness basis; keep native grid and benchmark grid PSD views as diagnostics and document the decimation operator explicitly.
Long-Λ texture suppression	Broad residual structures and poor fine-texture recovery despite enlarged ambiguity range	Q-diff, then H-diff	Figure 13; Supplementary Materials, Tables S16 and S19	Treat long-Λ information as a coarse prior or envelope cue, not as the primary nm-scale texture estimator.
Ambiguity/unwrap failure	Step-height error rises once the short-wavelength branch must bridge larger discontinuities; direct unwrap choice changes full-dataset medians	Classical direct branch and direct coincidence branches	Supplementary Materials, Tables S17 and S21	Prefer coarse-to-fine hybrid unwrapping or at least verify direct-branch claims against a global unwrap control.
Budget-convention sensitivity	Direct Q-noon2 moves toward or away from the classical floor when the resource normalisation is changed, while the hybrid ordering stays stable	Direct Q-noon2	Supplementary Materials, Table S20	Phrase low-light or fair-budget claims as convention-sensitive; do not infer a universal resource advantage from one normalisation.
Drift sensitivity and mitigation non-transfer	Representative zero-step cases are drift-dominated, but LSQ plus frame normalisation does not transfer safely to coincidence-derived branches on the full measured benchmark	All phase-stepped branches under drift; mitigation failure is strongest in the coincidence proxy and hybrid paths	Supplementary Materials, Tables S22 and S23; Supplementary Materials, Figure S6	Prioritise illumination stabilisation ahead of only tightening the phase-step actuator, and validate any drift-compensation workflow separately for classical and coincidence-derived branches.
Detector-side coincidence erosion	Accidentals, efficiency imbalance, and deadtime erode direct coincidence gains before they create a new hybrid win	Direct coincidence branch	Table 6; Supplementary Materials, Table S24	Propagate detector losses through the design model and judge coincidence channels mainly by whether they still help a hybrid estimator.
Selective envelope-only descriptor gain	Matched bandwidth S_z can remain favourable even when descriptor-scaled tolerance exceedance remains substantial	Direct coincidence proxy branch	Supplementary Materials, Tables S7 and S12	Interpret S_z wins as selective envelope following, not as broad roughness-fidelity improvement.

This distinction is important because it changes the level at which the coincidence-derived contribution should be judged. Much of the surrounding quantum and quantum-inspired sensing literature is framed in terms of phase sensitivity, enlarged unambiguous range, or idealised resolution arguments [22–24,27,29]. Those are valuable starting points, but they do not by themselves answer the metrology question addressed here: whether a given architecture improves the reconstructed surface map or the downstream texture descriptor that an engineer would actually report. By translating the comparison into endpoint-specific accuracy, uncertainty, limiting-case incidence, and PSD fidelity on realistic measured surfaces, and by setting that comparison against the corresponding classical robustness literature [11–14,21], the present study moves the contribution from abstract sensing potential to a stress-tested screening benchmark inside an explicitly limited proxy model.

Taken together, the grouped material, treatment, roughness parameter, paired-effect, dominance, and control comparisons define the follow-up prioritisation map summarised in Table 3. Within the present proxy benchmark, the first architecture to test experimentally is the hybrid branch when the primary endpoint is benchmark grid height RMSE. The long-Λ difference branches are retained for ambiguity-tolerant coarse step estimation, while direct coincidence proxy reconstruction is most relevant in the narrower envelope-dominated regime where matched bandwidth S_z is the endpoint of interest. The same regime map suggests a secondary methodological opportunity: uncertainty-aware method selection from observable surface descriptors, consistent with recent data-driven work on measurement technique preselection, calibrated surface parameter prediction, and broader coordinate metrology decision support [33–35].

The classical two-colour control makes that interpretation stricter: broad-envelope following alone does not justify a quantum claim, because much of that behaviour is already reproducible with a purely classical synthetic wavelength baseline. The strongest contribution that survives the full control set is therefore architectural and experimentally useful. The paper identifies which role remains defensible for coincidence-derived information after realistic texture endpoints and stronger classical controls are imposed, and that narrower reading is also the one most compatible with the current implementation trajectory of integrated quantum photonics and correlated photon instrumentation [30–32].

Beyond surface metrology itself, the phase-screen formulation used here may also be useful as a structured robustness benchmark for interferometric architectures that target extremely small phase or path-length perturbations, where stability, ambiguity handling, and readout design matter alongside nominal phase sensitivity [36–38].

The experimental validation subset in Table 8 contains six core and two extended surfaces that span the best hybrid case, a representative median case, the least favourable direct-branch case, the grouped rough-turning and rough-WEDM exceptions, the classical two-colour comparator case, the detector-sensitivity case, and the strongest native grid S_z envelope-following case. The corresponding protocol specifies repeated acquisition of the classical, classical two-colour, direct coincidence, and hybrid branches on the same field of view, together with matched bandwidth reporting, an uncertainty log, and an independent reference cross-check on every core case. That cross-check is required for experimental validation and should use an appropriate independent reference route, for example calibrated CSI, stylus profilometry, AFM, or another traceable areal surface-measurement chain matched to the surface scale.

Table 8. Reduced experimental validation subset for the next revision-critical laboratory follow-up. The core rows are the minimum set needed to challenge the main architectural claim on real coincidence measurements; the extended rows add the second grouped exception family and the strongest native grid envelope-following case. Raw stems retain the original measurement-file identifiers, while the English material and treatment descriptions are given below each stem.

Tier	Case	Selected Surface	Purpose in Reduced Validation
Core	Best hybrid height	`P1-1.4301_wedm_zgru` Stainless steel 1.4301 (AISI 304); Wire EDM (roughing)	Anchor best-case replication of the main fixed-workflow height–RMSE claim.
Core	Median hybrid height	`P1-Ti6A14V_frez_zgr` Titanium; Milling (roughing)	Anchor a typical-case replication near the benchmark median.
Core	Catastrophic direct-Q failure	`1.4301_szlifowane` Stainless steel 1.4301 (AISI 304); Grinding	Demonstrate the failure mode that most strongly limits direct coincidence-only reconstruction.
Core	Turning-roughing exception	`P1-C45_t_zgrubne` Steel C45; Turning (roughing)	Test one treatment family where the direct branch is a grouped height–RMSE exception.
Core	Two-colour non-uniqueness	`P1-1.4301_frez_zgrub` Stainless steel 1.4301 (AISI 304); Milling (roughing)	Test whether the envelope-following claim can be reproduced by a purely classical synthetic wavelength baseline.
Core	Detector fragility	`Graphite_szkielkowane` Graphite; Glass bead blasted	Test whether realistic detector non-idealities hurt the direct branch more than the hybrid branch.
Extended	Wire-EDM roughing exception	`P1-Ti6A14V_wedm_zgru_1prz` Titanium; Wire EDM (roughing)	Add the second grouped height–RMSE exception family from the manuscript.
Extended	Native grid S_z envelope	`C45_t_wyk` Steel C45; Turning (finishing)	Add a native grid Sz case where the direct branch has its clearest envelope-following advantage.

Scope of the Benchmark

The study combines measured FV topographies with simulated classical and coincidence-style observations. The coincidence branch is represented by a proxy observable and rate model sensitivity controls; experimental follow-up will require a calibrated fourth-order transfer function, measured mode overlap, measured visibility closure, detector timing response, dark counts, accidentals, channel efficiencies, deadtime response, and alignment stability. The FV maps serve as operational benchmark surfaces, so the roughness and PSD results reflect FV bandwidth, outlier sensitivity, and the deterministic low-pass effect of the benchmark grid reduction. The matched bandwidth benchmark grid descriptors are therefore used as the primary roughness comparison, while native grid descriptors provide a reference sensitivity view. The conclusions should be read within the audited simulation regimes and not as predictions for detector-dark-count-limited or poorly mode-matched coincidence hardware.

The statistical summaries describe the fixed measured surface dataset. Bootstrap intervals, material/treatment holdouts, paired effects, and Holm-adjusted Wilcoxon tests are reported as dataset-level evidence. The treatment holdouts are strongest for height RMSE and remain less stable for roughness, with matched bandwidth agreement of 70% for $|\Delta S_a|$, 60% for $|\Delta S_q|$, and 60% for $|\Delta S_z|$. The matched-count convention, detector efficiency, accidentals, and deadtime also shape the direct coincidence proxy branch, as shown by the non-ideal rate control. Within the tested surface family, the evidence therefore supports the hybrid branch as the preferred fixed-workflow height fidelity architecture and treats roughness outcomes as process- and endpoint-dependent descriptors.

The approximate Gaussian S/L roughness filter control gives the same boundary map. Filtered treatment holdout agreement remains at 60% for all three descriptors, and filtered $|\Delta S_q|$ material portability weakens rather than strengthens. The filtered comparison therefore preserves the roughness endpoint split observed in the detrend-only descriptor path.

5. Conclusions

This benchmark establishes a specific architecture selection result for rough surface interferometric profilometry. Using 59 real FV height maps exported as Mountains/

DigitalSurf `.sur` files as operational geometric priors, together with simulated classical and coincidence-style observations generated from those surfaces, the study shows that coincidence-derived information is most useful when it supplies a coarse ambiguity-resolving prior inside a hybrid estimator. The final fine texture is still reconstructed by the short-wavelength PSI branch.

The fixed-workflow height fidelity result is robust across the main controls. In the measured surface benchmark, the hybrid branch gives a surface-level median detrended height RMSE of 314.0 nm and 32 of 59 per-surface wins. A strengthened classical frontier comparator remains higher at 559.1 nm, and the rate-based coincidence control preserves the same ordering, with the hybrid median at 290.9 nm under ideal matched-count rates and 376.3 nm under detector non-idealities. Roughness endpoints define the portability boundary of the result: matched bandwidth S_a and S_q favour hybrid within the primary branches, while the matched bandwidth S_z advantage of the direct coincidence proxy branch remains treatment dependent.

The resulting contribution is a controlled photonic architecture benchmark rather than a hardware-validation claim. The classical two-colour and classical frontier controls show that broad long-wavelength envelope following is not sufficient evidence for overall architecture-level superiority within this simulation benchmark, while the hybrid result identifies the most credible configuration for experimental follow-up when benchmark grid height fidelity is the primary endpoint. A first hardware study should replace the proxy and rate-control hierarchy with a measured fourth-order transfer function, propagate detector-side losses, accidentals, and phase-stepping imperfections under laboratory conditions, and test the hybrid branch against an independent reference chain on the reduced measured surface subset in Table 8. The public repository and companion release archive preserve the manuscript sources, derived benchmark grid surfaces, public CSV/table record, and regeneration commands for that follow-up stage.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/photonics13060526/s1, Supplementary File S1: detailed benchmark audit tables and diagnostic figures.

Funding: This work was supported by a grant from the Polish Ministry of Science under the programme "Polish Metrology II" (Polska Metrologia II, Poland), project no. PM-II/SP/0090/2024/02.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The project repository contains the manuscript sources, simulation code, and generated benchmark artefacts cited in the paper. The public repository at https://github.com/dawidkucharski/Quantum_interference (accessed on 25 May 2026) provides the reproducibility code base and release archive for the benchmark workflow. Canonical regeneration is described in the public reproducibility companion at `docs/reproducibility_companion.md` and in the command record distributed with the companion archive. The main executable scripts are in `scripts/`; reusable Python modules (Python 3.9.6; qiprof 0.1.0; NumPy 2.0.2; SciPy 1.13.1; Matplotlib 3.9.4; Pillow 11.3.0; scikit-learn 1.6.1) are in `src/qiprof/`; derived measured surface CSVs, tables, figures, and the resolution-sensitivity summaries are under `outputs/paper_alicona_benchmark/`; synthetic optical property, non-ideality, and detector-control artefacts are under the corresponding `outputs/paper_optprops_*` directories; generated manuscript table sources are under `manuscript/tables/`. The reduced experimental validation subset is included under `outputs/paper_alicona_benchmark/experimental_validation/` together with the accompanying protocol in `docs/experimental_validation_protocol.md`. The public release archive contains derived 256 × 256 benchmark grid surfaces, public CSV summaries, generated table sources, regeneration commands, environment information, and checksum information. The repository states a split

open-access license: source code is distributed under the MIT License, while public documentation and derived benchmark artefacts are distributed under CC BY 4.0. The full native `.sur` topographies are not distributed through the Git repository because the dataset is too large for a practical public GitHub snapshot; these native files are available from the corresponding author upon reasonable request and are not covered by the public repository license unless separately distributed with an explicit license notice. Future experimental validation data should be archived separately with the corresponding independent reference measurements.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DOI	digital object identifier
FV	focus variation
GUM	Guide to the Expression of Uncertainty in Measurement
NOON	path-entangled multi-photon state shorthand used here for an effective half-wavelength proxy
PSD	power spectral density
PSI	phase-shifting interferometry
RMSE	root mean square error
S/L	short-/long-wavelength areal roughness filtering notation
SUR	Mountains/DigitalSurf surface file format
WEDM	wire electrical discharge machining

References

1. Vorburger, T.; Teague, E. Optical techniques for on-line measurement of surface topography. *Precis. Eng.* **1981**, *3*, 61–83. [CrossRef]
2. Dong, Y. An overview of optical methods for in-process and on-line measurement of surface roughness. In Proceedings of the 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, Changchun, China, 24–26 August 2010; Volume 4, pp. 35–37. [CrossRef]
3. Harding, K.G. Laser scatter surface finish measurement techniques. In *Proceedings of the ICALEO '91: Proceedings of the Optical Sensing and Measurement Symposium*; Laser Institute of America: Orlando, FL, USA, 1992; Volume 73, pp. 125–135. [CrossRef]
4. Mozer, W. Surface metrology for stent and implant manufacturing. *Med. Device Diagn. Ind.* **2006**, *28*, 72–79.
5. Zhang, Y.; Ba, D.; Yang, Y.; Dong, Y. Generalized Doppler effect for high-accuracy frequency shift measurement. *Light. Sci. Appl.* **2026**, *15*, 197. [CrossRef]
6. Zhang, Y.; Li, H.; Ba, D.; Dong, Y. Measuring angular velocity through atmospheric turbulence with rotational Doppler-shifted intervals. *Opt. Laser Technol.* **2025**, *192*, 114124. [CrossRef]
7. *ISO 25178-2:2021*; Geometrical Product Specifications (GPS)–Surface Texture: Areal–Part 2: Terms, Definitions and Surface Texture Parameters. International Organization for Standardization: Geneva, Switzerland, 2021.
8. *ISO 16610-61:2015*; Geometrical Product Specifications (GPS)–Filtration–Part 61: Linear Areal Filters: Gaussian Filters. International Organization for Standardization: Geneva, Switzerland, 2015.
9. *ISO 25178-606:2015*; Geometrical Product Specifications (GPS)–Surface Texture: Areal–Part 606: Nominal Characteristics of Non-Contact (Focus Variation) Instruments. International Organization for Standardization: Geneva, Switzerland, 2015.
10. *JCGM 100:2008*; Evaluation of Measurement Data–Guide to the Expression of Uncertainty in Measurement. GUM 1995 with Minor Corrections. Joint Committee for Guides in Metrology: Sevres, France, 2008.
11. McDonnell, E.M.; Deck, L.L. Solutions for environmentally robust interferometric optical testing. In Proceedings of the Optical Manufacturing and Testing XIII, Online, 24 August–4 September 2020; Volume 11487. [CrossRef]
12. Ri, S.; Takimoto, T.; Xia, P.; Wang, Q.; Tsuda, H.; Ogihara, S. Accurate phase analysis of interferometric fringes by the spatiotemporal phase-shifting method. *J. Opt.* **2020**, *22*, 105703. [CrossRef]
13. Medina, O.; Estrada, J.C. Full-field two-dimensional least-squares method for phase-shifting interferometry. *Opt. Eng.* **2014**, *53*, 114106. [CrossRef]
14. De Groot, P.J. 101-Frame algorithm for phase shifting interferometry. In *Proceedings of the Optical Inspection and Micromeasurements II*; SPIE: Bellingham, WA, USA, 1997; Volume 3098, pp. 283–292. [CrossRef]

15. Creath, K. Measurement considerations for precise, highly reflective surfaces using a phase-measuring Fizeau interferometer. In Proceedings of the 18th Congress of the International Commission for Optics, San Francisco, CA, USA, 2–6 August 1999; Volume 3749, pp. 182–183. [CrossRef]
16. Younus, I. Image enhancement for the phase stepped interferometric process by appropriate filtering. In *Proceedings of the IEEE 1998 National Aerospace and Electronics Conference, NAECON 1998, Celebrating 50 Years (Cat. No.98CH36185)*; IEEE: Minneapolis, MN, USA, 1998; pp. 600–603. [CrossRef]
17. Jeon, J.; Kim, Y.; Sugita, N. Deep learning-driven virtual phase shifting for enhanced interferometric surface profiling. *Precis. Eng.* **2026**, *97*, 757–766. [CrossRef]
18. Kim, H.; Kim, Y.; Ito, Y.; Sugita, N. Robust Zernike aberration sensing method based on deep learning for precision interferometric glass thickness profiling. *Meas. J. Int. Meas. Confed.* **2026**, *263*, 120062. [CrossRef]
19. Kotsiuba, Y.; Petrovska, H.; Fitio, V.; Bobitski, Y. Digital interferometry methods for the surface relief study. In *Proceedings of the Nanooptics, Nanophotonics, Nanostructures, and Their Applications*; Springer International Publishing: Cham, Switzerland, 2018; Volume 210, pp. 207–217. [CrossRef]
20. Zhang, S.; Lou, Y.; Zhou, Y.; Chen, B.; Yan, L. Self-correction of air refractive index in distance measurement using two-color sinusoidal phase modulating interferometry. *Opt. Commun.* **2022**, *505*, 127521. [CrossRef]
21. Okamoto, R.; Tahara, T. Precision limit for simultaneous phase and transmittance estimation with phase-shifting interferometry. *Phys. Rev. A* **2021**, *104*, 033521. [CrossRef]
22. Kapale, K.T.; Didomenico, L.D.; Lee, H.; Kok, P.; Dowling, J.P. Quantum interferometric sensors. In *Proceedings of the Noise and Fluctuations in Photonics, Quantum Optics, and Communications*; SPIE: Bellingham, WA, USA, 2007; Volume 6603. [CrossRef]
23. Didomenico, L.D.; Lee, H.; Kok, P.; Dowling, J.P. Quantum interferometric sensors. In *Proceedings of the Quantum Sensing and Nanophotonic Devices*; SPIE: Bellingham, WA, USA, 2004; Volume 5359, pp. 169–176. [CrossRef]
24. Xu, L.; Zhang, L. Progress and perspectives on weak-value amplification. *Prog. Quantum Electron.* **2024**, *96*, 100518. [CrossRef]
25. Rarity, J.; Burnett, J.; Tapster, P.; Paschotta, R. High-visibility two-photon interference in a single-mode-fibre interferometer. *EPL* **1993**, *22*, 95–100. [CrossRef]
26. Richards, R.K. Quantum-entangled photon interferometry. In *Proceedings of the Interferometry XII: Techniques and Analysis*; SPIE: Bellingham, WA, USA, 2004; Volume 5531, pp. 17–23. [CrossRef]
27. Jha, A.K.; Agarwal, G.S.; Boyd, R.W. Supersensitive measurement of angular displacements using entangled photons. *Phys. Rev. A-At. Mol. Opt. Phys.* **2011**, *83*, 053829. [CrossRef]
28. Vitelli, C.; Spagnolo, N.; Toffoli, L.; Sciarrino, F.; De Martini, F. Enhanced resolution of lossy interferometry by coherent amplification of single photons. *Phys. Rev. Lett.* **2010**, *105*, 113602. [CrossRef] [PubMed]
29. Sharma, A.N.; Krafczyk, C.; Jordan, A.N.; Kwiat, P.G. Enhanced-Precision Displacement Measurements Using Position-Entangled Photon Pairs. In *Proceedings of the Conference on Lasers and Electro-Optics*; Optica Publishing Group: Washington, DC, USA, 2022. [CrossRef]
30. Mahmudlu, H.; Johanning, R.; van Rees, A.; Khodadad Kashi, A.; Epping, J.P.; Haldar, R.; Boller, K.J.; Kues, M. Fully on-chip photonic turnkey quantum source for entangled qubit/qudit state generation. *Nat. Photonics* **2023**, *17*, 518–524. [CrossRef]
31. Labbe, F.; Ekici, C.; Zhdanov, I.; Muthali, A.L.; Oxenlowe, L.K.; Ding, Y. Thin-film lithium niobate quantum photonics: Review and perspectives. *Adv. Photonics* **2025**, *7*, 044002. [CrossRef]
32. Yin, L.; Hu, Y.; Zheng, X. Measured uncertainty analysis of detection efficiency of single photon detectors with correlated photons calibration. *Appl. Opt.* **2022**, *61*, 1316–1322. [CrossRef]
33. Kucharski, D.; Gąska, A.; Kowaluk, T.; Stępień, K.; Rępalska, M.; Gapiński, B.; Wieczorowski, M.; Nawotka, M.; Sobecki, P.; Sosinowski, P.; et al. Multi-Task Deep Learning for Surface Metrology. *Sensors* **2025**, *25*, 7471. [CrossRef]
34. Kucharski, D.; Gapiński, B.; Wieczorowski, M.; Gąska, A.; Sładek, J.; Kowaluk, T.; Rępalska, M.; Tomasik, J.; Stępień, K.; Makięła, W.; et al. Machine Learning-Based Selection of Measurement Technique for Surface Metrology: A pilot study. *Metrol. Hallmark* **2024**, *1*, 1–9.
35. Wieczorowski, M.; Kucharski, D.; Śniatała, P.; Pawlus, P.; Królczyk, G.; Gapiński, B. A novel approach to using artificial intelligence in coordinate metrology including nano scale. *Measurement* **2023**, *217*, 113051. [CrossRef]
36. The LIGO Scientific Collaboration. Advanced LIGO. *Class. Quantum Gravity* **2015**, *32*, 074001. [CrossRef]
37. Amelino-Camelia, G. Gravity-wave interferometers as quantum-gravity detectors. *Nature* **1999**, *398*, 216–218. [CrossRef]
38. Chou, A.; Glass, H.; Gustafson, H.R.; Hogan, C.J.; Kamai, B.L.; Kwon, O.; Lanza, R.; McCuller, L.; Meyer, S.S.; Richardson, J.W.; et al. Interferometric constraints on quantum geometrical shear noise correlations. *Class. Quantum Gravity* **2017**, *34*, 165005. [CrossRef]