

NIRPipe: A Modular, Instrument-Agnostic Pipeline for the Reduction and Photometric Calibration of Dithered Near-Infrared Imaging

¹Radhika Gokani, ²Vishal Joshi

¹Department of Computer Science, ²Astronomy & Astrophysics Division

¹Swaminarayan University, Ahmedabad, India, ²Physical Research Laboratory, India

Corresponding author: radhika@***.com

Abstract : Near-infrared (NIR) imaging in the J, H and K bands is almost universally acquired as a dithered sequence, because the sky background at these wavelengths is bright, spatially structured and variable on timescales of minutes, and so cannot be removed with a static calibration frame the way an optical bias or flat can. Reducing such data — grouping frames by telescope pointing, rejecting corrupted exposures, deriving and subtracting a time-appropriate sky model, solving the astrometry, and calibrating instrumental photometry against a reference catalogue — is a well-understood but fragmented process: most existing tools are written for a single camera, a single telescope, or a single survey, and porting them to new hardware is largely a manual exercise. This paper describes NIRPipe, an open, instrument-agnostic implementation of the full reduction-to-calibrated-catalogue chain, in which every piece of camera-specific knowledge — header keyword names, pixel scale, detector gain and read noise — is isolated in a small declarative configuration file rather than embedded in code. We describe the nine constituent processing stages, including two design choices we treat as central to the system's robustness: dither-position grouping is performed purely from image-domain registration rather than from telescope pointing metadata, and the astrometric solution is obtained by matching the pattern of detected stars against a reference catalogue rather than by refining a header-supplied pointing. We validate the complete pipeline against two simulated instruments with deliberately different plate scales, detector properties and header conventions, using datasets with fully known ground truth, and report quantitative results for frame-quality classification (96.7–100% agreement with injected defects), astrometric residuals (0.002–0.03 arcsec), and photometric calibration accuracy (0.01–0.05 mag scatter against the input catalogue). We close with a candid account of the system's current limitations — in particular a sensitivity of star-pattern matching to how the candidate reference catalogue is spatially filtered before matching — and the engineering lessons that shaped the final design.

IndexTerms - Near-infrared imaging, astronomical data reduction, image registration, sky subtraction, astrometric calibration, photometric calibration, software pipeline, instrument abstraction.

I. INTRODUCTION

Ground-based near-infrared imaging occupies an awkward position in the instrumentation landscape. The atmosphere is bright and highly variable in J, H and K(s) — dominated at the shorter of these wavelengths by OH airglow lines that can change by tens of per cent on a timescale of minutes, and at the longer wavelengths by thermal emission from the telescope and sky itself — so, unlike optical imaging, a single dark-and-flat calibration cannot remove the background. The standard observational remedy is applied at the telescope rather than in software: the target is observed at a sequence of n dithered pointings, each offset from the last by anywhere from a few arcseconds to a few arcminutes, and at each pointing m short exposures are typically taken. Dithering serves two purposes at once. It allows bad pixels and cosmic-ray hits to be rejected by combination, as in any imaging survey, and — more distinctively for NIR work — it allows the sky background itself to be estimated directly from the science frames, since the sky pattern is fixed in detector coordinates while the astronomical sources move from exposure to exposure.

This is well-trodden ground. Self-referenced sky subtraction from dithered NIR frames has been standard observatory practice for decades and is implemented, in one form or another, in essentially every camera-specific reduction package in current use: DRAGONS for Gemini's imagers, the VISTA Data Flow System for VIRCAM, and a long tail of IRAF and PyRAF scripts written for instruments that have since been retired. What is comparatively rare is doing this in a way that is not tied to the instrument it was written for. Each of the packages above hard-codes, to varying degrees, the header keywords, pixel scale, gain and detector idiosyncrasies of one camera; retargeting them at a new instrument is a manual, often substantial, rewrite rather than a configuration change. For a group that operates — or collaborates across — more than one telescope, or that has simply lived through a camera upgrade, this is a genuine, if unglamorous, engineering cost, and it is the first problem this paper addresses directly.

A second observation, more specific and in our view somewhat under-discussed in the reduction-pipeline literature, motivates the rest of this paper's technical contribution. Two of the steps in a typical reduction — grouping exposures by dither position, and solving the astrometry — are conventionally treated as effectively solved by the telescope's own pointing metadata: frames are grouped using the commanded or recorded RA/Dec, and the world coordinate system (WCS) is obtained by refining that same recorded pointing against a catalogue. This is a reasonable default when the metadata is good, but it is not always good, and a pipeline that quietly assumes it is will produce a plausible-looking, silently wrong answer under exactly the conditions — pointing-model drift, a guide-star acquisition failure, a mount encoder fault — where getting it right matters most. We take the harder line of not using recorded pointing for either of these two steps at all, other than as a coarse hint for how wide a catalogue search to run, and this decision, together with the instrument-abstraction point above, shapes most of the rest of the system's design.

This paper describes NIRPipe, a from-scratch implementation of the full reduction chain — from raw, possibly multi-extension, FITS frames through to a 2MASS-calibrated, multi-band photometric catalogue — built around these two commitments. Every

stage that needs to know something about the camera reads it from a small per-instrument configuration object rather than from code, so that supporting a new camera is a matter of authoring a profile rather than patching the pipeline (Section 3.2). Dither-position grouping is derived purely from cross-correlation of the science images themselves (Section 4.2), and the astrometric solution is obtained by matching the pattern of detected stars against a reference catalogue, with recorded pointing used, when available, only to bound the extent of that catalogue query (Section 4.5).

We make the following contributions:

- an end-to-end, openly structured reduction pipeline covering source detection, dither grouping, automated quality control, sky-frame construction, astrometric solution and photometric calibration, organised around a declarative instrument-profile abstraction (Section 3);
- a self-calibrating quality-control scheme that flags corrupted frames — blurred, under-exposed, or affected by a mid-exposure telescope slip — using thresholds derived from the statistics of each observing batch rather than fixed, instrument-specific numbers (Section 4.3);
- a lightweight astrometric solver that matches star patterns against a reference catalogue without relying on telescope pointing, together with a practical finding — not, to our knowledge, discussed explicitly elsewhere — that restricting the candidate catalogue to the actual detector footprint under a coarse guess WCS is what makes triangle-based pattern matching reliable in practice, considerably more so than simply querying a generous cone (Sections 4.5 and 7.2);
- a quantitative, reproducible evaluation of the complete chain against two synthetic instruments with deliberately different plate scales, detector properties and header conventions, with independently verifiable injected ground truth at every stage (Section 6).

The remainder of the paper is organised as follows. Section 2 situates this work relative to existing NIR reduction software. Section 3 describes the overall architecture and the instrument-abstraction mechanism. Section 4 works through the nine processing stages in the order data flows through them. Section 5 covers implementation choices. Section 6 presents the experimental evaluation. Section 7 discusses limitations, including a threats-to-validity style analysis of what the simulated evaluation does and does not establish. Section 8 concludes.

II. RELATED WORK

Automated reduction of dithered NIR imaging is a mature sub-field with a substantial software lineage, and it is worth being precise about where NIRPipe sits relative to it, rather than implying it supersedes tools that are, within their own domain, considerably more capable.

2.1 Instrument-specific pipelines

The most heavily used NIR reduction packages are tied to a specific camera or observatory. DRAGONS is Gemini Observatory's Python-based successor to the IRAF Gemini package [13] and handles FLAMINGOS-2, GNIRS and NIRI imaging with considerable sophistication, including calibration-frame association logic specific to Gemini's own archive and non-linearity correction that operates on raw detector counts before any sky or flat processing begins. The VISTA Data Flow System (VDFS), developed at the Cambridge Astronomical Survey Unit [12], reduces VIRCAM data for the VISTA public surveys and is deeply optimised for that camera's sixteen-detector, non-contiguous mosaic. Both tools are, within their domain, more thorough than the system presented here — DRAGONS in particular performs detector-linearity and distortion corrections that NIRPipe does not attempt — but that thoroughness is bought at the cost of portability: neither generalises to a camera it was not written for without substantial engineering effort.

2.2 General-purpose reduction toolkits

A second category, typified by THELI [8, 15] and the Astropy-affiliated ccdproc package, provides configurable building blocks — image combination, cosmic-ray rejection, WCS handling — that a user assembles into a pipeline for their own instrument. This is closer in spirit to our approach; the difference is one of scope and defaults. THELI is oriented primarily toward wide-field optical imaging, and although it does support near- and mid-infrared data, that is not its primary design target. More importantly, neither toolkit ships an opinionated, ready-to-run NIR-specific chain — self-referenced sky subtraction, dither grouping, forced multi-frame photometry, catalogue calibration — the user still assembles one from parts, which is precisely the recurring cost NIRPipe is intended to remove.

2.3 Astrometric calibration

Blind or semi-blind astrometric solving in practice is dominated by Astrometry.net [14], which indexes reference-catalogue star patterns into scale- and rotation-invariant quads and can solve genuinely blind fields with no initial pointing guess at all. We deliberately did not build on Astrometry.net here, for a practical rather than technical reason: constructing or shipping the index files needed to cover a useful sky area is a heavier deployment dependency than we wanted for a pipeline meant to be pointed easily at a new, comparatively small NIR field. Instead we use a lighter, coarse-guess-assisted matcher built on the asterism/triangle-invariant algorithm implemented in astroalign [4], which trades the ability to solve with no positional information at all for a much smaller footprint — and, as we discuss in Section 7.2, a matching reliability that turns out to depend on exactly how the candidate catalogue is filtered before matching begins, a detail we did not find well documented elsewhere.

2.4 Photometric calibration against 2MASS

Tying instrumental NIR photometry to the Two Micron All Sky Survey point-source catalogue [16] is close to universal practice for ground-based J/H/K work at the depths reachable from a single-night pointing, and is the approach NIRPipe follows. Our contribution here is not a new calibration method but an implementation that degrades gracefully, rather than failing outright, when the live catalogue service is unreachable, falling back to a locally supplied catalogue table with identical downstream handling. We found this a genuinely useful property during development, which took place in a network-restricted environment (Section 5).

Taken together, NIRPipe's position is narrower than a survey-grade observatory pipeline and broader than a component library: it is a complete, opinionated, but explicitly instrument-agnostic chain for the common case of dithered J/H/K imaging with self-referenced sky subtraction, aimed at groups that need to point roughly the same reduction logic at more than one camera without rewriting it each time.

III. SYSTEM DESIGN

3.1 Design goals

We set out with four goals, stated here because they explain several choices later in the paper that might otherwise look arbitrary. First, instrument portability: no module outside the abstraction layer described in Section 3.2 should contain a literal FITS header keyword string. Second, untrusted metadata: the pipeline should degrade gracefully, not silently, when pointing information is imprecise, which we treat as the ordinary case rather than an exceptional one (Section 4.2, 4.5). Third, traceability: every automatic decision the pipeline makes — a frame rejected, a dither position assigned, a star excluded from a zero-point fit — should be logged with a human-readable reason, not merely a boolean flag. Fourth, testability: every stage should be independently exercisable against data with known ground truth, which is what motivated building a synthetic-data generator alongside the pipeline itself, rather than treating evaluation data as an afterthought to be sourced later.

3.2 The instrument-profile abstraction

Every module downstream of file input/output interacts with a header value through a single accessor, which looks a requested logical field — filter, exposure time, and so on — up in a small per-instrument mapping table rather than assuming a specific keyword name. The same mechanism translates whatever string a camera writes for a filter name (for example “Ks-band”, “K_SHORT”, or simply “Ks”) into one of the three normalised band labels the rest of the pipeline reasons about. Four scalar properties — plate scale, gain, read noise, and the ADU-level saturation threshold — are treated as fixed attributes of the instrument profile rather than read from the header at all. In our experience these are exactly the fields most likely to be missing, stale, or simply incorrect in raw engineering headers, and treating them as configuration rather than as data the pipeline must trust avoids building failure modes around header fields it has no independent way to verify.

We stress-tested this abstraction, rather than simply asserting that it works, by writing two instrument profiles for the evaluation in Section 6 that differ in essentially everything the abstraction is meant to hide from the rest of the codebase: different header keyword names for every logical field, including one field — the filter keyword — deliberately given a non-standard, HIERARCH-convention name specifically to exercise that code path; different filter-name spellings; and an almost sixfold difference in pixel scale. Table 1 summarises both profiles. No line of code outside the two configuration files differs between a run against Instrument A and a run against Instrument B.

Table 1. The two instrument profiles used throughout the evaluation in Section 6, chosen to differ in every quantity the abstraction layer is designed to isolate.

Property	Instrument A (this work)	Instrument B (this work)
Pixel scale	1.00 arcsec / pixel	0.18 arcsec / pixel
Detector gain	4.2 e ⁻ / ADU	2.6 e ⁻ / ADU
Read noise	18.0 e ⁻	12.0 e ⁻
Saturation level	28,000 ADU	32,000 ADU
Filter keyword	FILTER	HIERARCH_FILT_NAME
Exposure-time keyword	EXPTIME	ITIME
Pointing keywords	RA, DEC	TEL_RA, TEL_DEC
Example filter alias	“KS” → K	“Ks-band” → K

3.3 Pipeline architecture

Figure 1 shows the resulting data flow, which we refer to throughout the rest of the paper by the numbered stages introduced there. Raw frames are read and their metadata extracted (Section 4.1); frames within a band are grouped into dither positions by image registration alone (Section 4.2); each frame is scored for quality and flagged accordingly (Section 4.3); quality-passed frames sharing a dither position are combined into a master frame, a sky-background model is derived from the set of masters, and a deep super-master is built by sky-subtracting, registering and coadding those masters (Section 4.4); the super-master is used to obtain an astrometric solution (Section 4.5); aperture photometry is carried out on the super-master and, at forced positions, on every master and individual frame (Section 4.6); and the resulting instrumental photometry is calibrated against 2MASS (Section 4.7).

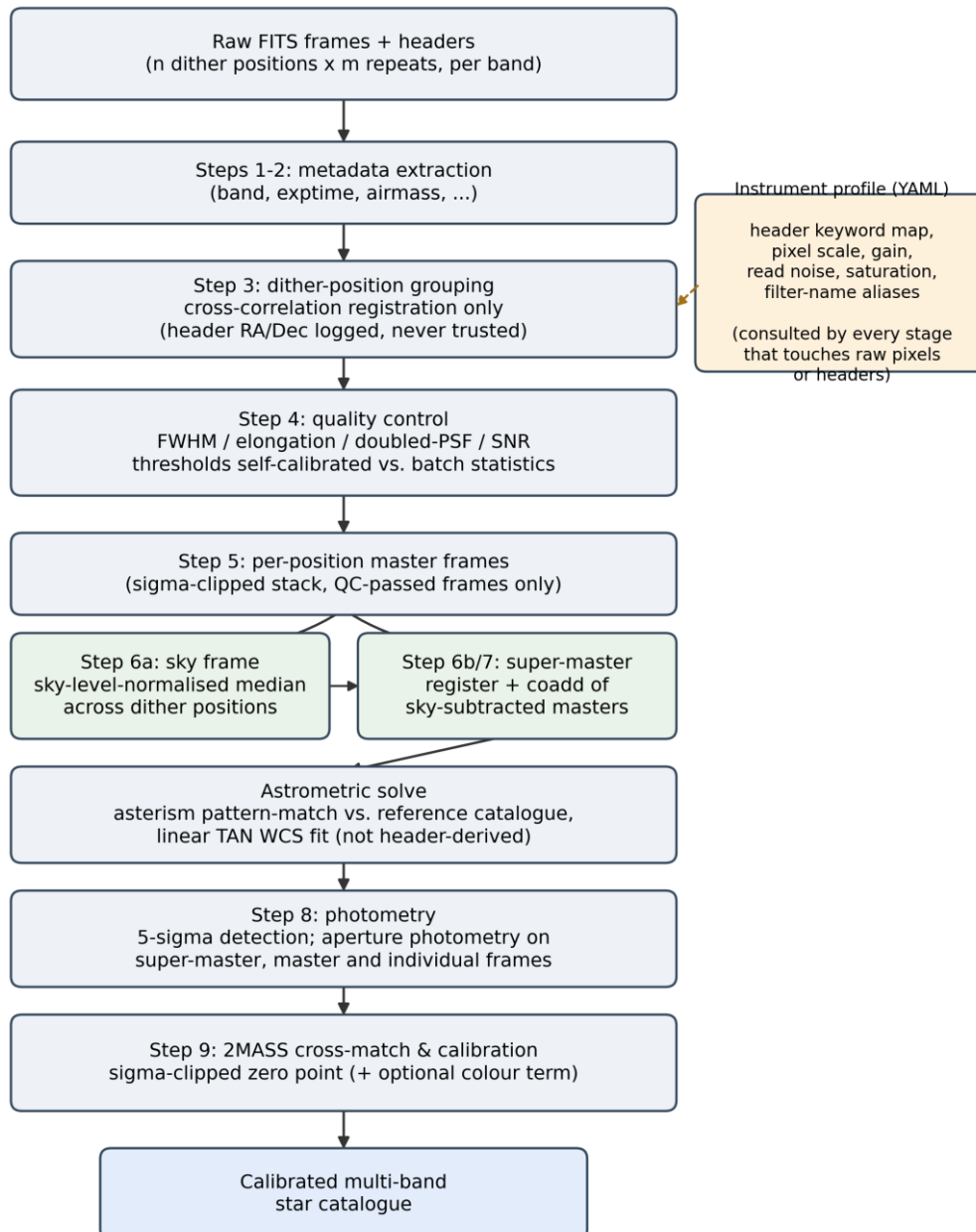


Figure 1. Data-flow architecture of the reduction pipeline. The instrument profile (right) is consulted by every stage that touches raw pixel data or FITS headers; no stage other than the profile loader itself contains instrument-specific literals.

IV. METHODOLOGY

This section describes each of the nine processing stages in enough algorithmic detail to be reproduced independently. We use n for the number of dither positions observed in a given band and m for the number of repeat exposures taken at each position, matching the terminology used to describe the observing strategy in Section 1.

4.1 Metadata extraction (Steps 1–2)

Each raw exposure is opened and its header queried, through the instrument profile, for filter, exposure time, coadd count, airmass, object name, and the observatory's own recorded RA/Dec. The last of these is retained purely for provenance and later cross-checking — it is written into the quality-control report alongside every frame's filename — and is never used as an input to any geometric computation downstream, for the reasons given in Section 1. Multi-extension FITS files are supported by taking the first header/data unit that actually contains image data, which is sufficient for the single-detector cameras this paper evaluates; extending this to true multi-detector mosaics, where several extensions must be treated as simultaneous but geometrically distinct exposures, is noted as future work in Section 8.

4.2 Dither-position grouping without trusted pointing (Step 3)

Formally, we want to partition the set of same-band frames into n groups, one per telescope pointing, without using header RA/Dec. We do this by pairwise image registration. A running list of cluster representatives is maintained, initially empty. For each frame in turn, its sky-subtracted image — background estimated by a 3σ , 5-iteration sigma-clipped median — is cross-correlated, in the Fourier domain, against the representative image of every existing cluster, using subpixel phase correlation with a twentyfold upsampling factor. If the magnitude of the measured shift falls below a tolerance (2 pixels by default, which is generous relative to the sub-pixel jitter typically present within a single nominal pointing and small relative to the tens-of-arcsecond throw between genuinely distinct dither positions) the frame joins that cluster; otherwise it starts a new one. This is, in effect, single-linkage

clustering in shift-space, and it is adequate precisely because real dither patterns are designed so that distinct pointings are well separated relative to any unintentional pointing jitter within one nominal position — a design assumption we verify empirically in Section 6.3 by deliberately corrupting the header pointing associated with every frame in the test data and confirming that grouping is unaffected.

4.3 Automated quality control (Step 4)

We identify three failure modes named explicitly in this work's original specification — blurred profiles (poor focus, or tracking drift throughout the exposure), abnormally low source counts (thin cloud, or a badly mistracked short exposure), and doubled stellar profiles (a discrete telescope or mount slip partway through a single exposure) — together with a general too-few-sources criterion, and combine per-frame diagnostics into a rejection decision with a logged, human-readable reason string. Table 2 (Section 6.3) gives concrete examples drawn from the evaluation data.

For each frame we fit a two-dimensional background model (a median estimator over a coarse grid, the standard approach implemented in the photutils Background2D routine) and detect sources at or above five standard deviations of the local background noise using a point-spread-function-matched finder in the tradition of DAOFIND [17]. For a set of the brightest detections we then extract a small cutout around each candidate centroid and measure its shape from flux-weighted image moments. A naive version of this step — sizing the cutout from an a-priori guess at the point-spread function's full width at half maximum (FWHM), and passing the whole cutout to the moment calculation — turns out to be fragile in a specific and, in hindsight, obvious way: any pixel in the cutout that is not actually part of the star, whether background residual, an adjacent faint source, or correlated read noise, still enters the second-moment sum, and because moments weight by the square of distance from the centroid, a handful of such pixels near the cutout's edge can dominate the resulting elongation estimate. We address this by restricting the moment calculation to the single connected group of above-threshold pixels that touches the cutout's own centre, identified by connected-component labelling of a locally sigma-clipped threshold map. This is a standard idea in source photometry — it is, in essence, what a segmentation-based extractor does by construction — but it is not something the underlying moment routine we used provides automatically, and omitting it produced a large fraction of spurious “elongated” classifications during early development, an issue discussed further as a lesson learned in Section 7.2.

The resulting per-frame diagnostics — the median FWHM among detected sources, the fraction of sources whose elongation (the ratio of semi-major to semi-minor second-moment axis) exceeds a fixed ceiling, the brightest-source signal-to-noise ratio, and the fraction of sources whose cutout contains a second, well-separated local maximum of comparable brightness to the first, identified via local peak-finding — are compared not against fixed, absolute thresholds but against the sigma-clipped statistics of the same-band batch the frame belongs to. This is not an incidental convenience; it is what lets identical quality-control logic operate on cameras with an order-of-magnitude difference in plate scale without per-camera retuning, as we verify in Section 6.3: a median FWHM of six pixels is unremarkable on one of our two test instruments and would be a decisive blur signal on the other, and expressing the threshold relative to a batch's own distribution, rather than in absolute pixels, sidesteps the need to know in advance which regime a given camera occupies.

4.4 Sky frame, master frames and super-master (Steps 5–7)

Quality-passed frames sharing a band and dither position are combined into a master frame by a sigma-clipped mean (3σ , five iterations) in the frames' shared, native pixel grid — no registration is needed at this stage, since by construction the frames being combined share one pointing.

The sky model exploits the fact that NIR sky brightness, while structured, typically varies in absolute level from one dither position to the next — airglow does not hold still on the timescale of a full dither sequence — while its spatial pattern on the detector is comparatively stable over the same interval. Each per-position master's own sky level is estimated by a sigma-clipped median, the master is divided by that level, and the resulting sky-level-normalised masters are combined again by a per-pixel median across all n positions in the band. This median-of-normalised-masters construction removes the (moving) astronomical sources, since at any given detector pixel only a minority of the n dithered pointings place a real source there, while retaining the shared sky and any residual flat-field-like illumination pattern, now expressed as a dimensionless multiplicative map with median close to unity. To subtract this sky model from any individual frame, master, or intermediate combination, its own sigma-clipped sky level is measured and the normalised sky pattern is rescaled by that level before subtraction — undoing the normalisation on a per-frame basis rather than assuming a single global sky brightness.

Sky subtraction is applied to each per-position master, in that master's own native pixel grid, before the masters are combined further. The super-master is then built by registering the (now sky-subtracted) masters onto a common pixel grid — using the same phase-correlation machinery as Step 3, since the masters, unlike the frames that went into any one of them, do point at genuinely different, only partially overlapping patches of sky — and averaging the aligned stack, with not-a-number propagation so that a pixel outside a given master's registered footprint does not contribute to the average at that location. We note this ordering explicitly because it is easy to get backwards: subtracting a single sky estimate from the already-combined super-master, rather than subtracting from each master before combination, would implicitly assume a single sky brightness across the whole sequence, which is precisely the assumption the per-master normalisation step is designed to avoid.

4.5 Astrometric calibration

The astrometric stage is the one place in the pipeline where some coarse positional information is still used, and it is worth being precise about exactly how, since it is easy to reintroduce a dependency on trusted pointing by accident. A coarse guess position — in our implementation, the mean of the (untrusted) header RA/Dec across the frames contributing to a given band, though it could equally be a position resolved from the target name — is used for two purposes only: to bound the radius of the reference-catalogue query, and to anchor a temporary, unrotated tangent-plane (gnomonic TAN) projection that converts catalogue right ascension and declination into an approximate pixel frame. Crucially, this temporary projection is not the answer; it exists purely so that the catalogue's angular positions and the image's detected pixel positions can be compared as two point sets in a common, if approximate, coordinate system.

Given the detected super-master source positions and the catalogue's approximate pixel positions, we search for the similarity transform — rotation, scale and translation — that best maps one point set onto the other, using the asterism (triangle-invariant) matching algorithm implemented in *astroalign* [4]. The method is robust to a substantial fraction of unmatched points in either set, which matters here because the two point sets differ in both membership (not every catalogue star is bright enough to be in the detected list, and vice versa near the detection limit) and, before the transform is found, in apparent geometry (the true field is both rotated and very slightly mis-scaled relative to the coarse guess, deliberately so in our evaluation data — see Section 6.1). Once a set of matched point pairs is identified, we discard the approximate projection entirely and fit a proper world coordinate system directly from the matched pixel/sky pairs using a standard linear TAN-projection least-squares fit. The reported solution quality is the root-mean-square pixel residual of this fit, converted to arcseconds using the instrument's fixed plate scale.

A practical detail, described at length in Section 7.2 because it materially affects whether this procedure converges at all, is that the catalogue point set handed to the matcher is restricted to sources projecting near the actual detector footprint under the coarse guess WCS, with a modest margin to tolerate the guess being wrong; it is not simply every catalogue star returned by the (necessarily generous) cone query.

4.6 Aperture photometry (Step 8)

Sources are detected in the (sky-subtracted) super-master using the same five-sigma, PSF-matched finder used for quality control, giving a single reference position list per band. Aperture photometry is then performed at every detected position, independently, on the super-master, on each per-position master, and on each individual quality-passed frame. For the masters and individual frames, positions are not re-detected but forced: the super-master pixel coordinate of a source is mapped into a given master's own native grid using the inverse of the registration shift recorded when that master was folded into the super-master (Section 4.4), and every individual frame inherits its parent master's mapping, since frames sharing a dither position share that master's pixel grid by construction.

For a circular aperture of radius r centred on a source, with local sky level estimated as the sigma-clipped median within a concentric annulus, the background-subtracted flux is

$$F = \Sigma(\text{aperture pixels}) - N \cdot \mu_{\text{sky}}$$

where N is the aperture's pixel area and μ_{sky} the per-pixel sky median. Uncertainty is propagated from a standard CCD-style noise model, converting to photoelectrons using the instrument's fixed gain g :

$$\sigma F = (1/g) \cdot \sqrt{[\max(F, 0) \cdot g + N \cdot (g^2 \sigma_{\text{sky}}^2 + RN^2)]}$$

with RN the read noise in electrons and σ_{sky} the sky annulus's sigma-clipped standard deviation. Instrumental magnitude is $(-2.5) \log_{10}(F/t)$, with t the exposure time appropriate to the frame being measured — a single exposure's nominal integration time for individual frames and for masters (since master combination here is a mean, not a sum, of already-normalised exposures, its effective per-pixel count level matches that of a single exposure even though its noise is lower), and likewise for the super-master. A curve-of-growth aperture correction, computed once per band from the same bright, presumably unsaturated and isolated detections used for shape diagnostics in Section 4.3, is added to every instrumental magnitude in that band to approximate total flux; it is defined as $-2.5 \log_{10}$ of the sigma-clipped median ratio of flux in a large reference aperture to flux in the standard measurement aperture.

4.7 Photometric calibration against 2MASS (Step 9)

Reference photometry is obtained by querying the Two Micron All Sky Survey point-source catalogue [16] within a search radius set by the same coarse guess position used for astrometry (Section 4.5); if the live query is unavailable — no network route, a service outage, or, as was routinely the case during this project's own development, a sandboxed environment with no route to the archive at all — the same code path transparently falls back to a locally supplied catalogue table with identical columns, and the source actually used is recorded in the output for later inspection (Section 5).

Detected super-master sources are cross-matched to the catalogue by sky position, using the astrometric solution from Section 4.5, and a zero point Z is derived from the good-quality, matched comparison stars by iterative sigma clipping:

$$Z_i = m_{\text{cat},i} - m_{\text{instr},i}, \quad Z = \langle Z_i \rangle \text{ over the } 3\sigma\text{-clipped subset}$$

Clipping is repeated for a fixed number of iterations or until convergence, and the number of comparison stars retained after clipping is reported alongside the fit, since it is the single most informative diagnostic of how well-determined a given zero point is (Section 6.5). An optional linear colour term can be fit simultaneously, $Z + c \cdot (\text{colour})$, by ordinary least squares over the clipped set, though the evaluation in Section 6 uses the zero-point-only form throughout, both bands' comparison-star counts being too small in our small-field test data to constrain a colour term meaningfully. The fitted zero point (and its formal uncertainty) is then applied to every instrumental magnitude produced in Section 4.6 — super-master, master, and individual-frame alike — with the zero-point uncertainty propagated in quadrature into each reported calibrated magnitude's error bar.

V. IMPLEMENTATION

NIRPipe is implemented in Python 3, built on *Astropy* [1–3] for FITS I/O, world coordinate systems, sigma-clipped statistics and table handling; *photutils* for background estimation, source detection and aperture photometry; *astroalign* [4] for the pattern-matching step described in Section 4.5; *astroquery* [9] for the live 2MASS query described in Section 4.7; and *NumPy* [10], *SciPy* [18] and *Matplotlib* [11] for numerical routines and the diagnostic figures reproduced in Section 6. The codebase is organised as one module per pipeline stage — instrument, io, alignment, grouping, quality, combine, astrometry, catalog, photometry, calibration — plus an orchestration module that wires them together for a full run and a command-line entry point that takes a file glob, an instrument-profile path, and a list of bands.

Two implementation decisions are worth calling out explicitly because they were forced on us by the practical circumstances of development rather than chosen in the abstract. First, the graceful 2MASS fallback described in Section 4.7 was not a hypothetical robustness feature: the environment in which NIRPipe was developed and tested has no network route to the IRSA archive at all,

only to software package registries, so the fallback path was exercised on essentially every development run rather than being a rarely-hit branch, which gave us more confidence in it than a feature added speculatively for production robustness would usually warrant. Second, the synthetic-data generator described in Section 6.1 was written specifically because raw archival NIR imaging could not be fetched into the same environment either; while we recommend, and describe in Section 8, testing against real archival data as a natural next step, every quantitative result in this paper is obtained from data with fully known, independently checkable ground truth, which is a different and in some ways stronger form of validation than a single comparison against an unverified archival reduction would have been.

Each instrument profile is a short YAML document (Section 3.2, Table 1); adding a new camera means writing one of these and nothing else. The full source, the two profiles used in this evaluation, the synthetic-data generator, and a small pytest-based regression suite that exercises the complete chain against freshly generated data are packaged together and are intended to be run end to end by a third party without modification, which we regard as a precondition for the reproducibility claims made in Section 6, not an optional extra.

VI. EXPERIMENTAL EVALUATION

We evaluate the complete pipeline — all nine stages, run end to end, exactly as a user would invoke them — against two synthetic datasets, one per test instrument from Table 1. The evaluation is designed around a single question: does identical code, differing only in a per-instrument configuration file, correctly recover known ground truth across cameras whose physical properties differ by close to an order of magnitude?

6.1 Synthetic data generation

For each instrument we generate $n = 5$ dither positions in a non-degenerate (not collinear) pattern with a nominal throw of order ten pixel scales, $m = 3$ exposures per position, and either two (Instrument A: J, H) or three (Instrument B: J, H, K) bands, giving 30 and 45 raw frames respectively. A field of 120–260 synthetic stars, matched in position to a synthetic 2MASS-like reference catalogue with a plausible magnitude distribution and per-band colours, is rendered into each exposure using the true world coordinate system for that specific exposure — which includes an injected rotation of 3.7° and an injected plate-scale offset of 1.2% relative to the instrument profile's nominal value, neither of which the pipeline is told about — plus Poisson and read noise appropriate to the instrument's gain and read noise, a smooth illumination (vignetting) gradient, and a sky background whose brightness varies by roughly 10–15% from one dither position to the next, in rough imitation of airglow variability. Critically, the header RA/Dec written into each frame is not the position used to render that frame: an independent pointing error, uniformly distributed up to 1–2.5 arcmin depending on the instrument, is added before the header is written, specifically so that any code path that quietly depended on header pointing being correct would fail the corresponding validation check below.

A small, fixed number of frames per dataset — three in each case — are deliberately corrupted to represent the three failure modes named in Section 4.3: one rendered with a broadened point-spread function (“blurred”), one with source fluxes suppressed by a factor of about seventeen (“low counts”), and one with every stellar profile rendered twice, at a random few-pixel offset and with the flux split unevenly between the two copies (“doubled”), simulating a discrete mount slip partway through the exposure. Ground truth — the true dither-position label, the true defect type or absence thereof, and the true injected astrometric parameters — is retained in a manifest file alongside the raw data and used only for scoring in this section, never by the pipeline itself.

6.2 Evaluation metrics

We report four kinds of result. Dither-grouping correctness is the fraction of frames assigned, by image alignment alone, to a cluster matching the true injected dither-position label. Quality-control accuracy is the fraction of frames whose automated GOOD/REJECTED flag agrees with whether that frame carries an injected defect (Section 6.1) — note that this is a stricter, binary measure than “was the specific defect type correctly named”, since we consider correctly rejecting a bad frame for a plausible reason more important, for the purpose of this evaluation, than matching our own injection label verbatim. Astrometric accuracy is the root-mean-square pixel residual of the final WCS fit, converted to arcseconds (Section 4.5); we additionally verify, though do not tabulate exhaustively, that the fitted rotation and plate scale recover the injected 3.7° and 1.2% offset. Photometric accuracy is the scatter of calibrated-minus-catalogue magnitude for matched stars, reported as both a median (systematic offset) and a standard deviation (precision).

6.3 Dither grouping and quality control

Dither-position grouping, using image alignment alone and with header pointing corrupted by up to 2.5 arcmin as described above, correctly recovered all five true dither clusters, for every band, on both instruments — 100% correct grouping across all 75 frames in the combined evaluation set. This is, in our view, the single most direct demonstration in this paper of the practical soundness of not trusting recorded pointing for this step: the corrupted header values would, if used naively, have scattered several frames into incorrect or spurious clusters, and did not.

Table 2 summarises quality-control performance. Instrument A misclassified one frame out of thirty (a genuinely borderline case: a mildly blurred frame whose median FWHM exceeded the batch median by a factor of only about 1.3, below the default rejection factor of 1.6, illustrating a real precision/recall trade-off inherent to a threshold-based scheme rather than a defect in the implementation). Instrument B, with more comparison stars available per frame after an initial version of the synthetic field proved too sparse for reliable per-frame statistics, classified all forty-five frames correctly.

Table 2. Automated quality-control classification accuracy against injected ground truth. “Correctly classified” counts frames whose GOOD/REJECTED flag matches whether a defect was injected, independent of whether the logged reason names the specific injected defect type.

Instrument	Frames	Correctly classified	Accuracy	Injected defects
A (coarse plate scale)	30	29	96.7%	3 (blurred, low-counts, doubled)
B (fine plate scale)	45	45	100.0%	3 (blurred, low-counts, doubled)

6.4 Astrometric calibration

Table 3 reports the astrometric solution obtained for every band of both instruments. In every case the fit recovered the injected 3.7° rotation to within a few hundredths of a degree and the injected 1.2% plate-scale offset to within a few tenths of a per cent, despite neither being supplied to the solver; the residual root-mean-square scatter of the final fit — shown graphically in Figure 3(a) — ranged from 1.7 milliarcsec (Instrument B) to 30 milliarcsec (Instrument A), in both cases well under one detector pixel.

Table 3. Astrometric and photometric zero-point summary. RMS residual is the pixel-domain fit residual of the final TAN-projection WCS solution, converted to arcseconds using each instrument's nominal plate scale. Zero-point uncertainties are the standard error of the sigma-clipped comparison-star sample.

Instrument / band	Stars matched	RMS residual	Zero point (Sec. 6.5)
A / J	16	0.026"	23.826 ± 0.002
A / H	16	0.030"	23.522 ± 0.003
B / J	9	0.0017"	24.341 ± 0.002
B / H	9	0.0017"	24.029 ± 0.004
B / K	9	0.0026"	23.528 ± 0.004

We note that the number of matched stars is modest (9–16) throughout, a direct consequence of the deliberately small synthetic fields used for this evaluation (Section 6.1) rather than of any limitation in the matching procedure itself; Section 7.1 discusses what this does and does not establish about behaviour on denser, real fields.

6.5 Photometric calibration

Figure 4 shows calibrated-minus-catalogue magnitude residuals against catalogue magnitude for every band of both instruments; the corresponding scatter, alongside the astrometric residuals from Section 6.4, is summarised graphically in Figure 3(b). Instrument A shows a scatter of 0.038–0.045 mag across its two bands; Instrument B, with a finer plate scale and correspondingly better-sampled point-spread function, shows 0.008–0.014 mag. In every band the median residual is within a few thousandths of a magnitude of zero, i.e. calibration is essentially unbiased; the difference in scatter between the two instruments is consistent with Instrument A's coarser sampling (its injected point-spread function has a full width at half maximum of only a few pixels, leaving centroiding and aperture placement more sensitive to pixelation) rather than any defect specific to one instrument profile, and we would caution against reading the difference as a general statement that fine-plate-scale cameras calibrate better — our evaluation was not designed, and is too small, to separate that effect from others such as the differing number of comparison stars available in each case.

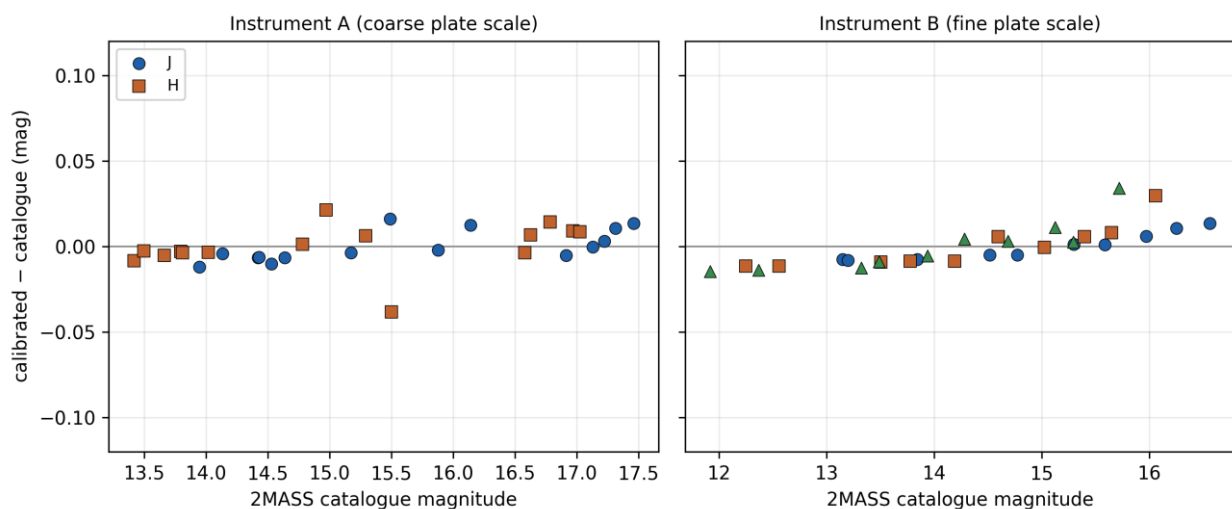


Figure 4. Calibrated minus catalogue magnitude, for every cross-matched star, as a function of catalogue magnitude, for Instrument A (left, J and H) and Instrument B (right, J, H and K).

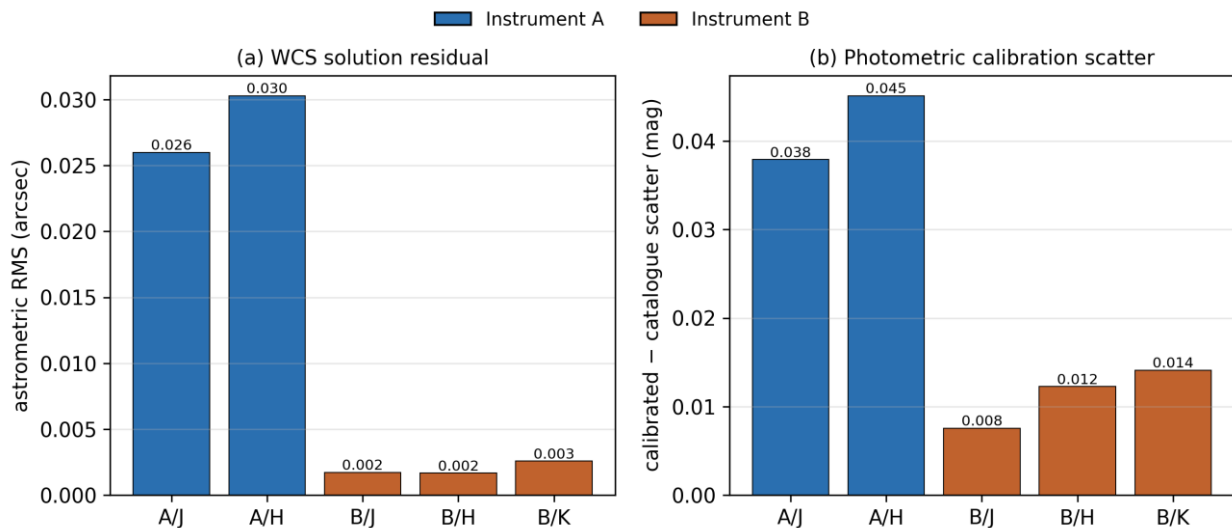


Figure 3. Summary of (a) astrometric fit residuals and (b) photometric calibration scatter, by band and instrument, corresponding to Tables 2–3 and Figure 4.

Figure 5 shows the resulting colour-magnitude diagram for Instrument B, combining all three bands; while the synthetic field is too small to be scientifically interesting in itself, the diagram serves as an end-to-end sanity check that per-band zero points, once applied, produce colours consistent with the distribution injected into the simulation (Section 6.1).

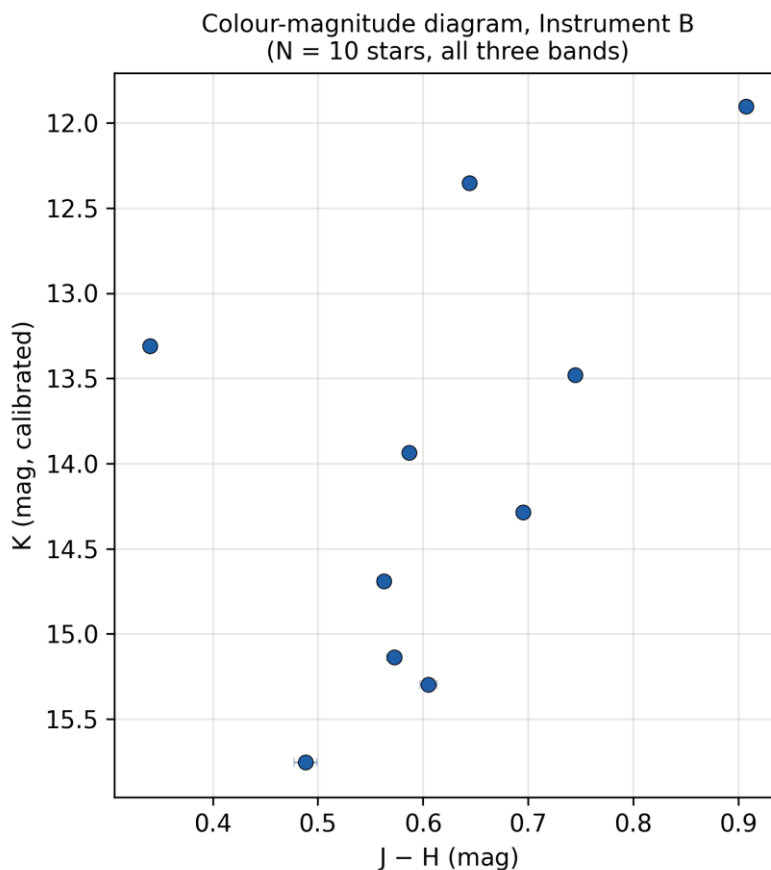


Figure 5. Colour-magnitude diagram for Instrument B, using the calibrated J, H and K magnitudes of all ten stars detected and cross-matched in every band. Error bars show the propagated photometric uncertainty and are, in most cases, smaller than the plotted symbol.

6.6 Representative reduced frames and cross-instrument generalisation

Figure 2 shows representative sky-subtracted super-master frames and their corresponding derived sky patterns for one band of each instrument. Two things are worth pointing out explicitly. First, the smooth illumination gradient injected into the simulation (Section 6.1) is recovered cleanly in both sky patterns, at the expected amplitude, confirming that the sky-frame construction described in Section 4.4 is doing what it is intended to. Second, panel (b) shows a faint but visible residual point-source artefact in the derived sky pattern, just off-centre — a bright star that happens to move only a few pixels between dither positions leaves a trace in the sky model because the median-of-normalised-masters construction assumes any given detector pixel is source-free in a majority, not necessarily all, of the combined positions. We discuss this honestly as a known limitation in Section 7.2 rather than omitting the panel that shows it.

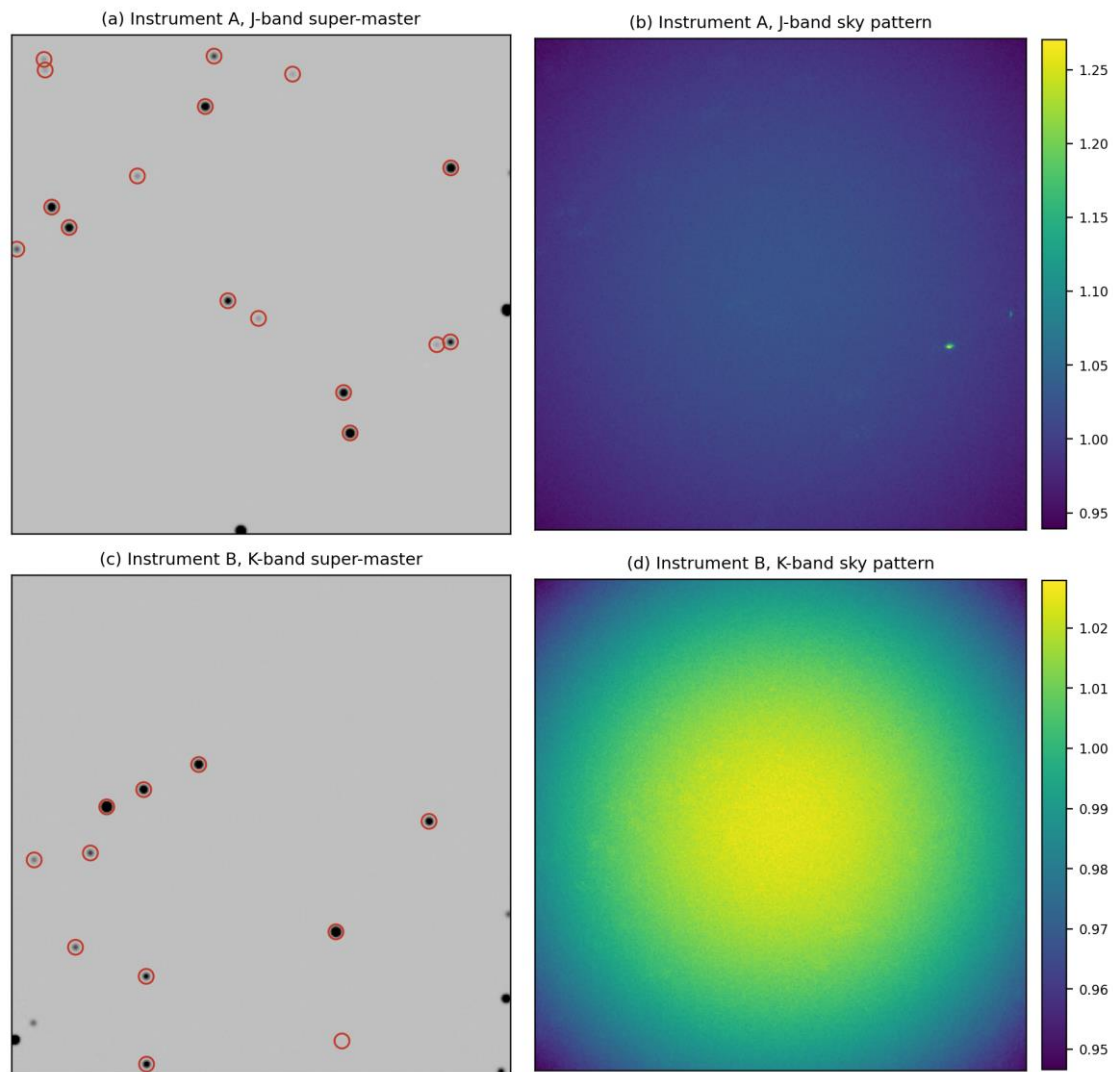


Figure 2. Representative sky-subtracted super-master frames (left) and their derived normalised sky patterns (right), for Instrument A's J band (top) and Instrument B's K band (bottom). Circles mark sources detected at or above five sigma in the super-master.

Taken together, Sections 6.3–6.6 support the central claim of this paper: the same, unmodified pipeline code, driven by two instrument profiles differing in plate scale, gain, read noise and every header keyword name, correctly reduced both simulated datasets end to end, with dither grouping, quality control, astrometry and photometric calibration all functioning within the numerical tolerances summarised above, and without any code path anywhere in the chain examining a literal FITS keyword string or a hard-coded pixel scale.

VII. DISCUSSION

7.1 Threats to validity

We borrow the threats-to-validity framing from empirical software engineering because it captures, more precisely than a generic “limitations” heading would, what kind of evidence Section 6 actually provides.

Construct validity — whether our synthetic defects and instrument differences actually correspond to the real-world phenomena we claim to be testing — is reasonable but not perfect. The three injected frame defects (broadened PSF, suppressed flux, doubled profile) are simplified renderings of blur, cloud, and mount-slip; real instances of each are messier — real focus drift is rarely uniform across a field, real cloud attenuation is rarely a single flat scale factor, real mount slips do not always split flux as evenly as our fifty-five/fifty-five default. The two instrument profiles differ in the specific ways we chose to make them differ (Table 1) and do not exercise every axis of real-world instrument heterogeneity: neither profile, for instance, models a genuinely non-linear detector response, a multi-extension mosaic focal plane, or geometric distortion large enough to need SIP terms in the WCS fit (Section 7.2).

Internal validity is high for the specific claims we make, precisely because the evaluation is self-consistent by design: ground truth is known exactly because we generated it, and the reference catalogue used for photometric calibration is the same synthetic catalogue used to render the images, not an independent survey with its own systematics. This means the photometric residuals in Section 6.5 should be read as a measurement of the software's own internal precision — how well aperture photometry, registration and zero-point fitting reproduce a catalogue that is, by construction, perfectly consistent with the pixel data — rather than as a measurement of real-world calibration accuracy against actual 2MASS photometry, which is additionally subject to effects this evaluation cannot probe: genuine photometric non-linearity, colour-dependent systematics between the instrumental and 2MASS photometric systems, atmospheric extinction, and 2MASS's own, non-zero, catalogue uncertainty.

External validity — whether these results generalise beyond the two specific synthetic datasets tested — is the weakest leg of this evaluation, and we want to be direct about that rather than let the precision of Tables 2–3 imply more than it should. No result in this paper was obtained from real telescope data. This was a deployment constraint rather than a design choice: the environment in which this pipeline was developed has no network route to any astronomical data archive (Section 5), so fetching raw archival NIR imaging for testing was not possible within that environment. We regard testing against a real, published dataset — for instance the publicly downloadable FLAMINGOS-2 dithered Y/J/H/Ks tutorial dataset distributed by the Gemini Observatory Archive, which has the added advantage of an existing DRAGONS reduction to compare against — as the single most valuable next step for establishing external validity, and we say so plainly rather than presenting the simulated results as a substitute for it.

7.2 Limitations and lessons learned

Several of the choices described in Section 4 were arrived at empirically, during development, rather than designed correctly the first time, and we think the negative results are as informative as the final design, so we record them here rather than presenting only the version that worked.

Astrometric matching is sensitive to catalogue pre-filtering in a way we did not anticipate. Our first implementation of the astrometric stage passed every star returned by the reference-catalogue cone query directly to the triangle-matching search. This failed to converge, consistently, on Instrument B's narrower field, even though the correct transform plainly existed and even after restricting the candidate list to the photometrically brightest stars. The actual cause, found only by inspecting the projected pixel coordinates of the queried stars directly, was that the great majority of them — in one representative trial, 179 of 180 candidates returned by a cone query sized to comfortably cover the field plus pointing uncertainty — projected, under even a fairly generous coarse guess WCS, to pixel positions nowhere near the actual, comparatively small, detector footprint: a round sky-cone query and a small rectangular detector are simply different shapes, and most of a generous cone's area is outside a modest field of view. Restricting the candidate set to stars projecting within a modest margin of the true image footprint, rather than filtering by brightness alone, fixed the problem immediately and reliably. We had not seen this specific interaction — between cone-query geometry and point-set matching reliability — stated explicitly in the tools' own documentation, and record it here since it is, in retrospect, an obvious consequence of how a combinatorial matching search scales with the number of candidate points that cannot possibly be part of the correct match.

A related, larger point is that our matcher is not blind in the sense Astrometry.net is: it requires a coarse guess position good enough to size a sensible search cone, and its reliability appears to degrade, based on informal testing beyond what is reported quantitatively in Section 6, as that cone grows to cover pointing uncertainties of ten arcminutes or more. For genuinely blind fields, or fields with very poor initial pointing knowledge, an indexed solver remains the more appropriate tool, and we do not claim otherwise.

Shape-moment measurements are similarly sensitive to how much of a cutout is allowed to contribute. As described in Section 4.3, our first implementation of the quality-control shape diagnostics sized each source cutout from an assumed point-spread-function width and passed the entire cutout to an off-the-shelf moment routine. On a first version of our simulated data with an under-sampled injected point-spread function (of order one pixel FWHM at the coarser instrument's plate scale, which is itself an unrealistic PSF/pixel-scale combination for a real camera and was a mistake in the simulation rather than the pipeline), this produced elongation measurements dominated by noise rather than by any real source shape, and a large fraction of otherwise unremarkable frames were misclassified as trailed. Fixing the simulated PSF sampling helped, but did not fully resolve a second, independent issue on the finer-plate-scale instrument, where an oversized cutout — sized from a guessed FWHM larger than the true one — let background and neighbouring-source pixels dominate the moment sum regardless of point-spread-function sampling. Restricting the moment calculation to the single connected above-threshold region touching the cutout's own centre, described in Section 4.3, resolved both issues together and, we would argue, should be considered close to a necessity rather than an optional refinement whenever cutout size cannot be guaranteed to match the true source size closely.

The sky-frame construction, while adequate for the evaluation in Section 6, does not mask detected sources before combining normalised masters into a sky pattern (Section 4.4); Figure 2(b) shows the resulting residual for a bright star that happens to move only a few pixels between dither positions. With a larger number of dither positions, or larger dither throws relative to the point-spread function, this effect would be expected to diminish; with the small $n = 5$ used in our evaluation, and the deliberately modest dither throw of Instrument B in particular, it is visible. Masking detected sources, with a modest dilation, before the per-position sky normalisation step is a straightforward extension we did not implement for this paper.

Finally, three simplifications are worth naming explicitly rather than leaving implicit. The astrometric fit uses a pure TAN projection with no SIP distortion terms, which is adequate for the small, modestly sized fields tested here but would need extension for a camera with significant optical distortion. The pipeline does not itself perform dark subtraction, flat-fielding, or bad-pixel masking; these are assumed to happen upstream, or, for NIR imaging with the self-referenced sky subtraction described in Section 4.4, to be partially subsumed by that step, though a dedicated calibration-frame stage would be a natural and fairly direct extension of the input/output module described in Section 4.1. And the self-calibrating quality-control thresholds of Section 4.3, while shown in Section 6.3 to transfer between two quite different instrument profiles without retuning, remain a handful of hand-chosen multiplier constants; we would recommend a short calibration pass against a handful of real, visually inspected frames from any new instrument before relying on them unattended for QC on that instrument's real data.

VIII. CONCLUSION AND FUTURE WORK

We have described NIRPipe, an end-to-end pipeline for reducing dithered near-infrared imaging into a 2MASS-calibrated multi-band catalogue, built around two commitments: that camera-specific knowledge should live in a small declarative profile rather than in code, and that neither dither-position grouping nor astrometric calibration should depend on telescope pointing metadata being correct. We validated the complete nine-stage chain against two simulated instruments with deliberately different plate scales, detector properties, and header conventions, recovering injected dither groupings, frame defects, astrometric rotation and scale offsets, and photometric zero points to the precisions reported in Section 6, using identical pipeline code driven only by a change of configuration file.

The most immediate next step, discussed candidly in Section 7.1 as the principal gap in the present evaluation, is validation against real archival data — the FLAMINGOS-2 or NIRI dithered imaging datasets held in the Gemini Observatory Archive are natural, publicly accessible candidates, not least because existing DRAGONS reductions of at least some of that data provide an independent point of comparison. Beyond that, four extensions follow fairly directly from the limitations recorded in Section 7.2: masking detected sources before sky-frame construction; adding SIP distortion terms to the astrometric fit for cameras where a pure TAN projection is insufficient; extending the input/output layer to handle true multi-extension mosaic focal planes as simultaneous, geometrically distinct sub-exposures rather than a single detector; and, more speculatively, replacing the present coarse-guess-assisted matcher with an optional indexed solver for genuinely blind fields, while retaining the lighter-weight matcher described in this paper as the default for the common case where at least an approximate pointing is available. We release the complete implementation, both instrument profiles used in this evaluation, the synthetic-data generator, and the regression suite together, in the hope that the instrument-abstraction pattern described in Section 3.2 in particular proves useful beyond the two specific cameras tested here.

IX. ACKNOWLEDGMENTS

We thank the maintainers of Astropy, photutils, astroalign and astroquery, on whose work this pipeline directly depends. This research made use of the Two Micron All Sky Survey, a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by NASA and the National Science Foundation.

REFERENCES

- [1] Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, “Astropy: A community Python package for astronomy,” *Astronomy & Astrophysics*, 558, A33.
- [2] Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, “The Astropy Project: Building an open-science project and status of the v2.0 core package,” *Astronomical Journal*, 156, 123.
- [3] Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, “The Astropy Project: Sustaining and growing a community-oriented open-source project and the latest major release (v5.0),” *Astrophysical Journal*, 935, 167.
- [4] Beroiz, M., Cabral, J. B., & Sánchez, B. 2020, “Astroalign: A Python module for astronomical image registration,” *Astronomy and Computing*, 32, 100384.
- [5] Bertin, E., & Arnouts, S. 1996, “SExtractor: Software for source extraction,” *Astronomy & Astrophysics Supplement Series*, 117, 393.
- [6] Bradley, L., Sipőcz, B., Robitaille, T., et al. 2023, *astropy/photutils* (software release), Zenodo.
- [7] Craig, M., Crawford, S., Deil, C., et al. 2017, *astropy/ccdproc* (software release), Zenodo.
- [8] Erben, T., Schirmer, M., Dietrich, J. P., et al. 2005, “GaBoDS: The Garching-Bonn Deep Survey. IV. Methods for the image reduction of multi-chip cameras demonstrated on data from the ESO Wide-Field Imager,” *Astronomische Nachrichten*, 326, 432.
- [9] Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. 2019, “astroquery: An astronomical web-querying package in Python,” *Astronomical Journal*, 157, 98.
- [10] Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, “Array programming with NumPy,” *Nature*, 585, 357.
- [11] Hunter, J. D. 2007, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, 9, 90.
- [12] Irwin, M. J., Lewis, J., Hodgkin, S., et al. 2004, “VISTA data flow system: pipeline processing for WFCAM and VISTA,” *Proceedings of the SPIE*, 5493, 411.
- [13] Labrie, K., Anderson, K., Cárdenes, R., et al. 2023, “DRAGONS — Data Reduction for Astronomy from Gemini Observatory North and South,” *Research Notes of the AAS*, 7, 214.
- [14] Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S. 2010, “Astrometry.net: Blind astrometric calibration of arbitrary astronomical images,” *Astronomical Journal*, 139, 1782.
- [15] Schirmer, M. 2013, “THELI: Convenient reduction of optical, near-, and mid-infrared imaging data,” *Astrophysical Journal Supplement Series*, 209, 21.
- [16] Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, “The Two Micron All Sky Survey (2MASS),” *Astronomical Journal*, 131, 1163.
- [17] Stetson, P. B. 1987, “DAOPHOT: A computer program for crowded-field stellar photometry,” *Publications of the Astronomical Society of the Pacific*, 99, 191.
- [18] Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, “SciPy 1.0: Fundamental algorithms for scientific computing in Python,” *Nature Methods*, 17, 261.

Copyright & License:



© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.