COMPOSE: Hypergraph Cover Optimization for Multi-view 3D Human Pose Estimation

Tony Danjun Wang[1,2] **Tolga Birdal**[3] **Nassir Navab**[1,2] **Lennart Bastian**[1,2,3]
[1]Technical University of Munich [2]Munich Center for Machine Learning
[3]Imperial College London
tony.wang@tum.de

Abstract

3D human pose estimation from sparse multi-view camera rigs is an essential task for numerous applications, including action recognition, sports analysis, and human-robot interaction. While learned methods dominate the field on benchmarks, they require large annotated datasets. Training-free optimization-based methods remain promising as they circumvent 3D supervision by solving a correspondence problem across views from 2D detections. Existing combinatorial formulations rely on pairwise associations to model this correspondence problem and enforce global consistency across views only as a downstream constraint. However, reconciling locally plausible pairwise matches becomes brittle under occlusion and noisy detections, where local errors propagate globally. We propose COMPOSE, which recasts multi-view 3D human pose estimation as a weighted exact-cover optimization over a hypergraph of person hypotheses. Our formulation replaces pairwise association and post-hoc consistency enforcement with a single global combinatorial objective. To address the exponentially large candidate space, we introduce a geometric pruning strategy alongside two complementary solvers: an exact Integer Linear Programming formulation and a scalable relaxation via Belief Propagation. Without any 3D supervision, COMPOSE improves AP_{25} on Panoptic by 22 points over the best optimization-based method and 11 points over the best self-supervised learned method, demonstrating the effectiveness of higher-order combinatorial association for training-free multi-view 3D human pose estimation. Project page and code are available at wngtn.github.io/COMPOSE.

1 Introduction

Human pose estimation is a fundamental task in computer vision, yet its deployment in safety-critical scenarios remains challenging. While significant progress has been made in monocular 2D settings due to the availability of large-scale annotated datasets [37, 40, 65], these methods inherently lack metric depth information and can be prone to occlusions [62]. In real-world applications such as collaborative human-robot interaction [17] and operating room monitoring [53], precise spatial localization thus requires lifting 2D detections across multiple camera views into a joint 3D coordinate system [13, 26]. By capturing activities from diverse viewpoints, these setups enable triangulation, thereby resolving depth ambiguities and noise inherent to single-view imaging.

Even with advances in 2D backbones, the community has largely shifted toward end-to-end learning-based methods for multi-view 3D human pose estimation [49, 58, 35]. These approaches have significantly advanced the field, but require large annotated datasets. However, obtaining ground-truth 3D annotations is labor-intensive and technically challenging, as skeletal joints are internal to the human body and cannot be directly observed or accurately annotated from surface images alone. Furthermore, learning-based models frequently suffer from domain gaps, failing to generalize to

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

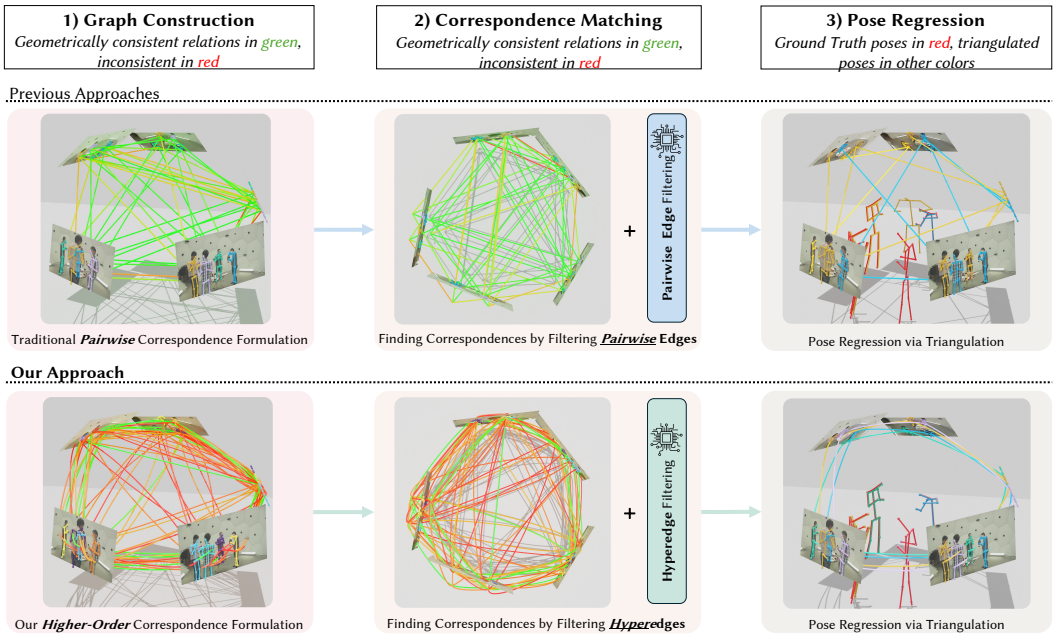


Figure 1: **Top:** Traditional approaches rely on pairwise geometric constraints [13, 56]. As illustrated, these methods generate pairwise associations that, while *locally* consistent between two views, often fail to form a *globally* coherent structure. Consequently, algorithms face the difficult task of reconciling these locally plausible but globally conflicting edges via *synchronization* to recover the correct 3D poses. **Bottom:** We propose COMPOSE, a hypergraph formulation that jointly models higher-order relationships across views. We re-frame correspondence matching as a *hypergraph partitioning* problem, where hyperedges encode multi-view consistency. This global formulation effectively resolves ambiguities by enforcing consensus across the entire set of views.

unseen environments or novel camera configurations [35, 11, 47]. Domain-agnostic methods that do not rely on 3D supervision are thus highly sought after.

Training-free optimization techniques offer a complementary paradigm. Rather than amortizing inference cost over a training distribution by learning model parameters, these methods explicitly solve the multi-view assignment problem at test time for each instance, without extensive supervision or annotations. As such, they do not require domain-specific 3D pose detectors and can encode geometric and combinatorial constraints, such as reprojection consistency and multi-person exclusivity, directly in the objective. Moreover, they preserve a modular separation between 2D perception and 3D reconstruction: improved off-the-shelf keypoint detectors can be incorporated without retraining the geometric solver. These properties are particularly appealing for deployments with novel camera layouts, limited access to 3D ground truth, or environments that exhibit a significant distribution shift from the training data of learned 3D pose models.

We present COMPOSE, a hypergraph formulation for 3D human pose estimation that models multi-view relations between *2D observations* as a weighted exact-cover over a hypergraph (see Fig. 1). Existing methods model multi-view relationships by *synchronizing* pairwise relations to recover a cycle-consistent matching between views [13, 7, 10] (see Section A for a comprehensive overview). However, these methods are notoriously reliant on pairwise matches computed in isolation (Fig. 1, top), **making it challenging to resolve ambiguities when views are occluded or noisy** [20].

To alleviate these challenges, COMPOSE extends modeling beyond pairwise relationships by abstracting 2D image correspondence in a hypergraph formulation, where each hyperedge represents a candidate multi-view person hypothesis. By scoring sets of 2D detections simultaneously, COMPOSE enforces a holistic consensus across all cameras, improving robustness to outliers in individual views (see Fig. 1, where hyperedges better express global geometric consistency). To solve this higher-order partitioning problem, we propose two complementary optimization strategies: an exact Integer Linear Programming (ILP) formulation and a scalable probabilistic relaxation using loopy Belief Propagation (BP). The ILP formulation recovers the globally optimal assignment, while the BP relaxation produces

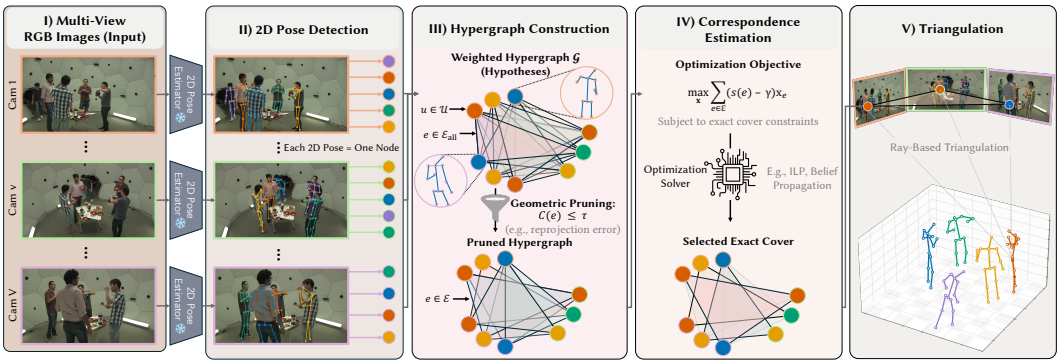


Figure 2: **COMPOSE**. Multi-view images are taken as input (I). We employ an off-the-shelf 2D pose estimator to extract 2D keypoints (II). Next, we construct the weighted hypergraph (III). We then solve the exact cover problem to optimally partition the graph and establish unique correspondences (IV). Finally, we triangulate the 3D human poses from the obtained correspondences (V).

continuous marginal beliefs over hyperedges. This enables fully parallelizable, GPU-accelerated inference and yields soft confidence scores over the associations, providing an uncertainty-aware alternative to hard assignments. With these optimization-based approaches, the proposed method successfully reconstructs accurate 3D poses from multi-view 2D observations, even in the presence of severe occlusions. Furthermore, experiments demonstrate that our hypergraph-based approach outperforms state-of-the-art optimization-based and recent self-supervised learning methods.

Our main contributions can be summarized as follows:

- We cast multi-view 3D human pose estimation as a weighted exact cover over a hypergraph of person hypotheses, rethinking the problem to natively incorporate multi-view consistency.
- We introduce a combinatorial optimization objective with a geometric pruning strategy, enabling recovery of the globally optimal cover via Integer Linear Programming.
- We derive a probabilistic relaxation of the exact-cover objective and solve it with loopy Belief Propagation, yielding parallelizable GPU-accelerated inference and continuous marginal beliefs.
- We show COMPOSE improves over state-of-the-art optimization-based baselines under identical 2D detections and surpasses recent self-supervised methods without any 3D supervision.

2 Related Work

Optimization-based methods. 3D Pictorial Structures [6] combines multi-view observations and skeletal constraints in a conditional random field over 3D body-part hypotheses. MvPose [13] jointly optimizes pairwise geometric and appearance affinities across views, using a low-rank relaxation to promote cycle consistency. QuickPose [64] associates partial skeletons, and 4DA [61] jointly models spatial and temporal associations. Permutation synchronization methods recover globally consistent assignments from estimated pairwise correspondences [4, 7]. COMPOSE instead scores candidate multi-view person hypotheses directly through joint reprojection consistency and selects mutually exclusive groups through exact-cover constraints.

Learning-based methods. VoxelPose [49] aggregates multi-view features in a 3D voxel grid and uses 3D convolutions to localize people and regress their poses. Faster VoxelPose [58] reduces this computation by projecting the feature volume onto three coordinate planes. MvP [52] and MVGFormer [35] use transformers for multi-view feature fusion. These methods learn their 3D pose models with 3D supervision. SelfPose3d [47], DSP [38], and DisPOSE [55] learn 3D pose estimators without 3D supervision. COMPOSE instead uses off-the-shelf 2D detections and solves association geometrically without training a 3D pose model. Extended related work is provided in Section A.

3 Methodology

COMPOSE addresses multi-view multi-person 3D pose estimation by decomposing the task into 2D detection, higher-order association, and triangulation, as illustrated in Fig. 2. 2D poses are

first independently detected in each camera view using an off-the-shelf pose estimator. The central modeling choice in COMPOSE is to treat an entire candidate 3D person hypothesis as the atomic association, rather than pairwise matching. Pairwise edges only express *local compatibility* between detections. Hyperedges, on the other hand, group detections across multiple views, representing a consistent hypothesis between all 2D views of the same individual. This turns correspondence recovery into a *weighted exact-cover problem* over a hypergraph: the solver selects a set of mutually exclusive hypotheses that explains the observed 2D detections, with scores derived geometrically.

3.1 Problem Setting and Hypergraph Construction

We consider V calibrated RGB cameras with projection functions $\{\pi_v\}_{v=1}^{V}$. For each view v, a 2D pose detector produces pose detections $\mathcal{U}_v = \{U_i^v\}_{i=1}^{n_v}$. Each detection U_i^v contains J image-space joints, with joint j denoted by $\mathbf{u}_{i,j}^v \in \mathbb{R}^2$. The set of all 2D pose detections is $\mathcal{U} = \bigcup_{v=1}^{V} \mathcal{U}_v$.

We model multi-view correspondences as a hypergraph $\mathcal{G} = (\mathcal{U}, \mathcal{E})$, where the vertex set $\mathcal{U}$ consists of all 2D pose detections and each hyperedge $e \in \mathcal{E}$ is a candidate correspondence group. A valid hyperedge contains at most one detection per camera view: a valid correspondence should have **at most one detection per image**:

$$|e \cap \mathcal{U}_v| \leq 1, \qquad \forall v,$$

We denote $\mathcal{E}_{\text{all}}$ as the set of all non-empty hyperedges satisfying this constraint. A non-singleton hyperedge represents one candidate 3D person hypothesis: all its 2D detections are hypothesized projections of the same physical person. For instance, $e = \{U_2^1, U_5^3, U_1^4\}$ groups one detection from each of views 1, 3, and 4 into a single multi-view correspondence hypothesis. Singleton hyperedges are retained to explain unmatched, false-positive, or single-view detections. In practice, COMPOSE operates on a pruned hypergraph with $\mathcal{E} \subseteq \mathcal{E}_{\text{all}}$ (see Section 3.3).

3.2 Weighted Exact-Cover Optimization

Given a candidate hyperedge set $\mathcal{E}$ and a compatibility score $s(e)$ for each hyperedge, we propose to model multi-person association as a *weighted exact-cover problem* [27], optimizing over a disjoint subset of hyperedges, ensuring that every 2D pose detection is explained exactly once.

Definition 1 (Weighted Exact-Cover ILP)**.** *Let $\mathbf{x} = (x_e)_{e \in \mathcal{E}}$ with $x_e \in \{0, 1\}$ indicate whether hyperedge e is selected.* COMPOSE *solves*

$$\max_{\mathbf{x}} \quad \sum_{e \in \mathcal{E}} (s(e) - \gamma) x_e \tag{1}$$

$$\text{s.t.} \quad \sum_{e:\, u \in e} x_e = 1, \quad \forall u \in \mathcal{U}, \tag{2}$$

$$x_e \in \{0, 1\}, \quad \forall e \in \mathcal{E}. \tag{3}$$

Here, $s(e)$ is a hyperedge compatibility score, and $\gamma > 0$ penalizes the number of selected hyperedges, thereby favoring compact explanations that group geometrically consistent detections across views.

The resulting association is the selected cover $\mathcal{E}^\star = \{e \in \mathcal{E} : x_e = 1\}$, comprising vertex-disjoint hyperedges that jointly cover all observed 2D detections.

3.3 Hyperedge Scoring and Candidate Pruning

Hyperedge Scoring. To realize the ILP, we now specify the compatibility score $s(e)$ and prune $\mathcal{E}_{\text{all}}$ to a tractable candidate set $\mathcal{E}$. For calibrated cameras, we score non-singleton hyperedges by reprojection consistency. For each candidate hyperedge e with $|e| \geq 2$, let $\hat{\mathbf{y}}_j(e)$ be the DLT triangulation of joint j from the detections in e. We define the reprojection cost

$$\mathcal{C}(e) = \frac{1}{J|e|} \sum_{j=1}^{J} \sum_{U_i^v \in e} \left\| \pi_v(\hat{\mathbf{y}}_j(e)) - \mathbf{u}_{i,j}^v \right\|^2.$$

The corresponding compatibility score is

$$s(e) = \exp(-\lambda \cdot \mathcal{C}(e)),$$

where λ controls the sensitivity to reprojection error. Singletons receive a fixed prior score $s_{\texttt{single}}$.

The complete candidate set $\mathcal{E}_{\text{all}}$ contains all combinations of detections across non-empty subsets of views. We observe that the size of this problem grows exponentially with the number of views:

Proposition 1 (Candidate hyperedges)**.** *The total number of potential hyperedges M is given by:*

$$M = \sum_{\emptyset \neq S \subseteq \{1,\dots,V\}} \prod_{v \in S} n_v = \prod_{v=1}^{V} (1 + n_v) - 1$$

where S represents a subset of views and $n_v = |\mathcal{U}_v|$ is the number of detected poses in view v. This implies exponential growth with respect to the number of views V, yielding a complexity $\mathcal{O}((N+1)^V)$, where $N = \max_v n_v$ denotes the maximum number of detected poses in any single view.

Proof. For each view v, a valid hyperedge either selects one of the n_v detections or selects no detection from that view, giving $1 + n_v$ choices. Multiplying over all views and subtracting the all-empty choice yields

$$M = \prod_{v=1}^{V} (1 + n_v) - 1 \leq (1 + N)^V - 1 \in \mathcal{O}\big((N + 1)^V\big), \quad N = \max_v n_v.$$

□

Candidate Pruning. Proposition 1 shows that naive optimization over $\mathcal{E}_{\text{all}}$ is intractable as the number of views or detections grows. We therefore construct the candidate set $\mathcal{E}$ used in (1) by retaining (i) all singleton hyperedges, ensuring that the exact-cover constraints remain feasible, and (ii) multi-view hyperedges satisfying the geometric consistency criterion

$$\mathcal{C}(e) \leq \tau,$$

yielding a tractable set of physically plausible hypotheses. While we use reprojection error for calibrated cameras, our formulation is not tied to this specific metric: in weakly calibrated or uncalibrated settings, $\mathcal{C}(e)$ could instead be based on epipolar or trifocal consistency [21, 34, 24], and the compatibility score could incorporate off-the-shelf appearance descriptors when additional visual evidence is beneficial [63, 44]; we provide experiments in the supplementaries (Section D.7).

Solving the ILP. Although weighted exact-cover is NP-hard [27], the candidate pruning above substantially reduces the candidate set, enabling effective solution of (1) via an ILP solver with branch-and-cut [14], as shown in Fig. 5. The ILP returns the globally optimal cover over $\mathcal{E}$, after which selected non-singleton hyperedges are triangulated into 3D poses (see Section C.2 for details).

3.4 Probabilistic Relaxation via Belief Propagation

While the ILP returns a globally optimal hard assignment over the retained candidate set $\mathcal{E}$, its worst-case complexity remains exponential despite geometric pruning. Moreover, the ILP solution does not provide marginal association uncertainty, which is useful for downstream tasks such as tracking [1]. We therefore construct a probabilistic relaxation that associates each binary decision variable with a *continuous marginal belief*, which can be evaluated efficiently via message passing.

The key idea is to reinterpret the ILP objective as the energy of a Gibbs distribution over the binary assignment vector $\mathbf{x} = (x_e)_{e \in \mathcal{E}}$. The resulting factor graph defines a Markov random field over hyperedge-selection variables, and has the same MAP solution as the ILP under exact coverage constraints, while the relaxed version used for BP yields soft association beliefs that can be rounded into a discrete selection. We provide guiding intuitions in the supplementary material (Section B) and refer the reader to [59, 60] for the fundamentals of belief propagation.

Proposition 2 (Exact MAP–ILP equivalence)**.** *Define unary factors*

$$f_e(x_e) := \exp\big(\beta\,(s(e) - \gamma)\,x_e\big), \tag{4}$$

constraint factors

$$g_u\big(\mathbf{x}_{\mathcal{E}(u)}\big) = \begin{cases} 1, & \sum_{e \in \mathcal{E}(u)} x_e = 1, \\ 0, & \textit{otherwise}, \end{cases} \tag{5}$$

and the Gibbs distribution

$$p_\beta(\mathbf{x}) = \frac{1}{Z} \prod_{e \in \mathcal{E}} f_e(x_e) \prod_{u \in \mathcal{U}} g_u\big(\mathbf{x}_{\mathcal{E}(u)}\big), \tag{6}$$

where Z is a normalization constant. Then, for any $\beta > 0$, the MAP solution of (6) *coincides with the optimizer of the ILP* (1)*.*

Proof sketch. We observe that due to monotonicity of the objective function for $\mathbf{x} \in \{0,1\}$, $\arg\max_{\mathbf{x}} p_\beta(\mathbf{x}) = \arg\max_{\mathbf{x}} \log p_\beta(\mathbf{x})$. The constraint factors (5) enforce $g_u(\mathbf{x}_{\mathcal{E}(u)}) \in \{0,1\}$, sending any infeasible assignment (i.e., $\sum_{e \ni u} x_e \neq 1$ for some u) to $-\infty$ in log-space, so the feasible set is exactly that of the ILP. On this feasible set, $\log p_\beta(\mathbf{x}) = \beta \sum_{e \in \mathcal{E}} (s(e) - \gamma)\, x_e + \text{const}$. Since $\beta > 0$ is a positive scalar, it does not affect the $\arg\max$, which thus coincides with the ILP (1). We provide a formal analysis in the supplementary materials. □

3.5 Loopy Belief Propagation On The Factor Graph

Exact inference on the factor graph induced by (6) is intractable due to the loopy structure created by hyperedges that share detections. Thus, we use loopy belief propagation (BP) [60], which iteratively passes messages between variable nodes (hyperedges $e \in \mathcal{E}$) and factor nodes (detections $u \in \mathcal{U}$).

Relaxation Used For BP. To obtain a softer and more robust model for loopy BP, we relax the equality constraint with a tunable uncovered-node penalty:

$$g_u^{(\eta)}\big(\mathbf{x}_{\mathcal{E}(u)}\big) = \mathbb{1}\left[\sum_{e \in \mathcal{E}(u)} x_e \leq 1\right] \exp\left(-\eta\, \mathbb{1}\left[\sum_{e \in \mathcal{E}(u)} x_e = 0\right]\right), \qquad \eta \geq 0. \tag{7}$$

This defines a continuum: (i) $\eta = 0$ gives the pure "at-most-one" relaxation, (ii) $\eta \to \infty$ recovers exact coverage ($= 1$) in the limit. For BP inference, we set $\eta = 0$ for computational simplicity, yielding the standard exclusion factor used in our updates.

Since all variables are binary, each message reduces to a single scalar log-ratio. Let $m_{e \to u}$ denote the variable-to-factor log-ratio message from hyperedge e to detection u, and $n_{u \to e}$ the factor-to-variable message in the reverse direction:

$$m_{e \to u} \coloneqq \log \frac{\mu_{e \to u}(1)}{\mu_{e \to u}(0)}, \qquad n_{u \to e} \coloneqq \log \frac{\nu_{u \to e}(1)}{\nu_{u \to e}(0)} \tag{8}$$

where $\mu_{e \to u}$ and $\nu_{u \to e}$ are the standard BP messages in probability space.

BP Inference. Define the unary log-potential $\phi_e \coloneqq \beta\,(s(e) - \gamma)$. After T iterations, the approximate marginal belief is

$$b_e = \sigma\big(\phi_e + \sum_{u \in e} n_{u \to e}\big),$$

where σ is the logistic sigmoid, and b_e estimates the marginal probability that hyperedge e is selected, providing a soft association confidence for each hypothesis. The scalar BP update equations for $m_{e \to u}$ and $n_{u \to e}$ are given in the supplementary material (Section C.3). The final non-overlapping discrete selection is obtained by greedy rounding in decreasing order of b_e.

GPU-Accelerated Inference. Within each synchronous BP iteration, all variable-to-factor messages can be computed in parallel, followed by all factor-to-variable messages. In practice, we implement the updates as batched tensor operations in PyTorch [2], enabling GPU acceleration. The ILP solver, however, relies on branch-and-cut [14], which is not easily parallelizable, making COMPOSE-BP particularly efficient.

4 Experiments and Results

Datasets. We evaluate COMPOSE on four public multi-view human pose datasets covering controlled indoor, heavily occluded indoor, and outdoor sparse-camera settings. CMU Panoptic [26] and EgoHumans [28] are used for the main quantitative evaluation and camera-layout generalization experiments, while Shelf and Campus [6] evaluate robustness under indoor and fewer camera view settings (Section D.2).

Table 1: **Quantitative comparison on the CMU Panoptic dataset [26].** We report Average Precision (AP) at millimeter thresholds, Recall, and Mean Per Joint Position Error (MPJPE) in mm. † uses 9 temporal frames as input. ‡ uses the same 2D keypoint detector as our method. 4DA [61] uses CE bottom-up poses [51] and different 2D inputs from COMPOSE. Best results per supervision category (full-, self-, and optimization-based) are highlighted in **blue**, **orange**, and **green**.

Method	Average Precision (AP) (↑)				Recall (↑)	Error (↓)
	25	50	100	150	@500	MPJPE
Fully-Supervised						
Plane Sweep Pose [36]	92.12	98.96	**99.81**	99.84	–	16.75
Wu *et al.* [56]	93.93	98.93	99.78	99.90	**99.97**	15.63
TEMPO [12]	89.01	**99.08**	99.76	**99.93**	–	14.68
VoxelPose + 3DSA [9]	**94.20**	98.49	99.21	99.31	–	**13.98**
Self-Supervised						
SelfPose3d [47]	55.13	**96.44**	**98.46**	**98.98**	**99.60**	24.47
DSP† [38]	**57.60**	86.10	94.00	–	–	**23.10**
Optimization-Based						
ACTOR [46]	–	–	–	–	–	168.40
MvPose‡ [13]	37.63	95.70	97.84	98.28	99.60	26.46
4DA [61], frame-wise	42.40	88.90	–	–	92.17	26.65
4DA [61], temporal	47.15	90.87	–	–	91.76	24.28
COMPOSE-ILP (Ours)	66.70	98.23	**99.43**	**99.62**	**99.81**	**22.78**
COMPOSE-BP (Ours)	**68.88**	**98.37**	99.42	99.61	**99.81**	**22.78**

- **CMU Panoptic [26]** is a large-scale indoor dataset captured with a dense multi-camera system. Following prior work we use the standard evaluation protocol with cameras `3`, `6`, `12`, `13`, and `23`.
- **EgoHumans [28]** provides tagging, fencing, and lego assemble test sequences. We use eight static calibrated exocentric cameras, sampled at 2.5 FPS, for 1,942 frames and 6,032 ground-truth poses.

Metrics. For CMU Panoptic and EgoHumans, we report Average Precision (AP) at multiple 3D distance thresholds (AP_{25}, AP_{50}, etc.), Recall at 500 mm, and Mean Per-Joint Position Error (MPJPE). For camera-layout generalization, we report the mean AP (mAP) averaged over the evaluated thresholds.

MPJPE is computed over matched poses and should be interpreted alongside recall, since missing difficult people can reduce the reported error. Metric details, including matching rules and threshold definitions, are provided in the supplementary material (Section C.10).

Baselines. We evaluate two variants of COMPOSE: the exact ILP solver (-ILP) and the BP relaxation (-BP). We compare against fully supervised methods [36, 56, 12, 9], recent self-supervised methods [47, 38], and optimization-based methods [46, 13]. For MvPose [13], we run the official implementation using the same ViTPose++ 2D detections as COMPOSE. This isolates the association and reconstruction stage from differences in 2D detector quality. We additionally run 4DA [61] with Coupled Embeddings (CE) [51] in frame-wise and temporal modes. For generalization experiments, we run MVGFormer [35] and Faster-VoxelPose [58] using CMU0-trained weights without retraining. For other baselines, we report the numbers published by the respective authors.

4.1 Quantitative Results

CMU Panoptic. Table 1 reports the main quantitative comparison on CMU Panoptic, where COMPOSE consistently improves over optimization-based baselines and recent self-supervised methods while remaining competitive with fully supervised approaches.

Against MvPose [13], both COMPOSE variants improve across metrics under identical 2D detections. COMPOSE-ILP obtains the global optimum of the discrete objective over the retained set and achieves the lowest optimization-based MPJPE, reducing error from 26.46 mm to 22.78 mm. COMPOSE-BP yields soft hyperedge marginals for ranking predictions and achieves strong AP scores; AP_{25} improves from 37.63 for MvPose to 68.88.

Compared with fully supervised methods, COMPOSE remains competitive despite not training a 3D pose model. Although direct 3D supervision benefits strict precision metrics, COMPOSE surpasses

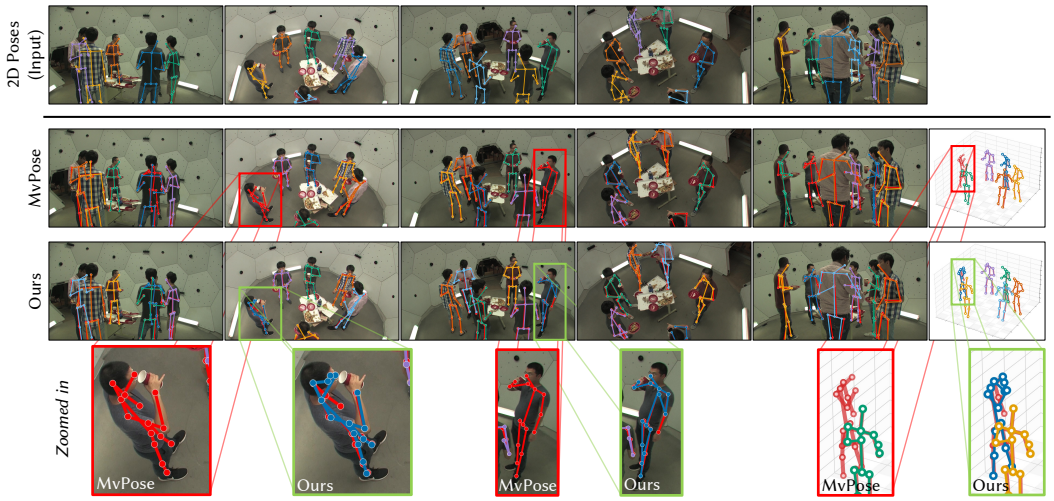


Figure 3: **COMPOSE recovers people missed by pairwise association**. Shown are Panoptic [26] 2D detections from ViTPose++ [57], and 3D predictions of MvPose [13] versus Ours. 3D GT is red; predictions are colored. MvPose misses the highlighted person; COMPOSE reconstructs everyone. recent self-supervised approaches, including SelfPose3d [47] and DSP [38], on all reported metrics. In particular, COMPOSE-ILP achieves lower MPJPE than DSP (22.78 mm vs. 23.10 mm), while COMPOSE-BP improves AP_{25} by 13.75 points over SelfPose3d. Temporal 4DA with CE reaches 47.15 AP_{25} and 24.28 mm MPJPE on Panoptic, versus 42.40 and 26.65 mm without temporality.

EgoHumans. Table 2 reports the evaluation on EgoHumans [28]. Under identical 2D detections, COMPOSE-ILP improves AP_{50} over MvPose from 62.06 to 69.50, recall from 95.86 to 99.49, and MPJPE from 37.22 to 32.68 mm. Temporal 4DA [61] achieves better AP_{25} (19.47 versus 17.23) and lower MPJPE (31.09 versus 32.68 mm), but recalls only 82.87% versus 99.49% of annotated people. We emphasize that EgoHumans is not fully annotated (Fig. 4), so plausible additional reconstructions may be false positives; AP and MPJPE should therefore be interpreted alongside recall.

4.2 Qualitative Results

Comparison on CMU Panoptic. Fig. 3 compares the reconstruction results of COMPOSE-ILP against MvPose [13] on the CMU Panoptic dataset. As illustrated in the zoomed-in regions, MvPose fails to establish correct correspondences for the highlighted individual, resulting in a missing reconstruction. In contrast, our method successfully processes the multi-view information and accurately reconstructs all individuals in the scene.

EgoHumans. Figure 4 illustrates the 2D human poses [57] taken as input as well as the reconstruction results of COMPOSE-ILP and MvPose [13] on EgoHumans. In the depicted scene there are four people in a sports hall with three people annotated (in red). While MvPose only reconstructs two people, and one of them inaccurately (green), COMPOSE-ILP recovers all of them more accurately, and even the non-annotated person (yellow).

4.3 Generalization to Camera Setups, Scalability, and Runtime

Generalization on Panoptic. Table 3 evaluates generalization across Panoptic camera setups with varying camera numbers and placements, testing how methods adapt to new layouts without

Table 2: **Quantitative comparison on EgoHumans [28].** AP/recall in %, MPJPE in mm. MvPose shares our 2D detections; 4DA uses CE [51]. Best values are **highlighted**.

Method	**Average Precision (AP)** (↑)				**Recall** (↑)	**Error** (↓)
	25	50	100	150	@500	MPJPE
MvPose	14.91	62.06	71.71	74.05	95.86	37.22
4DA (CE), frame-wise	15.37	58.72	64.26	66.13	79.68	36.17
4DA (CE), temporal	**19.47**	66.18	71.04	72.23	82.87	**31.09**
COMPOSE-ILP (Ours)	17.23	**69.50**	**79.86**	**81.09**	**99.49**	32.68
COMPOSE-BP (Ours)	14.31	68.55	77.76	78.90	99.30	32.36

Table 3: **Generalization across diverse camera setups on CMU Panoptic [26].** We report mAP and MPJPE in mm. The number of available cameras for each setup is indicated in parentheses. Supervised models use CMU0-trained weights without retraining. Best results per supervision category are highlighted in color.

Method	CMU1 (7)		CMU2 (7)		CMU3 (4)		CMU4 (4)	
	mAP ↑	MPJPE ↓	mAP ↑	MPJPE ↓	mAP ↑	MPJPE ↓	mAP ↑	MPJPE ↓
Fully-Supervised								
MVGFormer	**96.06**	**19.58**	**90.04**	**27.39**	**78.78**	45.43	**90.38**	**27.13**
Faster-VoxelPose	84.93	28.98	77.31	29.31	73.45	**34.96**	83.67	33.16
Self-Supervised								
SelfPose3d	**74.50**	**42.08**	**59.06**	**67.90**	**61.43**	**49.61**	**62.85**	**72.91**
Optimization-Based								
MvPose	84.62	29.44	80.07	39.01	59.74	68.32	74.85	41.68
COMPOSE-ILP (Ours)	**88.49**	26.45	**84.45**	32.65	73.83	46.90	**80.17**	**34.86**
COMPOSE-BP (Ours)	88.10	**26.44**	84.34	**32.59**	**74.43**	**46.85**	79.60	**34.86**

retraining. COMPOSE is stable across setups and consistently outperforms unsupervised baselines. MVGFormer [35] leads on all four layouts; compared with Faster-VoxelPose [58], COMPOSE-ILP has higher mAP on CMU1-3, while Faster-VoxelPose has lower MPJPE on CMU2-4. In sparse 4-view setups such as CMU3, COMPOSE achieves 74.43 mAP, versus 59.74 mAP for MvPose [13]. The BP relaxation is similar to COMPOSE-ILP across these settings, suggesting that the scalable relaxation preserves generalization behavior. Camera-count results are in Section D.6.

Generalization on EgoHumans. Table 4 reports the generalization performance on EgoHumans. We compare COMPOSE against fully-supervised methods trained on CMU panoptic [26]. COMPOSE-ILP reaches 68.35 mAP and 30.88 mm MPJPE, versus 49.19 and 76.97 mm for MVGFormer, and 46.39 and 75.39 mm for Faster-VoxelPose, without retraining on EgoHumans.

Table 4: **Generalization on the EgoHumans dataset [28].** Supervised models use CMU Panoptic [26] trained weights without retraining. Evaluation uses human joints common to all methods.

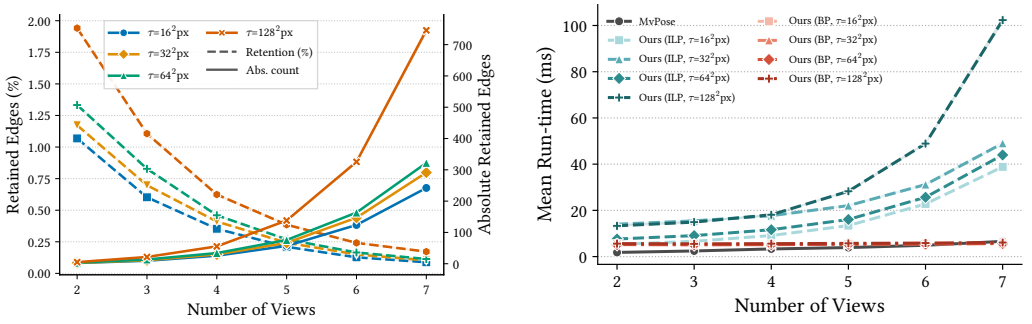
Method	Average Precision (↑)	Recall (↑)	Error (↓)
	mAP	@500	MPJPE (mm)
Fully-Supervised			
MVGFormer	49.19	88.26	76.97
Faster-VoxelPose	46.39	88.26	75.39
Optimization-Based			
COMPOSE-ILP (Ours)	**68.35**	**99.47**	**30.88**

Scalability and Runtime. Figure 5a analyzes geometric pruning. Although possible hyperedges increase with the number of views, only a small fraction satisfies the geometric consistency threshold,



Figure 4: **EgoHumans qualitative comparison.** Shown are EgoHumans [28] 2D detections from ViTPose++ [57], and 3D predictions of MvPose [13] versus COMPOSE (Ours). 3D GT is red; predictions are colored. MvPose misses the person highlighted in red. COMPOSE reconstructs everyone, even non-annotated people (blue). For clarity, only 4 of the 8 cameras are shown.

yielding a sparse candidate set for optimization. Figure 5b reports GH200 inference runtime as the number of cameras and pruning threshold τ vary. MvPose scales efficiently via pairwise correspondences, whereas COMPOSE-ILP slows as the retained hyperedge set grows, especially for larger V and τ, reflecting branch-and-cut worst-case complexity. In contrast, COMPOSE-BP uses batched GPU message passing and maintains an approx. constant runtime of 5 ms while closely tracking ILP accuracy. Runtimes exclude 2D pose detection and measure only association and reconstruction; candidate hyperedge-construction GPU memory is reported in the supplementaries (Section D.10).



(a) Hyperedge pruning vs. camera view count.

(b) Runtime vs. camera view count.

Figure 5: **Scalability analysis on Panoptic [26] across varying τ thresholds.** (a) Absolute count and percentage of hyperedges retained after pruning. (b) Runtime comparison of MvPose [13], COMPOSE-ILP, and COMPOSE-BP as the number of available cameras increases.

5 Concluding Remarks

We present COMPOSE, a higher-order combinatorial formulation for multi-view 3D pose estimation, solved at test time without 3D supervision. By recasting correspondence search as weighted exact-cover over a hypergraph of person hypotheses, our approach enforces multi-view consistency holistically rather than reconciling pairwise matches post-hoc. We provide two complementary solvers: an Integer Linear Programming formulation and a scalable Belief Propagation relaxation with parallelizable GPU inference and soft association beliefs. Extensive experiments across four benchmarks show state-of-the-art performance among optimization-based and self-supervised methods, while remaining competitive with fully supervised approaches. Our proposed method highlights the potential of higher-order graph formulations for multi-view 3D pose estimation, with applications to tracking [54] and team-behavior analysis [53].

Limitations and Future Work. COMPOSE requires calibrated cameras for triangulation, and cannot reconstruct persons visible in only a single view as there are no other observations to triangulate against. The formulation still operates frame-by-frame and does not exploit the temporal information readily available in multi-view recordings. Future work will explore weakly calibrated regimes, monocular priors for single-view recovery, and temporal hyperedges to enable multi-person tracking.

Acknowledgments and Disclosure of Funding. The authors gratefully acknowledge the computational resources of the LRZ AI service infrastructure provided by the Leibniz Supercomputing Center (LRZ) and Munich Center for Machine Learning (MCML), funded by the German Federal Ministry of Education and Research (BMBF) and Bavarian State Ministry of Science and the Arts (StMWK). T. W., L. B., and N. N. acknowledge support from the TUM Global Incentive Fund project TOPO-OR. T. B. was supported by a UKRI Future Leaders Fellowship (MR/Y018818/1).

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[2] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM international conference on architectural support for programming languages and operating systems, volume 2*, pages 929–947, 2024.

[3] Federica Arrigoni, Beatrice Rossi, and Andrea Fusiello. Spectral synchronization of multiple views in se (3). *SIAM Journal on Imaging Sciences*, 9(4):1963–1990, 2016.

[4] Federica Arrigoni, Eleonora Maset, and Andrea Fusiello. Synchronization in the symmetric inverse semigroup. In *International Conference on Image Analysis and Processing*, pages 70–81. Springer, 2017.

[5] Kristijan Bartol, David Bojanić, Tomislav Petković, and Tomislav Pribanić. Generalizable human pose triangulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11028–11037, 2022.

[6] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3D Pictorial Structures for Multiple Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, 2014.

[7] Tolga Birdal and Umut Simsekli. Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11105–11116, 2019.

[8] Tolga Birdal, Vladislav Golyanik, Christian Theobalt, and Leonidas J Guibas. Quantum permutation synchronization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13122–13133, 2021.

[9] Bo-Han Chen and Chia-chi Tsai. 3DSA: Multi-view 3D Human Pose Estimation With 3D Space Attention Mechanisms. In *Computer Vision – ECCV 2024*, volume 15085, pages 323–339. Springer Nature Switzerland, 2025.

[10] Keqi Chen, Vinkle Srivastav, Didier Mutter, and Nicolas Padoy. Learning from synchronization: Self-supervised uncalibrated multi-view person association in challenging scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24419–24428, 2025.

[11] Aviral Chharia, Wenbo Gou, and Haoye Dong. Mv-ssm: Multi-view state space modeling for 3d human pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11590–11599, 2025.

[12] Rohan Choudhury, Kris M. Kitani, and László A. Jeni. TEMPO: Efficient Multi-View Pose Estimation, Tracking, and Forecasting. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14704–14714. IEEE, 2023.

[13] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7792–7801, 2019.

[14] John Forrest and Robin Lougee-Heimer. Cbc user guide. In *Emerging theory, methods, and applications*, pages 257–277. INFORMS, 2005.

[15] Stuart Geman. Stochastic relaxation methods for image restoration and expert systems. In *Maximum-Entropy and Bayesian Methods in Science and Engineering: Volume 2: Applications*, pages 265–311. Springer, 1988.

[16] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6 (6):721–741, 1984.

[17] Michael A Goodrich and Alan C Schultz. Human–robot interaction: a survey. *Foundations and trends® in human–computer interaction*, 1(3):203–275, 2008.

[18] Venu Madhav Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004.

[19] Venu Madhav Govindu and A Pooja. On averaging multiview relations for 3d scan registration. *IEEE Transactions on Image Processing*, 23(3):1289–1302, 2014.

[20] Ruize Han, Yun Wang, Haomin Yan, Wei Feng, and Song Wang. Multi-view multi-human association with deep assignment network. *IEEE Transactions on Image Processing*, 31: 1830–1840, 2022.

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[22] Richard I Hartley. Lines and points in three views and the trifocal tensor. *International Journal of computer vision*, 22(2):125–140, 1997.

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[24] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 710–720. IEEE, 2021.

[25] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable Triangulation of Human Pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019.

[26] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[27] Richard M Karp. Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*, pages 219–241. Springer, 2009.

[28] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-Humans: An ego-centric 3D multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19807–19819, 2023.

[29] Jaihoon Kim, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Synctweedies: A general generative framework based on synchronized diffusions. *Advances in Neural Information Processing Systems*, 37:95198–95237, 2024.

[30] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[31] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[32] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

[33] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36: 50648–50660, 2023.

[34] Yu-Jhe Li, Yan Xu, Rawal Khirodkar, Jinhyung Park, and Kris Kitani. Multi-person 3d pose estimation from multi-view uncalibrated depth cameras. *arXiv preprint arXiv:2401.15616*, 2024.

[35] Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, and Steven L. Waslander. Multiple View Geometry Transformers for 3D Human Pose Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 708–717. IEEE, 2024.

[36] Jiahao Lin and Gim Hee Lee. Multi-View Multi-Person 3D Pose Estimation with Plane Sweep Stereo. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11881–11890. IEEE, 2021.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[38] Yang Liu and Zhiyong Zhang. DSP: Dense-Sparse Parallel Networks for Self-supervised 3D Multi-person Pose Estimation from Multiple Views. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4629–4638. ACM, 2025.

[39] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024.

[40] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2640–2649, 2017.

[41] Stuart Mitchell, Michael OSullivan, and Iain Dunning. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand*, 65:25, 2011.

[42] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[43] Ana Filipa Rodrigues Nogueira, Hélder P. Oliveira, and Luís F. Teixeira. Markerless multi-view 3D human pose estimation: A survey. *Image and Vision Computing*, 155:105437, 2025. ISSN 0262-8856.

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

[45] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[46] Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstruction. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[47] Vinkle Srivastav, Keqi Chen, and Nicolas Padoy. SelfPose3d: Self-Supervised Multi-Person Multi-View 3d Pose Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2512. IEEE, 2024.

[48] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[49] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. VoxelPose: Towards Multi-camera 3D Human Pose Estimation in Wild Environment. In *Computer Vision – ECCV 2020*, pages 197–212. Springer International Publishing, 2020.

[50] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[51] Haixin Wang, Lu Zhou, Yingying Chen, Ming Tang, and Jinqiao Wang. Regularizing vector embedding in bottom-up human pose estimation. In *Computer Vision – ECCV 2022*, pages 107–122, 2022. doi: 10.1007/978-3-031-20068-7_7.

[52] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct Multi-view Multi-person 3D Pose Estimation. In *Advances in Neural Information Processing Systems*, volume 34, pages 13153–13164. Curran Associates, Inc., 2021.

[53] Tony Danjun Wang, Lennart Bastian, Tobias Czempiel, Christian Heiliger, and Nassir Navab. Beyond role-based surgical domain modeling: Generalizable re-identification in the operating room. *Medical Image Analysis*, page 103687, 2025.

[54] Tony Danjun Wang, Christian Heiliger, Nassir Navab, and Lennart Bastian. Trackor: Towards personalized intelligent operating rooms through robust tracking. In *International Workshop on Collaborative Intelligence and Autonomy in Image-Guided Surgery*, pages 53–63. Springer, 2025.

[55] Tony Danjun Wang, Tolga Birdal, Nassir Navab, and Lennart Bastian. DisPOSE: Projected polystochastic diffusion for self-supervised multi-view 3D human pose estimation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 125100–125126. PMLR, 2026. URL `https://proceedings.mlr.press/v306/wang26l.html`.

[56] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-Based 3D Multi-Person Pose Estimation Using Multi-View Images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11128–11137. IEEE, 2021.

[57] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023.

[58] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection. In *Computer Vision – ECCV 2022*, pages 142–159. Springer Nature Switzerland, 2022.

[59] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Adv. Neural Inform. Process. Syst.*, 13(24), 2001.

[60] Jonathan S Yedidia, William T Freeman, Yair Weiss, et al. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236-239): 0018–9448, 2003.

[61] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D Association Graph for Realtime Multi-Person Motion Capture Using Multiple Video Cameras. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1321–1330. IEEE, 2020.

[62] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM computing surveys*, 56(1):1–37, 2023.

[63] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5056–5069, 2021.

[64] Zhize Zhou, Qing Shuai, Yize Wang, Qi Fang, Xiaopeng Ji, Fashuai Li, Hujun Bao, and Xiaowei Zhou. QuickPose: Real-time Multi-view Multi-person Pose Estimation in Crowded Scenes. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, pages 1–9. Association for Computing Machinery, 2022.

[65] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, pages 15085–15099, 2023.

Supplementary Material

In this supplementary material, we collect the additional context, derivations, implementation details, and empirical evidence supporting the main manuscript. We begin with related work in Section A, then expand on the ILP–MAP connection and BP relaxation in Section B; the following sections detail implementation and evaluation protocols (Section C), report extended experiments and ablations (Section D), and discuss broader societal impact (Section E).

A Extended Related Work

Multi-view multi-human 3D pose estimation can be stratified into two distinct paradigms: optimization-based and learning-based approaches [43]. Optimization-based methods rely on geometric triangulation and solver-based consensus to lift 2D priors (e.g., from an off-the-shelf 2D pose detector) into 3D space [6]. Conversely, learning-based approaches use deep neural networks to regress 3D poses from visual features, typically requiring annotated datasets [49].

Optimization-Based Approaches. As a pioneering work, Belagiannis et al. introduced *3D Pictorial Structures*, a conditional random field framework that uses a discrete state space and multi-view potential functions to resolve identity and body-part ambiguities across multiple camera views [6]. Dong et al. [13] establish multi-view correspondences by formulating the multi-way matching problem as a convex objective that simultaneously clusters 2D detections across all views using appearance similarity and geometric consistency cues while enforcing a *cycle-consistency* constraint. To address occlusions and crowded scenes, Zhou et al. [64] introduce multi-view association at the level of partial skeleton proposals instead of body-level. Zhang et al. [61] introduce the temporal dimension to the task and propose a spatio-temporal graph formulation for both spatial and temporal associations. These optimization-based methods are efficient and require minimal computational resources. However, compared to learnable methods, they struggle with noisy 2D detections and occlusions.

Learning-Based Approaches. The prevailing body of work leverages deep neural networks to regress 3D human poses. Early methods encode the environment as 3D voxel grids, learning with 3D CNNs [25, 49, 9, 47]. However, these are computationally expensive due to cubic complexity and often overfit to specific camera setups [35]. To address this, recent methods project 3D hypotheses onto 2D image planes to leverage 2D features, improving both speed and flexibility [52, 35, 11]. Despite these architectural advances, fully supervised models fundamentally depend on scarce, labor-intensive 3D-annotated data. While self-supervised approaches aim to reduce reliance on labeled data, they currently underperform and struggle to generalize to novel environments [47, 38]. Thus, optimization-based approaches remain a critical alternative, offering robust generalization in diverse settings without requiring costly 3D supervision.

Multi-View Synchronization. Reconciling locally consistent pairwise associations into a globally coherent structure has historically been treated as a *synchronization problem*, which is defined as the recovery of absolute quantities from a collection of ratios [19, 18, 4]. Pioneering works use Lie group theory to average rigid motions for *Structure-from-Motion*, extending to SE(3) via spectral decomposition and semidefinite programming, offering closed-form solutions and strong duality guarantees [3]. Recent advancements explore probabilistic synchronization [7], quantum permutation synchronization for non-convex optimization [8], and generative synchronization to align multiple joint diffusion processes [33, 29]. While cycle consistency can be mathematically enforced across pairwise matches, relying solely on synchronizing dyadic relations can propagate errors when detections in individual views are noisy or occluded.

B Background

We provide further intuition regarding the connection between integer linear programming (ILP), the Gibbs distribution, probabilistic relaxations such as Markov random fields, and the resulting belief propagation algorithm.

B.1 Candidate Hyperedge Count

For proposition 1, each view contributes one of its n_v detections or no detection, giving $1 + n_v$ choices. The all-empty selection is excluded.

Proof. For each view v, a valid hyperedge either selects one of the n_v detections or selects no detection from that view, giving $1 + n_v$ choices. Multiplying over all views and subtracting the all-empty choice yields

$$M = \prod_{v=1}^{V}(1 + n_v) - 1 \leq (1 + N)^V - 1 \in \mathcal{O}\big((N + 1)^V\big), \quad N = \max_v n_v.$$

□

B.2 Integer Linear Programming (ILP)

Let us first rephrase the ILP in definition 1 as a minimization problem:

$$\min_{\mathbf{x}} \quad -\sum_{e\in\mathcal{E}}(s(e) - \gamma)\, x_e \tag{9}$$

$$\text{s.t.} \quad \sum_{e:\, u\in e} x_e = 1, \quad \forall\, u \in \mathcal{U} \qquad \text{and} \qquad x_e \in \{0, 1\}, \quad \forall\, e \in \mathcal{E} \tag{10}$$

This can be interpreted as minimizing a corresponding energy function:

$$E(\mathbf{x}) = \sum_{e\in\mathcal{E}} c_e x_e, \qquad c_e := \gamma - s(e) \tag{11}$$

where c_e encodes the cost of selecting hyperedge e: a high-confidence hyperedge (large $s(e)$) yields a lower cost, making its selection energetically more favorable; the converse is true for low-confidence hyperedges. The key to relaxing the objective is that any optimization problem of the form $\min_{\mathbf{x}} E(\mathbf{x})$ can be cast as finding the mode of a corresponding probability distribution, the Gibbs distribution [15].

B.3 Gibbs Energy Formulation

Given our energy function in Eq. (11), we now define the Gibbs distribution (also called the Boltzmann distribution [42]) as:

$$p_\beta(\mathbf{x}) = \frac{1}{Z(\beta)} \exp\big(-\beta\, E(\mathbf{x})\big), \tag{12}$$

where $\beta > 0$ is an *inverse temperature* parameter and $Z(\beta) = \sum_{\mathbf{x}} \exp(-\beta\, E(\mathbf{x}))$ is a *partition function* to normalize the probability distribution [31].

The inverse temperature β controls the concentration of the distribution around low-energy states. In the *high-temperature regime* ($\beta \to 0$), the distribution becomes uniform over all configurations, assigning equal probability regardless of energy. In the *low-temperature regime* ($\beta \to \infty$), the distribution concentrates its mass on the global energy minimizer(s):

$$\lim_{\beta\to\infty} p_\beta(\mathbf{x}) = \begin{cases} \frac{1}{|\mathcal{X}^*|} & \text{if } \mathbf{x} \in \mathcal{X}^*, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $\mathcal{X}^* = \arg\min_{\mathbf{x}} E(\mathbf{x})$ is the set of optimal configurations [31]. Consequently, for any $\beta > 0$, the *maximum a posteriori* (MAP) estimate $\arg\max_{\mathbf{x}}\, p_\beta(\mathbf{x})$ coincides with $\arg\min_{\mathbf{x}}\, E(\mathbf{x})$, since β is a positive scalar. This relationship guarantees that solving the ILP is equivalent to MAP inference under the corresponding Gibbs distribution.

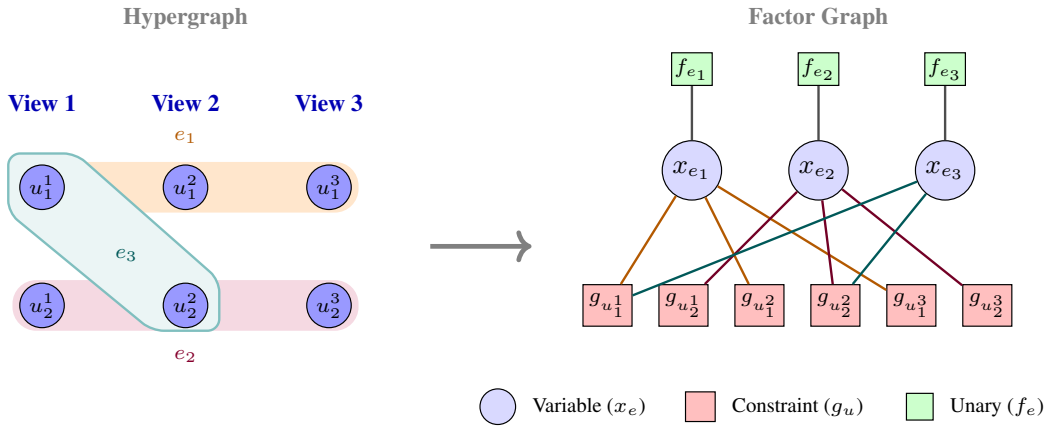


Figure 6: **From hypergraph to factor graph.** *Left:* A multi-view hypergraph with six detections across three views and three candidate hyperedges (e_1, e_2: cross-view correspondences; e_3: a partial hypothesis between two views). *Right:* The corresponding factor graph for belief propagation. Each hyperedge becomes a binary variable node x_e, each detection induces a constraint factor g_u enforcing exclusive assignment, and unary factors f_e encode hyperedge plausibility. Variables are coupled when their hyperedges share a detection (e.g., x_{e_1} and x_{e_3} via $g_{u_1^1}$).

Proposition 3 (ILP–MAP Equivalence)**.** *Let $E(\mathbf{x}) = \sum_{e\in\mathcal{E}}(\gamma - s(e))\, x_e$ be the energy function corresponding to the negated ILP objective, and let $\mathcal{F} = \{\mathbf{x} \in \{0,1\}^{|\mathcal{E}|} \mid \sum_{e:\, u\in e} x_e = 1,\ \forall\, u \in \mathcal{U}\}$ denote the feasible set. Define the constrained Gibbs distribution:*

$$p_\beta(\mathbf{x}) = \frac{1}{Z(\beta)} \exp\big(-\beta\, E(\mathbf{x})\big) \cdot \mathbb{1}[\mathbf{x} \in \mathcal{F}], \tag{14}$$

where $\mathbb{1}[\cdot]$ is the indicator function. Then, for any $\beta > 0$:

$$\arg\max_{\mathbf{x}}\ p_\beta(\mathbf{x}) = \arg\min_{\mathbf{x}\in\mathcal{F}} \sum_{e\in\mathcal{E}} (\gamma - s(e))\, x_e = \arg\max_{\mathbf{x}\in\mathcal{F}} \sum_{e\in\mathcal{E}} (s(e) - \gamma)\, x_e. \tag{15}$$

That is, MAP inference under p_β recovers the optimal solution of the ILP in definition 1.

Proof. The indicator $\mathbb{1}[\mathbf{x} \in \mathcal{F}]$ restricts the support to feasible configurations, and since $\log$ is strictly monotone and $\beta > 0$ is a positive constant, we have $\arg\max_{\mathbf{x}} p_\beta(\mathbf{x}) = \arg\min_{\mathbf{x}\in\mathcal{F}} E(\mathbf{x}) = \arg\max_{\mathbf{x}\in\mathcal{F}} \sum_{e\in\mathcal{E}} (s(e) - \gamma)\, x_e$, recovering the ILP in definition 1. This classical equivalence underlies a broad family of methods bridging combinatorial optimization and probabilistic inference [30, 16]. □

B.4 Markov Random Fields and Factor Graphs

We briefly review the graphical model formalism underlying our belief propagation solver, deferring to the seminal works of Pearl [45], Yedidia et al. [59, 60], and Wainwright & Jordan [50] for a thorough treatment.

A *Markov random field* (MRF) is an undirected graphical model in which a joint distribution factorizes over the cliques of a graph [45]:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c), \tag{16}$$

where ψ_c are non-negative potential functions and Z is the partition function. This factorization can be made explicit through a *factor graph* [32], a bipartite graph of variable nodes and factor nodes, where each factor ψ_c connects to exactly those variables it depends on. Factor graphs provide the natural domain for message-passing algorithms such as belief propagation [60].

Our Gibbs distribution in Eq. (14) admits precisely this structure. Specifically, it factorizes into *unary factors* $f_e(x_e)$, encoding the plausibility of each hyperedge, and *constraint factors* $g_u(\mathbf{x}_{\mathcal{E}(u)})$,

enforcing that each detection is explained exactly once:

$$p_\beta(\mathbf{x}) = \frac{1}{Z(\beta)} \prod_{e \in \mathcal{E}} f_e(x_e) \prod_{u \in \mathcal{U}} g_u\big(\mathbf{x}_{\mathcal{E}(u)}\big). \tag{17}$$

Two hyperedge variables are coupled if and only if they share a detection node, making the graph sparse. Moreover, this sparsity is further amplified by geometric pruning, which eliminates most candidate hyperedges, leaving only a few variable pairs to interact through shared detections. Figure 6 illustrates this construction on a small example with three views, two detections per view, and three candidate hyperedges. Each hyperedge e becomes a binary variable node x_e, each detection u induces a constraint factor g_u, and unary factors f_e sit atop each variable. Notably, x_{e_1} and x_{e_3} are coupled through $g_{u_1^1}$ because both hypotheses claim the same detection. The constraint factor enforces that at most one can be selected.

Unary (plausibility) factors. To encode the ILP objective in the Gibbs framework, we define each unary factor as:

$$f_e(x_e) := \exp\big(\beta\,(s(e) - \gamma)\,x_e\big), \tag{18}$$

so that $f_e(0) = 1$ and $f_e(1) = \exp(\beta(s(e) - \gamma))$. The resulting log-potential ratio is:

$$\phi_e := \log \frac{f_e(1)}{f_e(0)} = \beta\,(s(e) - \gamma), \tag{19}$$

which directly encodes the ILP coefficient scaled by the inverse temperature β.

Constraint factors. The ILP feasibility constraints are enforced by:

$$g_u\big(\mathbf{x}_{\mathcal{E}(u)}\big) = \begin{cases} 1, & \sum_{e \in \mathcal{E}(u)} x_e = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

which assigns zero probability to any configuration in which a detection is left uncovered or multiply assigned. Together with Eq. (18), MAP inference under Eq. (17) recovers the ILP solution, as established in proposition 3.

B.5 Constraint Relaxation for Belief Propagation (BP)

The hard equality constraint in Eq. (20) assigns zero probability to any configuration where a detection is uncovered ($\sum_{e \in \mathcal{E}(u)} x_e = 0$) or multiply assigned. To effectively apply loopy belief propagation, we must relax these to soft penalties, avoiding $\log 0$ and resulting numerical instabilities [50].

To obtain a well-behaved message-passing scheme, we relax the constraint factor to the "at-most-one" form with a tunable penalty $\eta \geq 0$ for uncovered detections, as introduced in the main paper (Eq. (7)):

$$g_u^{(\eta)}\big(\mathbf{x}_{\mathcal{E}(u)}\big) = \mathbb{1}\left[\sum_{e \in \mathcal{E}(u)} x_e \leq 1\right] \exp\left(-\eta\, \mathbb{1}\left[\sum_{e \in \mathcal{E}(u)} x_e = 0\right]\right). \tag{21}$$

This relaxation replaces the hard zero with a smooth penalty: uncovered detections are discouraged but not forbidden, providing a continuous landscape that BP can navigate. Setting $\eta = 0$ yields the pure exclusion constraint used in the BP updates derived in the main paper, while $\eta \to \infty$ recovers the exact coverage constraint in Eq. (20). We include Fig. 7 for a conceptual illustration of how BP works on the factor graph.

C Additional Methodological and Implementation Details

C.1 Normalized and Confidence-Weighted Hyperedge Compatibility

In the main manuscript (see Section 3.3), we define the hyperedge compatibility score as

$$s(e) = \exp(-\lambda \cdot C(e)),$$

where $C(e)$ measures the geometric inconsistency of the detections grouped by hyperedge e. In practice, we normalize this cost by the number of contributing joints and views, so that scores

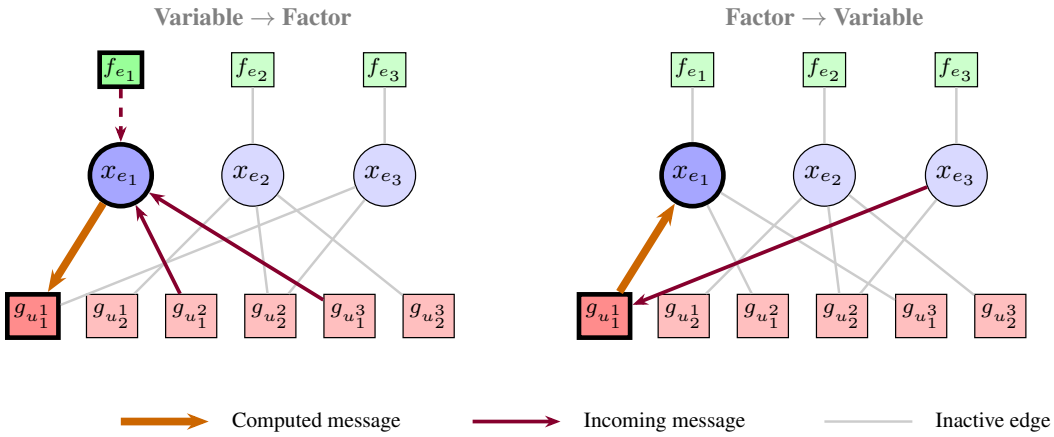


Figure 7: **Belief propagation message passing on the factor graph.** *Left:* A variable-to-factor message from x_{e_1} to $g_{u_1^1}$. The message aggregates the unary log-potential from f_{e_1} (dashed) and incoming messages from all other connected constraint factors. *Right:* A factor-to-variable message from $g_{u_1^1}$ back to x_{e_1}. The constraint factor collects the incoming message from the competing hypothesis x_{e_3} and sends a penalty reflecting how strongly e_3 claims the shared detection. Highlighted nodes are directly involved in the computation; dimmed edges are inactive.

remain comparable across hyperedges of different cardinalities. For the reprojection-based score, this corresponds to averaging the squared reprojection residuals over all scored joints and all detections contained in the hyperedge.

We additionally use the 2D detector confidence scores when computing the geometric cue. Let $c_{i,j}^v \in [0,1]$ denote the confidence of joint j in detection U_i^v. The reprojection cost is computed as a confidence-weighted average,

$$\mathcal{C}(e) = \frac{\sum_j \sum_{U_i^v \in e} c_{i,j}^v \left\| \pi_v(\hat{\mathbf{y}}_j(e)) - \mathbf{u}_{i,j}^v \right\|^2}{\sum_j \sum_{U_i^v \in e} c_{i,j}^v + \epsilon},$$

where ϵ is a small constant for numerical stability. Thus, unreliable joints contribute less to the hyperedge score, while the normalization prevents larger hyperedges from being penalized solely because they contain more observations. The resulting compatibility score $s(e)$ is used both for pruning geometrically implausible hyperedges and as the unary score in the ILP and BP objectives.

C.2 Confidence-Weighted Triangulation

After correspondence estimation, each selected non-singleton hyperedge $e \in \mathcal{E}^\star$ is converted into a 3D pose by triangulating its corresponding 2D joints. In the main manuscript, we denote this triangulated joint by $\hat{\mathbf{y}}_j(e)$. In practice, we use the confidence scores produced by the off-the-shelf 2D pose detector to perform confidence-weighted triangulation [25, 13].

Let $c_{i,j}^v \in [0,1]$ denote the confidence of joint j in detection U_i^v. For a selected hyperedge e and joint j, we triangulate $\hat{\mathbf{y}}_j(e) \in \mathbb{R}^3$ from the observations $\{\mathbf{u}_{i,j}^v : U_i^v \in e\}$ by solving the weighted least-squares problem

$$\hat{\mathbf{y}}_j(e) = \arg\min_{\mathbf{y} \in \mathbb{R}^3} \sum_{U_i^v \in e} w_{i,j}^v \left\| \pi_v(\mathbf{y}) - \mathbf{u}_{i,j}^v \right\|^2,$$

where $w_{i,j}^v$ is derived from the detector confidence $c_{i,j}^v$. In our implementation, we set $w_{i,j}^v = c_{i,j}^v$ and solve the corresponding weighted DLT system.

To reduce the influence of occasional outlier views, we apply a simple reprojection-error-based robustification step [21, 5]. After an initial confidence-weighted triangulation, we compute the per-view reprojection residual

$$r_{i,j}^v(e) = \left\| \pi_v(\hat{\mathbf{y}}_j(e)) - \mathbf{u}_{i,j}^v \right\|^2.$$

If at least three views are available for joint j, we discard the observation with the largest residual, and recompute $\hat{\mathbf{y}}_j(e)$ from the remaining observations.

C.3 BP Update Equations

With $\eta = 0$ and unary log-potential $\phi_e := \beta\,(s(e) - \gamma)$, the variable-to-factor and factor-to-variable log-ratio updates are

$$m_{e\to u} = \phi_e + \sum_{u' \in e \setminus \{u\}} n_{u'\to e}, \tag{22}$$

$$n_{u\to e} = -\log\left(1 + \sum_{e' \in \mathcal{E}(u)\setminus\{e\}} \exp\big(m_{e'\to u}\big)\right). \tag{23}$$

C.4 Probabilistic Relaxation via Belief Propagation

We solve the relaxed factor-graph formulation using damped loopy belief propagation. Algorithm 1 summarizes the procedure. We iteratively pass messages to estimate soft hyperedge beliefs b_e, and then greedily decode the final beliefs into a valid non-overlapping discrete selection. In practice, we terminate once the beliefs stabilize, i.e., when $\max_{e\in\mathcal{E}} |b_e^{t+1} - b_e^t| < \varepsilon$. Empirically, on the CMU Panoptic [26] dataset and with $\varepsilon = 10^{-3}$, the algorithm converges within three iterations for all evaluated instances. In Algorithm 1, we set $\gamma = 2.5$, $\beta = 0.5$, and $\alpha = 0.25$ for the Panoptic, Shelf, and Campus experiments.

C.5 2D Pose Estimator

For our top-down 2D pose estimation pipeline, we use ViTPose++ [57] with the `ViTPose-plus-huge` variant for pose estimation, together with RT-DETRv2 [39] (`r101` variant) for person detection. Both models are initialized from their official pretrained weights.

In Section D.8, we present experimental results analyzing the influence of various variants of the 2D person detector and 2D human pose estimator on overall performance. We also provide results using the default detectors used to create the 2D human pose labels in SelfPose3d [47].

C.6 ILP Solver

To solve the optimization problem using Integer Linear Programming (ILP), we implement our framework using the open-source Python library `PuLP` [41]. The underlying optimization process is driven by PuLP's `Coin-or branch and cut` (CBC) solver [14].

C.7 Hyperparameter Details

Table 5 summarizes the method hyperparameters $(\gamma, \lambda, \tau, \ldots)$ together with the BP-specific solver settings used in our experiments across the CMU Panoptic [26], Shelf [6], and Campus [6] datasets. We apply a consistent set of parameters across the Panoptic, Shelf, and Campus datasets and camera arrangements. The only exception is the Campus dataset [6], where a larger τ threshold is implemented to compensate for less accurate 2D pose detections, a consequence of the dataset's outdoor environment and lower image resolution.

C.8 Dataset Details

CMU Panoptic. Aligning with established evaluation protocols, we partition the CMU Panoptic dataset [26] identically to previous methodologies [49, 58, 56, 47, 35]. Hyperparameter tuning is conducted using a specific subset of sequences: `160422_ultimatum1`, `160224_haggling1`, `160226_haggling1`, `161202_haggling1`, `160906_ian1`, `160906_ian2`, `160906_ian3`, `160906_band1`, and `160906_band2`. The final assessment is performed on the sequences: `160422_haggling1`, `160906_pizza1`, `160906_ian5`, and `160906_band4`. Consistent with prior works, we extract every 12th frame from these test sequences, yielding 2,580 frames for the final evaluation.

Algorithm 1: Belief Propagation for Hypergraph Matching

Input: Hypergraph incidences (u, e), hyperedge scores $s(e)$, sparsity penalty γ, inverse temperature β, damping α, maximum iterations T, tolerance ε.

Output: Beliefs b_e and discrete selection $\mathbf{x}$.

Initialize variable-to-factor messages $m^0_{e\to u} \leftarrow 0$;

Initialize factor-to-variable messages $n^0_{u\to e} \leftarrow 0$;

Set log-potentials $\phi_e \leftarrow \beta\,(s(e) - \gamma)$;

for $t = 0, \dots, T-1$ **do**

`// Variable-to-factor update`

foreach *incidence* (u, e) **do**

$$\tilde{m}^{t+1}_{e\to u} \leftarrow \phi_e + \sum_{u' \in e \setminus \{u\}} n^t_{u'\to e}$$

$$m^{t+1}_{e\to u} \leftarrow \alpha \tilde{m}^{t+1}_{e\to u} + (1-\alpha) m^t_{e\to u}$$

`// Factor-to-variable update`

foreach *incidence* (u, e) **do**

if *sum-product* **then**

$$\tilde{n}^{t+1}_{u\to e} \leftarrow -\log\left(1 + \sum_{e' \in \mathcal{E}(u),\, e' \neq e} \exp(m^{t+1}_{e'\to u})\right)$$

else

$$\tilde{n}^{t+1}_{u\to e} \leftarrow -\max\left(0, \max_{e' \in \mathcal{E}(u),\, e' \neq e} m^{t+1}_{e'\to u}\right)$$

$$n^{t+1}_{u\to e} \leftarrow \alpha \tilde{n}^{t+1}_{u\to e} + (1-\alpha) n^t_{u\to e}$$

`// Belief computation`

foreach *hyperedge* e **do**

$$b^{t+1}_e \leftarrow \sigma\left(\phi_e + \sum_{u \in e} n^{t+1}_{u\to e}\right)$$

if $\max_e |b^{t+1}_e - b^t_e| < \varepsilon$ **then**

break

`// Greedy decoding`

$\mathbf{x} \leftarrow \textsc{GreedyRounding}(b)$;

return $b, \mathbf{x}$;

EgoHumans. We evaluate tagging, fencing, and lego assemble from the EgoHumans test set [28], using eight static calibrated exocentric cameras rather than wearable egocentric views. Sampling at 2.5 FPS yields 1,942 frames and 6,032 ground-truth poses. The primary comparison in Table 2 and the generalization comparison in Table 4 use different joint sets: the latter restricts evaluation to joints common to MVGFormer, Faster-VoxelPose, and COMPOSE.

Shelf Dataset. Evaluation on the Shelf dataset [6] is conducted using frames from frame index 300 to 600, which mirrors standard conventions established by recent literature [49, 58, 56, 47].

Campus Dataset. For the Campus dataset [6], we also adhere to the standard benchmark splits [49, 58, 56, 47]. The model is evaluated on frames 350 through 470, as well as 650 through 750.

Table 5: Summary of hyperparameters utilized for the CMU Panoptic [26], Shelf [6], and Campus [6] datasets.

Hyperparameter	Panoptic	Shelf	Campus
γ	2.5	2.5	2.5
λ	0.01	0.01	0.01
τ	$(32\,\mathrm{px})^2$	$(32\,\mathrm{px})^2$	$(64\,\mathrm{px})^2$
β (BP)	0.5	0.5	0.5
α (BP)	0.25	0.25	0.25
Max. BP iterations	10	10	10
ε (BP)	10^{-3}	10^{-3}	10^{-3}

C.9 Camera Configurations on CMU Panoptic

We assess the impact of varying camera configurations in our generalization experiments, utilizing the CMU Panoptic dataset [26]. We evaluate both the baseline models and our proposed approach under different camera arrangements and varying camera counts.

For these tests, we adopt the experimental framework introduced by Liao et al. (MVGFormer) [35], a standard subsequently also used by Chharia et al. (MV-SSM) [11]. The specific camera permutations deployed in our study are detailed in Table 6. For the `CMU4` configuration, we select the first four cameras, rather than the complete set of ten originally used in MVGFormer [35] to provide more settings with fewer cameras.

Table 6: Specification of camera setups used in our generalization studies. We list the specific Camera IDs alongside the total number of views for each configuration on the CMU Panoptic dataset [26].

Setup Name	**Camera IDs**	**# Views**
CMU0	3, 6, 12, 13, 23	5
CMU0 w/ 2 extra	3, 6, 12, 13, 23, 10, 16	7
CMU0(K)	First K cameras from "CMU0 w/ 2 extra"	K
CMU1	1, 2, 3, 4, 6, 7, 10	7
CMU2	12, 16, 18, 19, 22, 23, 30	7
CMU3	10, 12, 16, 18	4
CMU4	6, 7, 10, 12	4

C.10 Evaluation Metrics

We evaluate the quantitative performance of our 3D pose estimation framework using three primary metrics: Mean Per-Joint Position Error (MPJPE), Average Precision (AP), and Percentage of Correct Parts (PCP). These are the standard metrics used in each dataset.

- **MPJPE (Mean Per Joint Position Error):** MPJPE computes the Euclidean distance (reported in millimeters) between the predicted joint coordinates and their corresponding ground-truth locations. For any individual pose, this metric calculates the mean error across all visible GT joints:

$$E_{pose} = \frac{1}{J} \sum_{j=1}^{J} |\mathbf{p}_j - \mathbf{g}_j|_2, \tag{24}$$

  where $\mathbf{p}_j$ and $\mathbf{g}_j$ denote the estimated and actual 3D coordinates of the j-th joint, respectively, and J represents the total count of visible GT joints.

- **Average Precision (AP) and Recall:** A predicted human pose qualifies as a True Positive if its overarching pose-level error (E_{pose}) falls strictly below a defined threshold κ. We compute both Average Precision (AP) and Recall across a spectrum of thresholds, specifically $\kappa \in \{25, 50, \ldots, 150\}$ mm.

  Furthermore, we provide $\mathbf{Recall}_{500\mathbf{mm}}$, which highlights the fraction of ground-truth subjects successfully localized within a broader 500 mm error radius.

- **PCP (Percentage of Correct Parts):** This metric calculates the ratio of correctly predicted limbs. A specific limb (defined by a starting joint s and an ending joint e) is classified as "correct" provided that the mean positional error of its two endpoints does not exceed 50% of the actual ground-truth limb length. A limb satisfies this condition if:

$$\frac{\|\mathbf{p}_s - \mathbf{g}_s\|_2 + \|\mathbf{p}_e - \mathbf{g}_e\|_2}{2} \leq 0.5 \cdot \|\mathbf{g}_s - \mathbf{g}_e\|_2. \tag{25}$$

We note that PCP does not evaluate the precision of the predicted poses and only considers the recall of correctly localized limbs. This is because both the Shelf and Campus dataset [6] are not annotated exhaustively, and thus the precision of the predicted poses cannot be reliably measured.

D Additional Experiments

D.1 Extended CMU Panoptic Comparison

For completeness, Table 7 reports the CMU Panoptic comparison from Table 1 of the main manuscript with the additional fully supervised baselines (VoxelPose [49], MvP [52], Faster VoxelPose [58], MVGFormer [35], and MV-SSM [11]) that were omitted from the main paper for space. Including these methods does not alter the per-category best results as the highlighted entries in Table 1 remain the best within their respective supervision groups.

Table 7: **Extended quantitative comparison on the CMU Panoptic dataset [26].** Full version of Table 1 including all reported fully supervised baselines. We report Average Precision (AP) at millimeter thresholds, Recall, and Mean Per Joint Position Error (MPJPE) in mm. † uses 9 temporal frames as input. ‡ uses the same 2D keypoint detector as our method. Best results per supervision category (full-, self-, and optimization-based) are highlighted in **blue**, **orange**, and **green**.

Method	Average Precision (AP) (↑)				Recall (↑)	Error (↓)
	25	50	100	150	@500	MPJPE
Fully-Supervised						
VoxelPose [49]	83.59	98.33	99.76	99.91	–	17.68
Plane Sweep Pose [36]	92.12	98.96	**99.81**	99.84	–	16.75
MvP [52]	92.28	96.60	97.45	97.69	–	15.76
Faster VoxelPose [58]	85.22	98.08	99.32	99.48	–	18.26
Wu *et al.* [56]	93.93	98.93	99.78	99.90	**99.97**	15.63
TEMPO [12]	89.01	**99.08**	99.76	**99.93**	–	14.68
MVGFormer [35]	92.32	97.93	99.32	99.55	99.86	15.99
VoxelPose + 3DSA [9]	**94.20**	98.49	99.21	99.31	–	**13.98**
MV-SSM [11]	93.50	–	–	–	–	15.70
Self-Supervised						
SelfPose3d [47]	55.13	**96.44**	**98.46**	**98.98**	**99.60**	24.47
DSP† [38]	**57.60**	86.10	94.00	–	–	**23.10**
Optimization-Based						
ACTOR [46]	–	–	–	–	–	168.40
MvPose‡ [13]	37.63	95.70	97.84	98.28	99.60	26.46
4DA (CE), frame-wise	42.40	88.90	–	–	92.17	26.65
4DA (CE), temporal	47.15	90.87	–	–	91.76	24.28
COMPOSE-ILP (Ours)	66.70	98.23	**99.43**	**99.62**	**99.81**	**22.78**
COMPOSE-BP (Ours)	**68.88**	**98.37**	99.42	99.61	**99.81**	**22.78**

D.2 Extended Shelf and Campus Comparison

Shelf and Campus. Table 8 reports PCP results on Shelf and Campus. On Shelf, COMPOSE achieves an average PCP of 96.2%, outperforming the self-supervised baseline SelfPose3d (95.1%) and remaining competitive with the optimization-based MvPose baseline (96.9%). The proposed method performs particularly well on Actor 1, reaching 99.8% PCP.

Table 8: **Extended quantitative comparison on the Shelf [6] and Campus [6] datasets.** Best results per supervision category (full-, self-, and optimization-based) are highlighted in **blue**, **orange**, and **green**. A1, A2, and A3 refer to Actor 1, 2, and 3, respectively.

	Shelf (PCP %) (↑)				Campus (PCP %) (↑)			
Method	A1	A2	A3	Avg.	A1	A2	A3	Avg.
Fully Supervised								
VoxelPose [49]	99.3	94.1	97.6	97.0	97.6	93.8	**98.8**	96.7
Wu et al. [56]	99.3	**96.5**	97.3	97.7	–	–	–	–
MvP [52]	99.3	95.1	97.8	97.4	**98.2**	94.1	97.4	96.6
Faster VoxelPose [58]	**99.4**	96.0	97.5	97.6	96.5	94.1	97.9	96.2
TEMPO [12]	99.0	96.3	**98.2**	**98.0**	97.7	**95.5**	97.9	**97.3**
Self-Supervised								
SelfPose3d	**97.2**	**90.3**	**97.9**	**95.1**	**92.5**	**82.2**	**89.2**	**87.9**
Optimization-Based								
3DPS [6]	75.3	69.7	87.6	77.5	93.5	75.7	84.4	84.5
MvPose [13]	98.8	**94.1**	**97.8**	**96.9**	97.6	93.3	98.0	96.3
COMPOSE-ILP (Ours)	**99.8**	92.4	96.3	96.2	**99.4**	**94.3**	**98.1**	**97.3**
COMPOSE-BP (Ours)	**99.8**	92.4	96.3	96.2	**99.4**	**94.3**	93.6	95.7

On Campus, COMPOSE-ILP achieves 97.3% avg. PCP, outperforming MvPose (96.3%) and SelfPose3d (87.9%). This matches fully supervised TEMPO while requiring no 3D pose model training.

Shelf Ground-Truth Inaccuracies. Figure 8 compares our predictions with the public Shelf annotations. We observe several frames in which the annotated 3D pose appears misaligned with the image evidence. In the highlighted examples, the public annotations deviate from the visible actor location, whereas the COMPOSE prediction is visually consistent with the images. Such cases can penalize quantitatively correct predictions under PCP. For example, in the top example, the prediction receives 0% PCP for the right lower and upper arms despite visually matching the actor; in the bottom example, the right lower arm, upper arm, and head are similarly penalized.

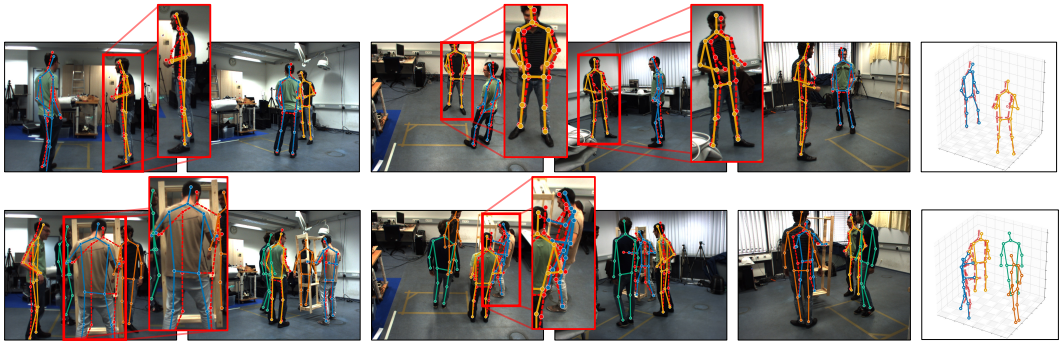


Figure 8: **Visual evidence exposes annotation noise**. In occlusion cases [6], COMPOSE aligns with image evidence even when the 3D annotation (dashed red) penalizes it (solid colored).

D.3 Pruning robustness

The geometric pruning discards hyperedges whose mean squared reprojection error exceeds τ, controlling the recall/precision balance of the candidate set fed to the matching solver. Table 9 sweeps $\sqrt{\tau}$ from 16 to 128 pixels on a subsample of the Panoptic training sequences (every 128th frame; 8 sequences, 545 frames), holding all other parameters at their default values. The mAP curve is unimodal at our default $\tau = (32\,\text{px})^2$: tightening to $\tau = (16\,\text{px})^2$ shifts the operating point toward precision – winning AP_{25} by 1.5 points but losing 2.3 AP_{50} – while loosening to $\tau \geq (64\,\text{px})^2$ admits noisy candidates that progressively erode AP_{25} from 53.92 down to 41.14 at $\tau = (128\,\text{px})^2$. Our default sits at the mAP peak and is the joint optimum for AP_{50}, AP_{100}, and $\text{Recall}_{500\text{mm}}$.

Table 9: **Pruning-threshold ablation on the CMU Panoptic train split [26].** Mean-squared reprojection threshold τ governs which hyperedges survive geometric pruning. All other parameters are kept at their default values (see Section C.7).

Threshold τ	**Average Precision (AP)** (↑)				**mAP**	**Recall**
	25	50	100	150	25–150	@500
$(16\,\mathrm{px})^2$	55.43	87.51	95.37	97.28	87.43	99.38
$(32\,\mathrm{px})^2$ (default)	53.92	89.78	96.51	98.15	88.35	99.55
$(64\,\mathrm{px})^2$	45.13	89.34	96.47	98.38	86.83	99.32
$(96\,\mathrm{px})^2$	42.67	88.53	96.24	98.33	86.11	99.27
$(128\,\mathrm{px})^2$	41.14	88.26	96.26	98.28	85.76	99.27

D.4 The Matching Solver Matters

To verify that joint matching provides additional benefit beyond what the geometric pruning already provides, we replace the optimizer with a naive greedy baseline. Greedy selects the highest-scoring hyperedge whose nodes are still unclaimed; ILP solves for the joint optimum, while BP provides a relaxed message-passing approximation. Greedy, given the same pruned graph, underperforms the optimizers by 5.2 AP_{50} and 3.3 mm MPJPE.

Table 10: **Solver evaluation on the CMU Panoptic test split.** The greedy heuristic, applied to the same pruned graph, cannot recover the ILP joint optimum.

Solver	**Average Precision (AP)** (↑)				**Recall** (↑)	**Error** (↓)
	25	50	100	150	@500	MPJPE
Greedy	61.30	93.19	97.51	98.02	99.80	26.05
COMPOSE-ILP	66.70	98.23	99.43	99.62	99.81	22.78
COMPOSE-BP	68.88	98.37	99.42	99.61	99.81	22.78

D.5 Robustness of The Matching Penalty

The ILP penalty γ is the only optimizer-side hyperparameter. Sweeping $\gamma \in [2.0, 3.0]$ at the default $\tau = (32\,\mathrm{px})^2$ produces similar AP at every reported threshold, indicating that the joint optimum is well-defined and insensitive to the penalty within a wide band. We use $\gamma = 2.5$ throughout. λ, in contrast, is cost-function-specific (Table 12) and is tuned per cost function rather than treated as a free parameter.

D.6 Generalization: Effect of Varying Camera Numbers

We study the effect of varying the number of available cameras on the standard CMU Panoptic setup (CMU0) by progressively adding views. Table 11 reports mAP and $\mathrm{Recall}_{500\mathrm{mm}}$ for subsets of 3, 4, 6, and 7 cameras. As expected, all methods benefit from additional views, yet COMPOSE exhibits the most consistent gains. Even in the challenging 3-camera regime, COMPOSE-ILP achieves 72.01 mAP, outperforming MvPose (58.05) by a large margin and surpassing the self-supervised SelfPose3d (66.42), which requires training supervision. As the number of views increases, our hypergraph formulation leverages a richer multi-view consistency signal: at 7 cameras, COMPOSE-BP achieves 95.17 mAP, compared to 89.80 for MvPose. Notably, the gap between COMPOSE-ILP and COMPOSE-BP remains small across all configurations, suggesting that the BP relaxation provides a close approximation to the exact ILP solution under different evaluated settings.

D.7 Effect of Different Cost Functions

We evaluate six geometric cost functions $\mathcal{C}$ for hyperedge weighting:

1. The *epipolar distance* [21] computes the average absolute epipolar constraint over all view pairs in normalized camera coordinates.

Table 11: **Generalization across camera numbers on CMU Panoptic [26].** We report mean Average Precision (mAP) and $\text{Recall}_{500\text{mm}}$ (Rec.). The number of available cameras for each setup is indicated in parentheses. Best results are highlighted in color.

Type	Method	CMU0(3)		CMU0(4)		CMU0(6)		CMU0(7)	
		mAP	Rec.	mAP	Rec.	mAP	Rec.	mAP	Rec.
Self-Sup.	SelfPose3d [47]	66.42	93.40	86.59	**99.44**	78.77	99.41	78.73	99.20
Optim.	MvPose [13]	58.05	96.65	79.13	98.99	89.00	99.68	89.80	99.73
	COMPOSE-ILP (Ours)	**72.01**	**97.92**	**86.65**	**99.44**	94.34	**99.86**	95.06	**99.86**
	COMPOSE-BP (Ours)	70.01	**97.92**	86.52	99.43	**94.58**	**99.86**	**95.17**	**99.86**

2. The *Sampson error* [21] normalizes the squared epipolar constraint by the epipolar line gradients, approximating the geometric distance to the epipolar line in pixel space using the fundamental matrix.
3. The *trifocal tensor* [22] generalizes to view triples by evaluating an algebraic constraint derived from the trifocal tensor using the known projection matrices, averaged over all triples within the hyperedge; for hyperedges containing only two views, we fall back to the Sampson error.
4. The *reprojection error* (our default) [21] triangulates a 3D point and measures the average squared pixel distance to the observed detections.
5. The *ray-to-ray distance* aggregates pairwise forward-ray distances in meters.
6. The *point-to-ray distance* triangulates a confidence-weighted consensus point and measures its RMS distance to the observation rays in meters.

Table 12 compares geometric and photometric cost functions for hyperedge weighting. Each cost function uses its own tuned λ. Within the geometric scores, the epipolar distance performs worse than the alternatives. Sampson error and the trifocal tensor yield similar results, whereas the reprojection error improves AP_{25} and MPJPE over both.

Table 12: **Different cost functions on CMU Panoptic [26].** We compare different cost functions $\mathcal{C}$ used for hyperedge weighting in COMPOSE-ILP. The first block lists geometric cost functions. The second extends the ablation with photometric alternatives: person re-identification features (OSNet-AIN [63], 512-d) and DinoV2 features (ViT-B/14 [44], 768-d). "Alone" uses the photometric cue both for pruning and for matching score; "$\times$" multiplies the photometric score onto the geometric one on the geometrically-pruned graph.

Score Function	Average Precision (AP) (↑)				Recall (↑)	Error (↓)
	25	50	100	150	@500	MPJPE
Epipolar Distance	62.93	96.22	97.45	97.70	97.98	23.01
Sampson Error	63.46	97.77	99.03	99.21	99.42	23.02
Trifocal Tensor	63.54	97.93	99.26	99.49	99.69	23.08
Reprojection Error (default)	66.70	98.23	99.43	99.62	99.81	22.78
Ray-to-Ray Distance	66.76	98.21	99.37	99.61	99.80	22.82
Point-to-Ray Distance	66.82	98.26	99.45	99.64	99.82	22.79
Person Re-ID (alone)	57.92	94.00	95.64	95.80	96.45	24.35
DinoV2 (alone)	22.35	32.86	34.30	34.80	44.94	57.26
Reprojection $\times$ Re-ID	66.69	98.25	99.45	99.64	99.82	22.78
Reprojection $\times$ DinoV2	66.68	98.25	99.45	99.63	99.82	22.79

Ray-to-ray and point-to-ray give nearly unchanged Panoptic performance. On EgoHumans (Table 13), ray-to-ray degrades performance overall, while point-to-ray improves AP_{25} and MPJPE with slightly lower recall.

Table 13: **Geometric cost comparison on EgoHumans using COMPOSE-ILP.** AP_{25} and R@500 are percentages; MPJPE is in mm. Reprojection remains the default. Best values are highlighted in **bold**.

Cue	AP_{25} ↑	R@500 ↑	MPJPE ↓
Pixel reprojection	17.23	**99.49**	32.68
Ray-to-ray	14.83	97.35	34.24
Point-to-ray	**18.54**	99.32	**32.02**

The point-to-ray parameters transfer unchanged between datasets, whereas reprojection and ray-to-ray use dataset-specific parameters (Table 14).

Table 14: **Cue-specific pruning thresholds and score bandwidths.** Each pair is (τ, λ) for $s(e) = \exp[-\lambda \mathcal{C}(e)]$.

Cue	Panoptic	EgoHumans
Pixel reprojection	$1024\,\mathrm{px}^2,\ 0.01\,\mathrm{px}^{-2}$	$2048\,\mathrm{px}^2,\ 0.0005\,\mathrm{px}^{-2}$
Ray-to-ray	$0.20\,\mathrm{m},\ 50\,\mathrm{m}^{-1}$	$0.025\,\mathrm{m},\ 0.5\,\mathrm{m}^{-1}$
Point-to-ray	$0.10\,\mathrm{m},\ 2.0\,\mathrm{m}^{-1}$	$0.10\,\mathrm{m},\ 2.0\,\mathrm{m}^{-1}$

For the photometric cues, "alone" applies a feature-distance threshold to prune candidate hyperedges and uses the same cosine-similarity score for matching; "×" applies the photometric score multiplicatively on top of the geometric pruning. Person re-ID features [63] alone produce a viable but inferior matching cue (-4.2 AP_{50}, $+1.6$ mm MPJPE vs. reprojection), suggesting that appearance carries some multi-view correspondence signal but is dominated by geometric consistency. DinoV2 features [44] alone fail (-65 AP_{50}, $+34$ mm MPJPE), implying that generic self-supervised features might not be discriminative enough between people for this task. When applied on top of the geometric pruning, photometric scores have no measurable effect.

D.8 Effect of Different 2D Pose Estimation Prior

We investigate the sensitivity of COMPOSE-ILP to the quality of the 2D pose prior by varying the backbone capacity of both the person detector and the top-down pose estimator. Specifically, we evaluate four configurations on the CMU Panoptic dataset: the `K-RCNN+HRNet` setup used by SelfPose3d [47] to generate its pseudo labels, a lightweight setup using a ResNet-18 backbone for RT-DETRv2 [39] (`R-18`) paired with the small version of the ViTPose++ pose estimator [57] (`Small`), an intermediate setup with a ResNet-50 backbone (`R-50`) and the base pose-estimator variant (`Base`), and our default configuration using ResNet-101 (`R-101`) with the huge pose-estimator variant (`Huge`).

Table 15: **Effect of different 2D pose estimation priors on the CMU Panoptic dataset [26].** We vary the person detector (`K-RCNN` [23], `RT-DETR` variants [39]) and pose estimator (`HRNet` [48], `ViTPose` variants [57]) backbones in COMPOSE-ILP. The `K-RCNN+HRNet` configuration corresponds to the 2D prior used by SelfPose3d [47] to generate pseudo labels. Best results are highlighted in **bold**.

Person Detector	Pose Estimator	Average Precision (AP) (↑)				Recall (↑)	Error (↓)
		25	50	100	150	@500	MPJPE
`K-RCNN`	`HRNet`	60.32	97.62	99.23	99.47	99.73	23.91
`R-18`	`Small`	63.45	97.79	99.26	99.51	99.75	23.43
`R-50`	`Base`	62.97	98.22	99.33	99.52	**99.81**	23.67
`R-101`	`Huge`	**66.70**	**98.23**	**99.43**	**99.62**	**99.81**	**22.78**

Table 15 reports the results. All configurations achieve comparable performance at relaxed AP thresholds (AP_{100} and above) and $\mathrm{Recall}_{500\mathrm{mm}}$, indicating that our hypergraph formulation is robust to moderate variations in 2D input quality. The primary differences emerge at the strictest threshold

(AP_{25}), where the stronger backbone yields a +3.25 improvement (66.70 vs. 63.45), and in MPJPE, where more precise 2D localizations translate into lower triangulation error (22.78 mm vs. 23.43 mm). This indicates that while the correspondence matching stage of COMPOSE is largely insensitive to the 2D backbone, the final triangulation accuracy naturally benefits from more precise 2D keypoint detections.

D.9 Cross-View Consensus Accuracy

We evaluate cross-view association on the full Panoptic test split [26], giving all methods the same corrupted detections with ten paired seeds per condition. The metric is the fraction of people whose surviving detections form one predicted identity without any detection from another person.

Table 16: **Association under controlled detection dropout.** Fractions of correctly associated people; higher is better. All methods receive the same corrupted inputs.

Condition	**COMPOSE-BP**	**COMPOSE-ILP**	**MvPose**
Clean	0.992	0.992	0.862
20% person-views dropped	0.988	0.989	0.860
40% person-views dropped	0.978	0.978	0.839
20% joints dropped	0.942	0.944	0.847
40% joints dropped	0.594	0.597	0.505

D.10 Hypergraph Construction

Materializing the full V-partite candidate set scales as $O(N^V)$ and exhausts a 16 GB GPU around $V = 8$ (Table 17); we instead build the hypergraph lazily. Order-2 hyperedges are formed from every pair of detections across distinct views and pruned by the threshold τ; each subsequent order is constructed by extending the surviving k-hyperedges with one detection from a new view and re-pruning under the same threshold. Peak memory, therefore, scales with the number of geometrically plausible survivors at each level rather than with the combinatorial space, enabling inference at $V \geq 10$ on commodity hardware.

Table 17: **Peak GPU memory of hyperedge construction across camera counts on CMU Panoptic [26]** Reported as the mean over all test frames, in GB. The flat enumerate-then-prune approach materializes the full $O(N^V)$ candidate space before applying the geometric threshold τ and runs out of memory on a 16 GB GPU at $V \geq 8$. Our construction prunes at every order and stays under 0.1 GB up to $V = 10$.

Construction	**Number of Views V**								
	2	3	4	5	6	7	8	9	10
Enumerate-Then-Prune	0.008	0.010	0.022	0.103	0.630	5.819	OOM	OOM	OOM
Lazily (Ours)	0.008	0.008	0.009	0.010	0.012	0.017	0.026	0.045	0.085

E Broader Societal Impact

This work advances optimization-based multi-view multi-person 3D pose estimation without requiring 3D supervision. It may benefit applications such as sports analysis, healthcare, operating-room monitoring, and human-robot collaboration, where calibrated multi-camera setups are common and accurate spatial localization is crucial. However, 3D human pose estimation raises privacy concerns, including biometric profiling and potential surveillance misuse. These risks are partly mitigated by the controlled acquisition setting required by COMPOSE, which assumes multiple synchronized calibrated cameras and is therefore less suited to unconstrained public deployment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the proposed hypergraph formulation, the exact ILP solver, the BP-based probabilistic relaxation, and the geometric pruning strategy. The methodological claims are supported by the ILP formulation and its factor-graph relaxation, while the empirical claims are supported by experiments on CMU Panoptic, EgoHumans, Shelf, and Campus.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The manuscript includes a dedicated "Limitations and Future Work" paragraph in the Concluding Remarks section discussing the main scope limitations: calibrated cameras, inability to reconstruct single-view-only persons without additional priors, and frame-by-frame inference without temporal modeling. It also identifies future directions addressing these limitations, including weak calibration, monocular priors, and temporal hyperedges.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for the theoretical results, including binary hyperedge-selection variables, exact-cover constraints, positive inverse temperature $\beta > 0$, and the retained candidate hyperedge set $\mathcal{E}$. The main paper provides proof sketches, and the supplementary material provides the corresponding formal derivations.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the datasets, evaluation protocols, metrics, baselines, 2D detector and pose estimator, hyperedge construction, scoring and pruning, ILP and BP solvers, and main hyperparameters needed to reproduce the reported results. The code is available at `https://github.com/wngTn/COMPOSE`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is publicly available at `https://github.com/wngTn/COMPOSE` To support reproducibility, the paper and supplementary material describe the datasets, preprocessing, evaluation protocols, hyperedge construction, ILP and BP solvers, hyperparameters, and implementation details used for the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the evaluated datasets, standard data splits and protocols, metrics, baselines, 2D detector and pose estimator, hyperedge scoring and pruning procedure, ILP solver, BP inference settings, and runtime hardware. Additional implementation and hyperparameter details are provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The main results are based on deterministic optimization pipelines and established evaluation protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports inference runtime on an NVIDIA GH200, and provides GPU-memory measurements in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research uses established public datasets and standard evaluation protocols, does not introduce new human-subject data collection, and does not release high-risk generative or surveillance datasets.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The manuscript contains a "Broader Societal Impact" section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release anything that requires special release safeguards. The experiments use existing public benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the existing datasets, models, and software packages used in the experiments, including CMU Panoptic, Shelf, Campus, ViTPose++, RT-DETRv2, PuLP, and CBC.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release a new dataset or pretrained model. The code is available at `https://github.com/wngTn/COMPOSE`.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or new research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve new data collection, crowdsourcing, or direct interaction with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.