SemaPop-GAN: Persona-Conditioned Population Synthesis with Marginal Consistency

Zhenlin Qin[1], Yancheng Ling[1], Leizhen Wang[3], Francisco Câmara Pereira[4], and Zhenliang Ma*[1,2]

[1]Department of Civil and Architectural Engineering, KTH Royal Institute of Technology
[2]Digital Future, KTH Royal Institute of Technology
[3]Department of Data Science and Artificial Intelligence, Monash University
[4]Department of Technology, Management and Economics, Technical University of Denmark

Abstract

Population synthesis is a core component of agent-based transport modeling, requiring realistic individual-level populations that preserve both statistical consistency and behavioral heterogeneity. Existing methods primarily rely on structured attributes and marginal constraints, limiting their ability to capture high-level behavioral semantics. We propose SemaPop-GAN, a semantic-conditioned generative framework that integrates large language model (LLM)-derived persona representations into population synthesis. To ensure statistical validity, we introduce marginal regularization during training and a post-hoc marginal calibration procedure for deployment. Extensive experiments demonstrate that SemaPop-GAN substantially improves generative performance, achieving closer alignment with both target marginal and joint distributions while preserving sample-level feasibility and diversity under semantic conditioning. Furthermore, through post-hoc marginal calibration, we show that SemaPop-GAN remains robust under additional calibration constraints. Overall, our framework provides a principled approach to bridging semantic expressiveness and statistical consistency in population synthesis, enabling more realistic and scenario-aware transport modeling.
Keywords: Population Synthesis; Persona Representation; Semantic-Conditioned Generative Modeling; Large Language Model

1 Introduction

Individual-level simulation frameworks, such as agent-based models and micro-simulation, play a central role in transportation and urban systems analysis by capturing heterogeneous behaviors and policy impacts (Publishing, 2017; Helbing, 2009; Batty, 2013). These approaches rely critically on synthetic populations, which initialize agents with realistic demographic and behavioral attributes (Bonabeau, 2002; Gilbert & Troitzsch, 2005; Miller, 2018; Wang et al., 2025). As a result, the quality of population synthesis directly affects the credibility of downstream simulation and policy evaluation.

A key challenge in population synthesis is to represent both statistical consistency and higher-level behavioral heterogeneity. Traditional methods, such as IPF/IPU and related approaches, focus on matching marginal distributions or dependency structures (Beckman et al., 1996; Norman, 1999; Jeong et al., 2016), but remain limited in capturing latent behavioral patterns and adapting to scenario-driven changes. Combinatorial optimization approaches generate populations under predefined constraints, yet do not learn underlying microdata distributions and thus exhibit limited generalization (Hermes & Poulsen, 2012; Müller & Axhausen, 2010; Barthelemy & Toint, 2013). Recent deep generative models improve statistical realism by learning latent representations from survey data, but they primarily operate on structured attributes and lack mechanisms to encode higher-level semantics such as lifestyles, preferences, and mobility tendencies (Aemmer & MacKenzie, 2022; Kim & Bansal, 2023; Garrido et al., 2020).

To address this limitation, we develop *SemaPop-GAN*, a persona-conditioned population synthesis framework that integrates semantic abstraction with statistical modeling. We leverage large language models (LLMs), which have demonstrated strong capabilities in semantic representation (Brown et al., 2020; Bommasani, 2021), to transform structured survey attributes into high-level personas. These personas provide interpretable descriptions of individual characteristics and are

encoded into continuous embeddings that condition the generative model, enabling semantic information to guide the synthesis of population agents. To maintain consistency with population-level statistics required in transportation applications, we incorporate marginal regularization during training, encouraging the model to capture key marginal distributions as implicit structural priors while preserving generative flexibility. In addition, to assess whether the learned model has effectively internalized these marginal-constrained priors, we introduce a post-hoc marginal calibration procedure that reweights generated samples to match target marginal distributions. This calibration is used as an analysis tool rather than a training component, allowing us to disentangle marginal mismatch from model-induced discrepancies and to evaluate the extent to which the generative model captures the desired population structure.

2 Preliminaries

Population synthesis in this study is based on individual-level survey data in tabular form, where each row represents a population agent described by demographic, household, and behavioral attributes. Unlike general tabular data generation, the goal is not to replicate the observed dataset, but to generate a synthetic population that is both behaviorally plausible at the individual level and statistically consistent with population-level distributions.

Population agent. A population agent $\mathbf{x}_i \in X$ represents an individual characterized by a set of attributes, including categorical (e.g., gender, occupation) and numerical variables (e.g., age, income). These attributes collectively describe the socio-demographic and behavioral profile of an individual in the population.

Persona. To provide a higher-level and interpretable representation, each agent $\mathbf{x}_i$ is associated with a persona π_i, defined as a textual description summarizing key characteristics such as lifestyle and mobility preferences. Personas are automatically generated from structured attributes using a large language model.

Problem formulation. Given a limited survey dataset $X = \{\mathbf{x}_i\}_{i=1}^{n}$, where $n \ll N$, the objective is to learn a conditional generator

$$\hat{\mathbf{x}} = F(\boldsymbol{\epsilon}, \pi), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

that approximates the underlying population distribution. The generated population should preserve realistic individual-level characteristics while matching key aggregate statistics of the target population.

3 Methodology

SemaPop-GAN Overview

We propose *SemaPop-GAN*, a persona-conditioned population synthesis framework show in Figure 1, which integrates semantic abstraction with statistical modeling. The framework aims to bridge structured survey-based representations with interpretable behavioral descriptions, while ensuring consistency with population-level statistics.

Given structured population agents $\mathbf{x}$, we first construct a high-level semantic representation (persona) using a frozen large language model (LLM). The persona is then encoded into a continuous embedding $\mathbf{e}$ and used to condition a generative model that synthesizes new agents $\hat{\mathbf{x}} = G(\mathbf{z}, \mathbf{e})$ from latent noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. During training, the model is optimized to produce realistic samples while matching target marginal distributions. During generation, new populations can be synthesized from persona inputs alone.

Persona Knowledge Representation

In population and survey research, individuals are often represented by high-dimensional demographic and behavioral attributes. While statistically expressive, such representations are difficult to interpret when analyzing heterogeneous populations. To address this, prior work has introduced higher-level abstractions—such as behavioral segments or latent constructs—to capture coherent patterns and link micro-level data to macro-level understanding (Anable, 2005; Walker & Ben-Akiva, 2002; Ben-Akiva et al., 2002). Personas extend this idea by providing an interpretable

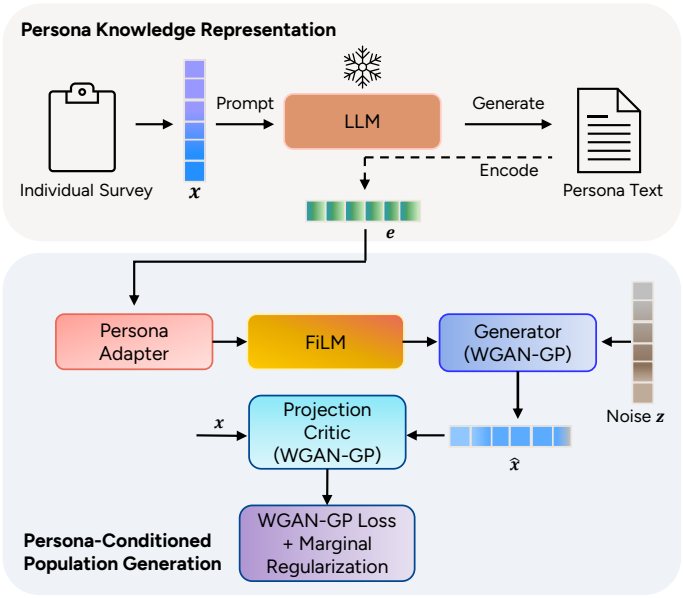


Figure 1: Overview of SemaPop-GAN. Structured survey attributes are transformed into persona representations using a frozen LLM, which are then encoded and used to condition the generative model. The framework integrates semantic conditioning with statistical constraints for population synthesis.

semantic layer that summarizes multiple attributes into representative population profiles, enabling more structured reasoning without replacing formal statistical representations (Cooper, 1999; Pruitt & Grudin, 2003).

In this study, we operationalize personas using a frozen LLM that transforms structured survey attributes into concise textual descriptions. The LLM extracts high-level semantic characteristics while abstracting away low-level numeric details, guided by a prompt template (Fig. 2) that emphasizes qualitative summaries of personal profiles and travel preferences. This abstraction aligns with the role of personas and is particularly suitable for population synthesis, where exact numerical attributes of target populations are often unavailable.

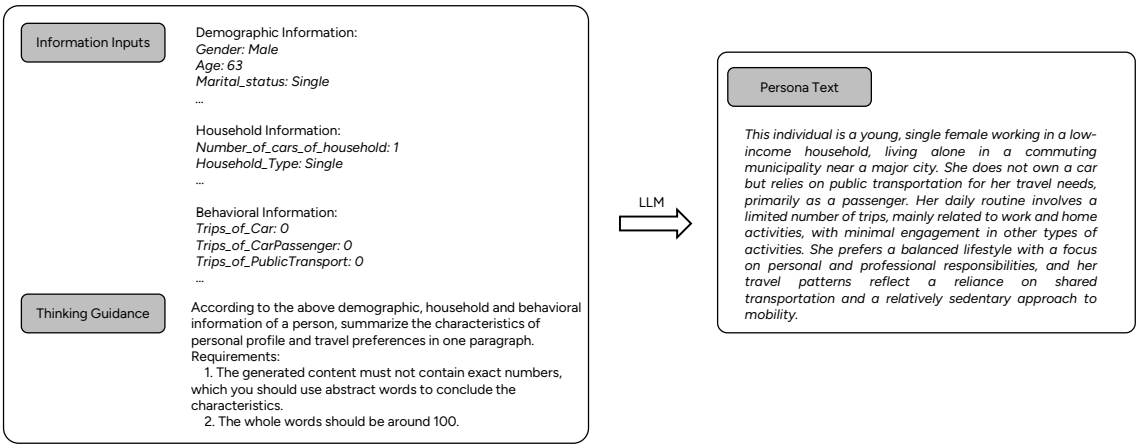


Figure 2: Persona generation prompt and corresponding generated persona text. The left panel shows structured demographic, household, and behavioral inputs with guidance and constraints, while the right panel presents the resulting persona description produced by the LLM.

To enable personas as conditioning signals in the generative modeling pipeline, we employ the same frozen LLM to encode persona text into continuous representations. Using a single frozen LLM for both persona construction and encoding ensures a stable and consistent semantic workflow from abstraction to representation. We extract high-level semantic information from the last L Transformer layers and aggregate them via a layer-wise mean operation. The resulting token-level

representations are then combined using masked mean pooling, followed by LayerNorm, to obtain the final fixed-dimensional persona embedding.
Formally, given a persona description π_i, this encoding process defines a persona embedding function

$$\boldsymbol{e}_i = \mathcal{E}_\pi(\pi_i), \tag{2}$$

where $\mathcal{E}_\pi(\cdot)$ denotes the fixed persona embedding module instantiated by the frozen LLM and the subsequent pooling and normalization operations.

Semantic Conditioning via FiLM Modulation

To incorporate high-level semantic information into the generative process while keeping the tabular data representation unchanged, we employ FiLM as the mechanism for semantic conditioning (Perez et al., 2018). Persona texts associated with individual agents are first encoded into continuous persona embeddings $\boldsymbol{e}$, which are then transformed by a lightweight linear adapter network into a compact conditioning vector $\boldsymbol{c}$,

$$\boldsymbol{c} = f_{\text{adapt}}(\boldsymbol{e}). \tag{3}$$

The conditioning vector is concatenated with the noise input $\boldsymbol{z}$ to form the initial hidden representation,

$$\boldsymbol{h}_0 = [\boldsymbol{z}; \boldsymbol{c}], \qquad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

and is subsequently used to generate feature-wise modulation parameters that apply affine transformations to intermediate feature representations within the generator. Given an intermediate feature representation $\boldsymbol{h}$, FiLM performs feature-wise scaling and shifting as

$$\boldsymbol{h}' = \big(\mathbf{1} + \boldsymbol{\gamma}(\boldsymbol{c})\big) \odot \boldsymbol{h} + \boldsymbol{\beta}(\boldsymbol{c}) \tag{5}$$

where $\boldsymbol{\gamma}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ denote learnable functions mapping the conditioning vector to modulation parameters. This formulation enables semantic information to influence generation in a continuous and differentiable manner. The modulation layers are initialized as identity mappings, ensuring stable training dynamics and allowing semantic effects to emerge progressively during training. By injecting semantic signals through internal feature modulation rather than explicit augmentation of the underlying population feature space, this design enables controllable semantic influence while preserving the original feature space and statistical modeling assumptions of the population synthesis framework.

Marginal Regularization for Statistical Consistency

Ensuring consistency with population-level statistics is essential in transportation applications, where marginal distributions are commonly used for validation and policy analysis. However, generative models trained solely on sample-level objectives may deviate from such aggregate statistics. To address this, we introduce marginal regularization, which aligns generated samples with reference marginal distributions estimated from training data. For each variable, we compute marginal distributions and penalize discrepancies using SRMSE (Müller & Axhausen, 2010; Pritchard & Miller, 2012):

$$\mathcal{L}_{\text{marg}} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \text{SRMSE}(\hat{p}^{(j)}, p^{(j)}). \tag{6}$$

For categorical variables, marginals correspond to category proportions, while for continuous variables we adopt a soft histogram approximation. In addition, we consider a post-hoc *marginal calibration* step to further align generated populations with target distributions, enabling robust statistical consistency under different scenarios.

Training Obejctive of SemaPop-GAN

We adopt a conditional Wasserstein generative adversarial network with gradient penalty (WGAN-GP) as backbone of SemaPop-GAN. The choice of this GAN-based backbone is motivated by its established effectiveness in modeling high-dimensional tabular microdata and its flexibility in incorporating external conditioning signals. In particular, the Wasserstein objective offers a stable training signal and a principled measure of distributional discrepancy, which aligns well with the goal of reproducing population-level statistical properties under explicit marginal constraints.

For conditional discrimination, we adopt the *projection discriminator* paradigm (Miyato & Koyama, 2018), which introduces conditioning through an inner-product projection rather than naive concatenation. In the context of WGAN-GP, this projection mechanism is applied to the critic, yielding a conditioning scheme that is fully compatible with the Wasserstein critic objective. Formally, the critic is expressed as

$$D(\mathbf{x}, \mathbf{e}) = h(\mathbf{x}) + \langle \phi(\mathbf{x}), \psi(\mathbf{e}) \rangle, \tag{7}$$

where $h(\mathbf{x})$ denotes an unconditional scoring function, $\phi(\mathbf{x})$ extracts feature representations of the input sample, and $\psi(\mathbf{e})$ maps the persona embedding to the same latent space. This formulation allows the critic to assess sample quality while explicitly accounting for semantic consistency with the conditioning persona.

Let $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote the latent noise vector and $\hat{\mathbf{x}} = G(\mathbf{z}, \mathbf{e})$ the generated population agent. The critic is trained using a conditional Wasserstein objective with gradient penalty,

$$\mathcal{L}_D = \mathbb{E}_{\mathbf{z}}[D(\hat{\mathbf{x}}, \mathbf{e})] - \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}}[D(\mathbf{x}, \mathbf{e})] + \lambda_{\text{gp}} \, \mathbb{E}_{\tilde{\mathbf{x}}} \left(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}}, \mathbf{e})\|_2 - 1 \right)^2, \tag{8}$$

where $\tilde{\mathbf{x}} = \alpha \mathbf{x} + (1 - \alpha)\hat{\mathbf{x}}$ denotes an interpolated sample used exclusively for gradient penalty computation, with $\alpha \sim \mathcal{U}(0, 1)$ sampled following the standard WGAN-GP formulation. In practice, the critic is updated for n_{critic} steps per generator update to ensure stable Wasserstein distance estimation.

The generator is optimized to both fool the critic and satisfy explicit population-level statistical constraints. Its objective is defined as

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z}}[D(G(\mathbf{z}, \mathbf{e}), \mathbf{e})] + \lambda_m \left(\mathcal{L}_{\text{cont-marg}} + \mathcal{L}_{\text{cat-marg}} \right), \tag{9}$$

where $\mathcal{L}_{\text{cont-marg}}$ and $\mathcal{L}_{\text{cat-marg}}$ denote marginal regularization terms for continuous and categorical variables, respectively, computed via soft histogram-based SRMSE. This composite objective ensures that the generated population remains statistically grounded while allowing semantic persona information to induce controlled distributional shifts.

This formulation enables the model to generate semantically meaningful and statistically consistent population agents, supporting both realistic microdata synthesis and scenario-oriented transportation analysis.

Post-hoc Marginal Calibration

Post-hoc marginal calibration is introduced as an analysis-time procedure to disentangle semantic effects from distributional mismatch in generated populations. Given a fixed set of synthetic agents generated by SemaPop, calibration adjusts only their sample weights—without modifying the generative model or persona conditioning—to match selected target marginal distributions.

For the purpose of post-hoc marginal calibration and evaluation, attributes in $\mathcal{J}_{\text{cont}}$ are discretized into a finite set of bins, such that both categorical and discretized numerical attributes are treated uniformly as categorical variables during calibration.

The calibrated population distribution is represented by the weighted empirical measure

$$\hat{p}_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{N} w_i \, \delta(\mathbf{x} = \hat{\mathbf{x}}_i), \tag{10}$$

where $\delta(\cdot)$ denotes the Dirac delta function. Here, $\mathbf{w} = (w_1, \ldots, w_N)$ denotes the vector of non-negative sample weights learned through post-hoc marginal calibration, representing the relative population mass assigned to each generated agent.

Calibration is performed with respect to a selected set of marginal constraints.

The objective of calibration is to adjust $\mathbf{w}$ such that $\hat{p}_{\mathbf{w}}(k) \approx p^*(k)$ for all constrained attributes and categories, while keeping the generated samples $\{\hat{\mathbf{x}}_i\}$ fixed.

Marginal calibration is implemented via an iterative proportional fitting (IPF), or raking, procedure based on multiplicative weight updates. Starting from uniform weights

$$w_i^{(0)} = \frac{1}{N}, \qquad i = 1, \ldots, N, \tag{11}$$

the algorithm sequentially enforces each marginal constraint through iterative updates.

A multiplicative adjustment ratio is defined as

$$r_j^{(t)}(k) = \frac{p_j^*(k) + \varepsilon}{\hat{p}_{\mathbf{w},j}^{(t)}(k) + \varepsilon}, \tag{12}$$

where $\varepsilon > 0$ is a small constant for numerical stability.
The weights are then updated according to

$$w_i^{(t+1)} = w_i^{(t)} \cdot \left[r_j^{(t)} \left(\hat{x}_i^{(j)} \right) \right]^{\alpha}, \tag{13}$$

where $\alpha \in (0, 1]$ is a damping factor controlling the strength of the update. After each update, the weights are renormalized to satisfy $\sum_{i=1}^{N} w_i^{(t+1)} = 1$.
When multiple marginal constraints are specified, the above update is applied sequentially for each constrained attribute j within a single iteration. This process is repeated for a fixed number of iterations or until convergence.

4 Experiments

Data Description

All experiments are conducted on a nationwide synthetic population dataset of Sweden, publicly available via Mendeley Data.[1] The dataset comprises a microscopically simulated population of over 10 million individuals and is designed for activity-based transport modeling and large-scale population simulation. The synthetic agents are calibrated to match official Swedish demographic and travel statistics for 2018, ensuring consistency between micro-level representations and observed aggregate patterns Tozluoğlu et al. (2023).
The data are organized as a relational database with three linked tables—*Person*, *Household*, and *Activity-travel*—capturing individual socio-demographic attributes, household context, and daily activity–travel behavior. Detailed descriptions of the attributes and their semantic groupings can be found in Qin et al. (2026).
For experiments, we construct a municipality-level stratified dataset. Sweden consists of 290 municipalities, from each of which we randomly sample 20% of agents. This proportional strategy preserves spatial heterogeneity while reducing computational cost. The resulting dataset is split into training (50,873), validation (20,262), and test (2,040,650) sets. The intentionally imbalanced split reflects a realistic setting in which models are trained on limited microdata and deployed to generate large-scale populations.

Evaluation Metrics

We assess the quality of population synthesis from two complementary perspectives: distributional fidelity and structural validity. To measure how well the generated population aligns with the reference (held-out test) data, we adopt the Standardized Root Mean Square Error at both the marginal level (SRMSE-M) and the bivariate marginal level (SRMSE-B) Garrido et al. (2020). To further evaluate the structural realism and diversity of the generated samples, we report Precision, Recall, and F1 scores Kim & Bansal (2023). In post-hoc marginal calibration analysis, we use the effective sample size (ESS) as a diagnostic metric to quantify weight concentration and the associated loss of effective information after reweighting (Deville & Särndal, 1992; Biemer & Christ, 2012). Implementation details are consistent with Qin et al. (2026).

Generative Performance Comparison

We first evaluate the generative performance of *SemaPop* against a set of representative baselines, focusing on their ability to reproduce both population-level statistics and individual-level joint distributions under a unified experimental protocol. The comparison includes representative baselines from multiple modeling paradigms, including GAN-based methods (**CTGAN** (Xu et al., 2019), **WGAN-GP** (Gulrajani et al., 2017), **WGAN-GP-ZCR** (Kim & Bansal, 2023)), diffusion and variational approaches (**TabDDPM** (Kotelnikov et al., 2023), **TVAE** (Xu et al., 2019)), and probabilistic graphical models (**BN** (Sun & Erath, 2015), **BN-Copula** (Jutras-Dubé et al., 2024)).
Table 1 summarizes the results in terms of marginal alignment, bivariate consistency, and feasibility–diversity metrics. Across models, SRMSE-M and SRMSE-B exhibit limited variation, suggesting that marginal-level agreement alone is insufficient to differentiate generative quality. In contrast, precision and recall reveal substantial differences. Classical probabilistic models (BN, BN-Copula)

[1] `https://data.mendeley.com/datasets/9n29p7rmn5/2`

Table 1: Generative performance comparison between the SemaPop and baseline models.

Model	SRMSE-M	SRMASE-B	Precision	Recall	F1
CTGAN	0.0119	0.0678	3.83	70.56	7.26
TabDDPM	0.0118	0.0679	18.71	91.54	31.07
TVAE	0.0119	0.0672	6.74	79.32	12.43
BN	0.0119	0.0679	4.57	79.60	8.64
BN-Copula	0.0119	0.0679	4.66	79.26	8.81
WGAN-GP	0.0111	0.0607	52.01	93.08	66.74
WGAN-GP-ZCR	0.0109	0.0593	51.88	93.33	66.69
SemaPop-GAN	**0.0104**	**0.0554**	**73.95**	**93.39**	**82.54**

and non-adversarial neural models (CTGAN, TVAE, TabDDPM) achieve high recall but consistently low precision (generally below 20%), indicating that while they cover many valid patterns, they also produce a large number of infeasible samples, leading to low F1 scores. GAN-based models markedly improve this trade-off: both WGAN-GP and WGAN-GP-ZCR achieve significantly higher precision (above 50%) while maintaining high recall, resulting in strong F1 performance. Overall, SemaPop-GAN achieves the best performance across all metrics, attaining the lowest SRMSE-M and SRMSE-B, as well as the highest precision, recall, and F1. Compared to the strongest baselines, it delivers substantial gains in precision and F1 while maintaining comparable recall, demonstrating a superior balance between feasibility and coverage.

Post-hoc Marginal Calibration

To assess the robustness of persona-conditioned generation under explicit distributional constraints, we conduct a post-hoc marginal calibration experiment. The objective is to determine whether population-level deviations induced by semantic conditioning stem from marginal mismatch, and to characterize the trade-off between marginal fidelity, joint realism, and population diversity under increasing constraint strength.

In this experiment, given a fully specified set of target marginal distributions, generated samples are fixed and reweighted at analysis time using a raking-based calibration procedure. Based on the hold-out test set, marginal constraints are imposed on key attributes capturing demographic structure (*Age*), socioeconomic status (*Income class*), household composition (*Household Type*), motorization level (*Number of cars of household*), and public transport usage intensity (*Trips of PublicTransport*). Calibration strength is varied across five levels (L0–L4), corresponding to 0, 5, 10, 20, and 40 raking iterations. We evaluate two model variants under identical calibration levels: SemaPop-GAN trained with marginal regularization and an otherwise identical model without it. All calibration experiments are conducted on synthetic population microdata. Target marginals are derived from the same synthetic reference population or constructed as hypothetical constraints for controlled analysis, rather than from external real-world statistics. Accordingly, post-hoc calibration is used strictly as an analytical tool, not as a mechanism for fitting the model to ground-truth marginals.

Figure 3 reports marginal consistency (SRMSE-M), bivariate joint consistency (SRMSE-B), and effective sample size (ESS) across calibration levels. Both models benefit from mild calibration (L1), but their behavior diverges as the calibration constraints strengthen. SemaPop-GAN with marginal regularization maintains stable marginal consistency and bivariate joint consistency from L1 to L4, indicating robust preservation of local joint-distribution structure under increasing constraints. In contrast, the unregularized model exhibits progressively degraded marginal consistency and bivariate joint consistency as calibration strength increases. Moreover, although ESS decreases for both models as expected, the regularized model consistently retains a higher effective sample size, indicating less severe weight concentration and improved robustness of the calibrated population. These results suggest that marginal regularization during training not only improves baseline marginal alignment, but also produces synthetic populations that better accommodate subsequent distributional constraints, preserving both higher-order structure and population diversity under semantic conditioning.

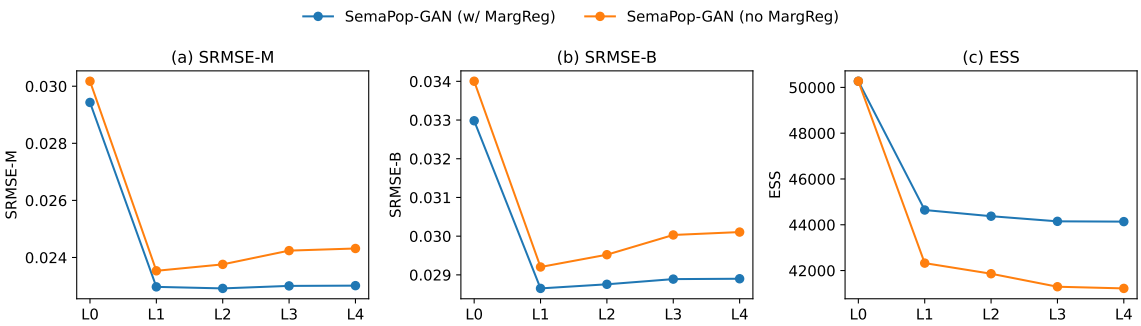


Figure 3: Trade-offs induced by post-hoc marginal calibration under increasing calibration strength (L0-L4). Panels report (a) marginal consistency (SRMSE-M), (b) joint realism (SRMSE-B), and (c) effective sample size (ESS). Results are shown for SemaPop-GAN trained with marginal regularization (*w/ MargReg*) and an otherwise identical model trained without marginal regularization (*no MargReg*).

5 Conclusions

This study presents SemaPop-GAN, a persona-conditioned population synthesis framework that integrates LLM-derived semantic representations into generative modeling. By introducing personas as a high-level abstraction, the proposed approach enables controllable and interpretable population generation while preserving statistical consistency through marginal regularization during training. Experimental results demonstrate that SemaPop-GAN improves both distributional fidelity and structural realism compared to strong baselines. In addition, post-hoc marginal calibration is used as an analysis tool to examine the robustness of the generated populations under explicit distributional constraints, showing that the proposed model better preserves joint structure and population diversity under increasing calibration strength. These results highlight the potential of semantic-conditioned generative modeling for advancing scenario-driven population synthesis and supporting more flexible, policy-aware transportation analysis.

References

Aemmer, Z., & MacKenzie, D. (2022). Generative population synthesis for joint household and individual characteristics. *Computers, Environment and Urban Systems*, *96*, 101852.

Anable, J. (2005). 'complacent car addicts' or 'aspiring environmentalists'? identifying travel behaviour segments using attitude theory. *Transport policy*, *12*(1), 65–78.

Barthelemy, J., & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, *47*(2), 266–279.

Batty, M. (2013). *The new science of cities*. MIT press.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*(6), 415–429.

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., & Polydoropoulou, A. (2002). Integration of choice and latent variable models. *Perpetual motion: Travel behaviour research opportunities and application challenges*, *2002*, 431–470.

Biemer, P. P., & Christ, S. L. (2012). Weighting survey data. In *International handbook of survey methodology* (pp. 317–341). Routledge.

Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, *99*(suppl_3), 7280–7287.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Cooper, A. (1999). The inmates are running the asylum. In *Software-ergonomie'99: design von informationswelten* (pp. 17–17). Springer.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376–382.

Garrido, S., Borysov, S. S., Pereira, F. C., & Rich, J. (2020). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, *120*, 102787.

Gilbert, N., & Troitzsch, K. (2005). *Simulation for the social scientist*. McGraw-Hill Education (UK).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, *30*.

Helbing, D. (2009). Managing complexity in socio-economic systems. *European Review*, *17*(2), 423–438.

Hermes, K., & Poulsen, M. (2012). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, *36*(4), 281–290.

Jeong, B., Lee, W., Kim, D.-S., & Shin, H. (2016). Copula-based approach to synthetic population generation. *PloS one*, *11*(8), e0159496.

Jutras-Dubé, P., Al-Khasawneh, M. B., Yang, Z., Bas, J., Bastin, F., & Cirillo, C. (2024). Copula-based transferable models for synthetic population generation. *Transportation Research Part C: Emerging Technologies*, *169*, 104830.

Kim, E.-J., & Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C: Emerging Technologies*, *148*, 104053.

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning* (pp. 17564–17579).

Miller, E. J. (2018). Agent-based activity/travel microsimulation: what's next? In *The practice of spatial analysis: Essays in memory of professor pavlos kanaroglou* (pp. 119–150). Springer.

Miyato, T., & Koyama, M. (2018). cgans with projection discriminator. In *International conference on learning representations (iclr)*.

Müller, K., & Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. *Arbeitsberichte Verkehrs-und Raumplanung*, *638*.

Norman, P. (1999). Putting iterative proportional fitting on the researcher's desk; 1999. *School of Geography, University of Leeds*.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685–704.

Pruitt, J., & Grudin, J. (2003). Personas: practice and theory. In *Proceedings of the 2003 conference on designing for user experiences* (pp. 1–15).

Publishing, O. (2017). *Systems approaches to public sector challenges-working with change*. OECD Publishing.

Qin, Z., Ling, Y., Wang, L., Pereira, F. C., & Ma, Z. (2026). Semapop: Semantic-persona conditioned population synthesis. *arXiv preprint arXiv:2602.11569*.

Sun, L., & Erath, A. (2015). A bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, *61*, 49–62.

Tozluoğlu, Ç., Dhamal, S., Yeh, S., Sprei, F., Liao, Y., Marathe, M., ... Dubhashi, D. (2023). A synthetic population of sweden: datasets of agents, households, and activity-travel patterns. *Data in Brief*, *48*, 109209.

Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical social sciences*, *43*(3), 303–343.

Wang, L., Duan, P., He, Z., Lyu, C., Chen, X., Zheng, N., ... Ma, Z. (2025). Agentic large language models for day-to-day route choices. *Transportation Research Part C: Emerging Technologies*, *180*, 105307.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*.