

Scalar-mediated inelastic dark matter as a solution to small-scale structure anomalies

Zihan Wang^{*}

Department of Physics, University of Oxford, Keble Road, Oxford, OX1 3PU, United Kingdom



(Received 23 December 2025; accepted 16 April 2026; published 11 May 2026)

We propose a novel self-interacting dark matter (SIDM) model mediated by a light leptophilic scalar boson to solve long-standing small-scale structure problems within the Lambda cold dark matter (Λ CDM) framework. Small-scale anomalies such as the core-cusp problem challenge the traditional CDM framework. Therefore, we introduce a scalar-mediated SIDM model as an alternative to CDM that naturally addresses these problems. It also exhibits p -wave-suppressed annihilation and avoids constraints from the cosmic microwave background. In our model, we assume pseudo-Dirac dark matter with a small mass splitting of order 10^2 eV to ensure kinematic scattering suppression in satellites. We also introduce a dimension-five transition magnetic dipole operator to satisfy big bang nucleosynthesis requirements. It allows the decay $\chi_2 \rightarrow \chi_1 \gamma$, so the excited-state abundance is primarily depleted by exothermic self-interactions ($\chi_2 \chi_2 \rightarrow \chi_1 \chi_1$) in the early Universe. This ensures that the excited-state population is negligible for structure formation at the dipole scale $\Lambda_{\text{eff}} \gtrsim 10^7$ GeV. The model is made inelastic by a dark discrete $\mathbb{Z}_2$ symmetry that prevents tree-level elastic scattering. Treating the scattering dynamics nonperturbatively in a Schrödinger framework yields a relatively small residual resonant benchmark window ($m_\chi \approx 40$ GeV, $\Delta m \approx 100$ eV, $m_\phi \approx 20$ MeV). In such a system, the kinematic threshold minimizes interactions in ultrafaint Milky Way satellites, while resonant effects produce large cross sections, $\sigma/m \sim \mathcal{O}(10)$ cm²/g, in field dwarf galaxies. Current direct-detection prospects are weak, and the leptophilic scalar mediator keeps the model away from existing nuclear recoil constraints. However, the low-threshold dipole operator provides a potential discovery channel for future xenon-based experiments.

DOI: [10.1103/lcbj-grg8](https://doi.org/10.1103/lcbj-grg8)

I. INTRODUCTION

The classical cosmological model Lambda cold dark matter (Λ CDM) has been successful in explaining the large-scale structure of the Universe [1]. Its predictions agree well with observations ranging from temperature anisotropies in the cosmic microwave background (CMB) to galaxy cluster distributions. However, on subgalactic scales, with distances $\lesssim 10$ kpc and halo masses $\lesssim 10^{11} M_\odot$, significant deviations from collisionless N -body simulations and astrophysical observations appear [2]. Classified as “small-scale crises,” these deviations suggest that the dark sector may depart from the collisionless assumption [3]. The anomalies include (i) the core-cusp problem: CDM simulations predict dark matter halos with high central-density cusps ($\rho \propto r^{-1}$) [4], while rotation curves

in dwarf and low surface brightness (LSB) galaxies favor low-density cores ($\rho \propto r^0$) [5–7]; (ii) the diversity problem: galaxies with similar maximum circular velocities (V_{max}), corresponding to similar halo masses, show large scatter in their inner rotation curves ($V_{2 \text{ kpc}}$) [8]; (iii) the too-big-to-fail problem: simulations predict a large population of massive subhalos in Milky-Way-like systems that is not observed [9,10]. Some subhalos may be dark due to baryonic feedback, and the dynamical properties of the most massive subhalos differ from the observed kinematics of the brightest satellite galaxies [11,12].

Self-Interacting dark matter (SIDM) [13,14] is one promising approach to these issues. If dark matter particles scatter with a large cross section ($\sigma/m \sim 1$ cm²/g), heat transfer from the dynamically hot outer halo to the colder interior can thermalize the central region, reduce central density, and transform cusps into cores [15,16]. SIDM interaction rates and halo concentration dynamics can also explain the diversity of rotation curves [17,18]. However, a robust SIDM model must satisfy velocity-dependent constraints. Observations of galaxy clusters, including halo shapes and mergers such as the bullet cluster, limit the high-velocity cross section ($v \sim 10^3$ km/s) to $\sigma/m \lesssim 0.6$ cm²/g

^{*}Contact author: zihan.wang@queens.ox.ac.uk

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

[19–22]. For dwarf galaxies ($\sim 30\text{--}60$ km/s), values around $\sigma/m \sim 1\text{--}10$ cm²/g are needed to significantly modify density profiles [23]. Ultrafaint Milky Way satellites (e.g., Draco, 10–20 km/s) provide some of the strongest upper limits on self-interaction [24]. Excessive scattering in these systems may trigger the Nishikawa instability, runaway gravothermal collapse, and central densities much larger than those initially studied in [25]. Taken together, observations require an effective model that is active at intermediate velocities (e.g., dwarfs) but suppressed at high (clusters) and very low (satellites) velocities. Light mediators with a Yukawa potential can provide cluster-scale suppression while exhibiting Rutherford-like behavior, $\sigma \propto v^{-4}$ [26].

However, as we can see with only a simple elastic cross section, there are problems when it comes to getting the satellite and dwarf requirements to match. Inelastic scattering is therefore an interesting solution. It sets a threshold of kinematics: The endothermic channel ($\chi_1\chi_1 \rightarrow \chi_2\chi_2$) is kinematically forbidden by the mass splitting Δm below the endothermic velocity $v_{\text{th}} \sim \sqrt{8\Delta m/m_\chi}$. This threshold acts as a switch that turns off interactions in the coldest halos [27,28] (the coupling must be chosen to reside in an antiresonance valley of the residual elastic cross section to ensure this kinematic switch remains effective, see Appendix A). The idea mainly applies to vector mediator known as dark photons [29]. However, vector mediators create unavoidable tension with cosmological variables. The thermal annihilation of vector-coupled dark matter are s -wave, as energy is injected into the plasma in the recombination epoch and warps the CMB [1,30,31]. To escape these constraints, we usually need to fine-tune the mass splittings or construct asymmetric dark matter models [32,33]. Therefore, we present a scalar-mediated inelastic dark matter model in this work. The answer is indeed enough to satisfy CMB constraints since we can plug the vector mediator into a scalar, which is already a p -wave suppressed annihilation without any fine-tuning. To prevent tree-level elastic scattering, we apply a simple $\mathbb{Z}_2$ symmetry in the dark sector. Moreover, to prevent violating any rules of big bang nucleosynthesis (BBN), we present a leptophilic coupling structure to a scalar mediator.

The paper is organized as follows. The theoretical framework and dipole mechanism are discussed in Sec. II. Then, we discuss the nonperturbative scattering phenomenology in Sec. III. Section IV describes cosmological consistency. Section V provides numerical results. Also, we constrained our model with direct search experiments in Sec. VI, and pointed out future detection possibilities. We then assess the robustness of the benchmark against parameter variations in Sec. VII. Finally, we discuss and conclude this paper in Sec. VIII. In addition, for detailed derivations of the coupled-channel scattering formalism, please refer to Appendix A.

II. THEORETICAL FRAMEWORK

We predict a dark sector that is outside the Standard Model (SM), interacting with SM weakly via necessary portal effects. The particles in this sector consist of Dirac fermion χ and light real scalar boson ϕ , giving a kinetic term:

$$\mathcal{L}_{\text{kinetic}} = \bar{\chi}(i\partial - m_\chi)\chi - \frac{\Delta m}{2}(\bar{\chi}^c\chi + \text{H.c.}), \quad (1)$$

where $\Delta m \ll m_\chi$. We predict that the Dirac fermion χ acquires a Majorana mass term, splitting into two separate mass eigenstates:

$$\mathcal{L}_{\text{kinetic}} = \bar{\chi}_1(i\partial - m_1)\chi_1 + \bar{\chi}_2(i\partial - m_2)\chi_2. \quad (2)$$

In this way, the matrix is diagonalized and two Majorana fermions χ_1 and χ_2 are of masses

$$m_1 = m_\chi - \frac{\Delta m}{2}, \quad m_2 = m_\chi + \frac{\Delta m}{2}, \quad (3)$$

respectively. We regard χ_1 as the stable dark matter candidate and χ_2 as the heavier excited state.

A. Symmetry and self-interaction

Suppression of elastic self-scattering ($\chi_1\chi_1 \rightarrow \chi_1\chi_1$) at low velocities is a very powerful phenomenological necessity when addressing the tension between satellite survival and dwarf galaxy cores [34]. If elastic and inelastic scatterings are equivalent, v^{-4} cross section in satellites would be expected to increase, which violates the Draco constraint [22].

To block elastic scattering, we impose a dark $\mathbb{Z}_2$ parity. The fields then undergo the following transformations:

$$\chi_1 \rightarrow +\chi_1 \quad (\text{even}), \quad (4)$$

$$\chi_2 \rightarrow -\chi_2 \quad (\text{odd}), \quad (5)$$

$$\phi \rightarrow -\phi \quad (\text{odd}). \quad (6)$$

Under this antisymmetric symmetry, a general renormalizable Yukawa-type interaction between the scalar and fermions is strictly off-diagonal:

$$\mathcal{L}_{\text{int}} = -g_S\phi(\bar{\chi}_1\chi_2 + \bar{\chi}_2\chi_1). \quad (7)$$

Note that while the $\mathbb{Z}_2$ symmetry forbids the tree-level elastic vertex $\bar{\chi}_1\chi_1\phi$, residual elastic scattering is induced at one loop via box diagrams involving two scalar exchanges. However, these contributions are negligible, as demonstrated in Appendix A, and therefore do not violate stringent self-interaction bounds from satellite galaxies [22].

This Lagrangian guarantees that the vertices $\chi_1\chi_1\phi$ and $\chi_2\chi_2\phi$ vanish identically. For $\chi_1\chi_2 \rightarrow \chi_1\chi_2$, our derivation in Sec. IV shows that the abundance of the excited state χ_2 is effectively small, making tree-level elastic scattering negligible. We will see that tree-level $\chi_1\chi_1 \rightarrow \chi_2\chi_2$ scattering (or $\chi_1\chi_2 \rightarrow \chi_2\chi_1$) provides the desired phenomenology, while other processes are highly suppressed, if not forbidden. While we treat the strict off-diagonal nature of the scalar interaction as an effective low-energy feature, its UV-complete origin requires careful construction. If the 20 MeV scalar mediator ϕ were simply the CP -even radial mode of a dark Higgs field that spontaneously breaks a dark $U(1)_D$ gauge symmetry, its couplings to the mass eigenstates would be strictly diagonal ($\phi\bar{\chi}_1\chi_1$), as demonstrated in [35]. A light radial mode would therefore induce unsuppressed tree-level elastic scattering, violating the Draco constraints. Furthermore, the coupling of such a radial mode is fixed by $g_S \propto \Delta m/m_\phi$, which for our benchmark parameters would require a wildly nonperturbative quartic coupling $\lambda_\phi \sim 10^{10}$. To rigorously forbid these diagonal terms and decouple the interaction strength from the mass splitting, we separate the symmetry-breaking mechanism from the light SIDM mediator. Following the framework outlined in [36], we postulate a dark sector containing a heavy dark Higgs Φ_H and a separate light real scalar singlet ϕ . The $U(1)_D$ gauge symmetry is broken at a high scale by $\langle\Phi_H\rangle$, generating the small Majorana mass splitting $\Delta m \approx 100$ eV. The dangerous diagonal elastic scattering mediated by the radial mode of Φ_H is heavily suppressed by its large mass ($m_{\Phi_H} \gg m_\chi$). The light mediator ϕ , which does not acquire a vacuum expectation value, is instead governed by a dark charge-conjugation symmetry C_χ under which the dark matter gauge eigenstates transform as $\chi \leftrightarrow \chi^c$. The mass eigenstates therefore transform with opposite C_χ parities: $\chi_1 \rightarrow -\chi_1$ (odd) and $\chi_2 \rightarrow +\chi_2$ (even). By assigning ϕ to be explicitly odd under this symmetry ($\phi \rightarrow -\phi$), the diagonal interactions $\phi\bar{\chi}_1\chi_1$ and $\phi\bar{\chi}_2\chi_2$ are strictly forbidden. The parity-even Lagrangian then uniquely permits the interaction $\mathcal{L}_{\text{int}} = -g_S\phi(\bar{\chi}_1\chi_2 + \text{H.c.})$. Because ϕ is independent of the symmetry-breaking VEV, its coupling g_S is a free parameter, smoothly accommodating the strongly coupled resonant regime required for SIDM. This symmetry structure protects the kinematic-switch mechanism from dangerous diagonal corrections at tree level.

B. SM portal and decay of scalar mediator

In the early Universe, the dark sector was in thermal equilibrium with the plasma formed by SM particles. We presume the scalar ϕ is leptophilic. It couples to SM leptons ($l = e, \mu, \tau$) but not quarks or the Higgs boson. The effective Lagrangian for the portal is

$$\mathcal{L}_{\text{portal}} = \sum_l g_l \phi \bar{l} l, \quad (8)$$

where g_l is the coupling constant.

We treat the dark $\mathbb{Z}_2$ parity as an approximate symmetry of the dark sector. While the dark sector fields transform nontrivially under this parity ($\chi_2 \rightarrow -\chi_2, \phi \rightarrow -\phi$), the Standard Model leptons are $\mathbb{Z}_2$ even. The portal interaction in Eq. (8) constitutes a source of explicit $\mathbb{Z}_2$ symmetry breaking. However, since the portal coupling is extremely weak ($g_e \sim 10^{-6}$) to satisfy beam-dump and stellar constraints, this breaking is parametrically small. We quantify the induced diagonal couplings and their impact on the stability of the kinematic switch in the following section.

The scalar mediator ϕ decays preferentially into electron-positron pairs (e^+e^-) [37]. When the mass of scalar mediator ranges from 10 to 100 MeV, the decay width is

$$\Gamma(\phi \rightarrow e^+e^-) = \frac{g_e^2}{8\pi} m_\phi \left(1 - \frac{4m_e^2}{m_\phi^2}\right)^{3/2}. \quad (9)$$

This leptophilic composition allows the mediator to rapidly decay prior to BBN by $\phi \rightarrow e^+e^-$ and falling within cosmological limits while simultaneously suppressing coupling to nucleons to avoid the limits introduced by direct search experiments [38–40], it gave m_χ from 10^{-2} to 10^2 GeV.

C. The electron-specific portal and UV realization

Throughout this work, we treat the electron portal coupling constant as a scan parameter in the range

$$g_e \in [10^{-7}, 10^{-5}], \quad (10)$$

chosen to ensure mediator decay well before BBN. Also, it enables a conservative discussion of stellar/supernova constraints without committing to a single finely tuned value. Unless stated otherwise, numerical examples use reference value $g_e = 10^{-6}$. In contrast, the dipole operator introduced in Sec. IID is controlled by an independent effective scale Λ_{eff} generated at loop level in a separate heavy charged sector and is not fixed by g_e .

Considering contributions to $(g-2)_\mu$ or $(g-2)_\tau$ factors, if the scalar mediator couples to the particles, the $(g-2)$ value will significantly deviate from SM prediction. Therefore, we make the assumption that ϕ only couples to electrons. We then construct a flavor-specific UV completion.

Rather than a type-IV two-Higgs doublet model (2HDM) [41], we foresee the heavy vectorlike leptons E should couple to both ϕ and e to make the coupling flavor specific. Our effective Lagrangian, below the mass scale M_E , is

$$\mathcal{L}_{\text{portal}} = g_e \phi \bar{e} e, \quad \text{with } g_e \sim \lambda_\phi \frac{v_{\text{EW}}}{M_E} \sin \theta_e. \quad (11)$$

We will discuss in Sec. IID that the transition magnetic dipole operator $\bar{\chi}_1 \sigma^{\mu\nu} \chi_2 F_{\mu\nu} / \Lambda$ is required to have a very

high Λ ($\sim 10^4$ GeV) in order to facilitate excited state decay. We note that generating this operator via loops involving the same weak portal coupling $g_e \approx 10^{-6}$ would imply a effective $\Lambda_{\text{eff}} \sim (16\pi^2 M)/(eg_{Sg_e}) \gg 10^4$ GeV, which is phenomenologically insufficient. We therefore postulate a split-sector UV completion in which the portal coupling g_e and the dipole operator are generated by different heavy states.

We parametrize the dipole operator as an EFT generated at one loop by electrically charged heavy states Ψ of mass

$$M_\Psi = 10^4 \text{ GeV}. \quad (12)$$

At energies $\mu \ll M_\Psi$, integrating out Ψ yields the transition dipole operator with a loop-suppressed coefficient,

$$\mathcal{L}_{\text{dipole}} \supset \frac{1}{\Lambda_{\text{eff}}} \bar{\chi}_1 \sigma^{\mu\nu} \chi_2 F_{\mu\nu} + \text{H.c.}, \quad \frac{1}{\Lambda_{\text{eff}}} \simeq \frac{e\lambda_\Psi}{16\pi^2 M_\Psi} \times \mathcal{O}(1), \quad (13)$$

where λ_Ψ denotes a representative product of dark-sector couplings in the UV completion and $\mathcal{O}(1)$ encodes loop-function dependence on mediator mass ratios and chiral structure. For $\lambda_\Psi = \mathcal{O}(1)$ and $M_\Psi = 10^4$ GeV, this gives parametrically $\Lambda_{\text{eff}} \sim \text{few} \times 10^6$ GeV. Throughout the phenomenological analysis, we treat Λ_{eff} as the effective suppression scale controlling both $\chi_2 \rightarrow \chi_1 \gamma$ and dipole-mediated scattering rates. To generate the effective dipole scale $\Lambda_{\text{eff}} \sim 10^7$ GeV, we introduce a heavy sector of vectorlike leptons $\Psi^\pm$ with mass $M_\Psi \sim 10\text{--}50$ TeV. These states carry dark charges and couple via $\mathcal{L}_{\text{UV}} \supset y\phi\bar{\Psi}\Psi + \text{H.c.}$ Integrating out Ψ at one-loop generates the dipole operator with coefficient $\Lambda_{\text{eff}}^{-1} \sim ey^2/(16\pi^2 M_\Psi)$. This naturally accommodates the PeV-scale suppression.

D. Dark dipole and excited state decay

For our benchmark parameters ($\Delta m \sim 100$ eV, $m_\phi \sim 20$ MeV), the decay $\chi_2 \rightarrow \chi_1 \phi$ is kinematically forbidden $\Delta m \ll m_\phi$, so the excited state cannot deexcite by emitting an on-shell mediator. We therefore include a dimension-five transition electromagnetic operator that permits radiative deexcitation, $\chi_2 \rightarrow \chi_1 \gamma$. Importantly, the mediator lifetime relevant for BBN is controlled by the leptophilic portal through $\phi \rightarrow e^+ e^-$, which is prompt for our benchmark g_e . The dipole operator instead controls the late-time fate of the excited state and the inelastic photon-mediated scattering phenomenology discussed in Sec. III. To stop the excited state χ_2 from contributing too much to the relic and to allow stable decay to the SM states we introduce a dimension-five transition magnetic dipole operator [42], which links the dark sector to the photon field strength tensor $F_{\mu\nu}$:

$$\mathcal{L}_{\text{dipole}} = \frac{1}{\Lambda_{\text{eff}}} \bar{\chi}_1 \sigma^{\mu\nu} \chi_2 F_{\mu\nu} + \text{H.c.} \quad (14)$$

Introducing this operator allows the decay $\chi_2 \rightarrow \chi_1 + \gamma$ to occur rapidly, giving a decay rate:

$$\Gamma(\chi_2 \rightarrow \chi_1 \gamma) \simeq \frac{1}{\pi} \frac{\Delta m^3}{\Lambda_{\text{eff}}^2}. \quad (15)$$

For our benchmark values Λ_{eff} and $\Delta m \approx 100$ eV, the excited-state lifetime is $\tau_{\chi_2} = 1/\Gamma(\chi_2 \rightarrow \chi_1 \gamma)$. For the direct-detection-safe parameter space of interest we typically have $\Lambda_{\text{eff}} \gtrsim 10^7$ GeV, so χ_2 can be long-lived. This does not endanger BBN [43] because the radiated energy per decay is only $\Delta m \ll \text{MeV}$, far below nuclear photodissociation thresholds; the relevant late-time limits are instead from electromagnetic energy injection such as spectral distortions [42] and ionization history, which are strongly suppressed by the small fractional energy release $\Delta m/m_\chi \ll 1$.

III. SCATTERING PHENOMENOLOGY

To solve the small-scale structure problems ($m_\chi \sim 10\text{--}100$ GeV, $m_\phi \sim 10$ MeV), the coupling strength $\alpha \equiv g_S^2/4\pi$ is typically large enough that the perturbative Born approximation fails. This failure occurs when the potential is strong enough to support bound states or resonances, formally when $\alpha m_\chi/m_\phi \gtrsim 1$. Our theory is in the non-perturbative “resonant” regime, where the cross section exhibits a rich velocity dependence defined as Sommerfeld enhancement [29].

A. Nonperturbative dynamics

To compute the cross section $\frac{\sigma_{T1 \rightarrow 2}}{m}$ of inelastic scattering $\chi_1 \chi_1 \rightarrow \chi_2 \chi_2$, we solve the nonrelativistic coupled-channel Schrödinger equation (see Appendix A). The system involves two states: the incoming $\chi_1 \chi_1$ pair and the excited $\chi_2 \chi_2$ pair. The off-diagonal interaction in Eq. (7) induces transitions between these states mediated by the Yukawa potential:

$$V(r) = -\alpha \frac{e^{-m_\phi r}}{r}. \quad (16)$$

The inelastic channel $\chi_1 \chi_1 \rightarrow \chi_2 \chi_2$ only opens when the kinetic energy exceeds the mass splitting threshold, $E_{\text{kin}} > 2\Delta m$. Below this kinematic threshold, the dominant heat-transferring inelastic channel is closed. We note that the elastic scattering cross section for $\chi_1 \chi_1 \rightarrow \chi_1 \chi_1$ (S3 in Fig. 1) does not vanish exactly: although it is forbidden at tree level by the $\mathbb{Z}_2$ symmetry, it is induced by virtual transitions through the closed $\chi_2 \chi_2$ channel [equivalently, by box-type contributions at $\mathcal{O}(\alpha^4)$ in the perturbative regime]. In the resonant regime, however, the Born expansion can fail and the below-threshold elastic amplitude can be

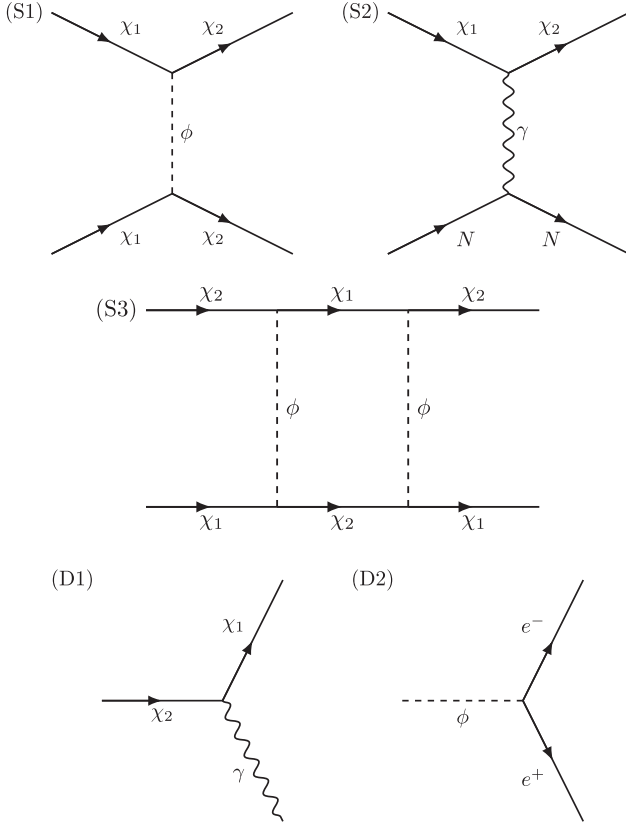


FIG. 1. Feynman diagrams presenting some of the scatterings and decay processes to investigate: $\chi_1\chi_1 \rightarrow \chi_2\chi_2$, the contributor to the inelastic scattering phenomenology (S1); $\chi_1N \rightarrow \chi_2N$, potential nuclear recoil in direct search experiments (S2); the possible elastic scattering below threshold (S3); $\chi_2 \rightarrow \chi_1 + \gamma$, the decay from excited state DM to stable state (D1); and $\phi \rightarrow e^+ + e^-$, decay of mediators to positron and electron pairs (D2).

nonperturbatively enhanced by virtual closed-channel effects [44]. We therefore evaluate $\sigma(\chi_1\chi_1 \rightarrow \chi_1\chi_1)$ nonperturbatively by solving the coupled-channel Schrödinger equation with an exponentially decaying boundary condition in the closed channel (Appendix A.5), and select benchmarks that lie in an antiresonance valley. For our benchmark, we obtain $\sigma/m_\chi \lesssim 0.01 \text{ cm}^2/\text{g}$ at satellite velocities, safely below the Draco bound $\sigma/m_\chi \lesssim 0.1 \text{ cm}^2/\text{g}$, thereby realizing the kinematic switch.

B. Resonant enhancement

Above the kinematic threshold, the cross section does not simply rise monotonically. It is heavily influenced by the presence of “darkonium” bound states in the spectrum [45]. When the set of parameters $\{m_\chi, m_\phi, \alpha\}$ is tuned such that a bound state exists near zero energy, the scattering cross section undergoes a resonant enhancement (the Sommerfeld effect). This resonance manifest as a sharp peak in the cross section just above the threshold velocity [46].

For our benchmark parameter ($m_\chi \approx 40 \text{ GeV}$), we reside near a resonance. This leads to the following behavior:

- (a) Subthreshold regime ($v < v_{\text{th}}$): In ultrafaint satellites ($v \sim 20 \text{ km/s}$), the kinetic energy is insufficient to create the heavier χ_2 state [47]. Inelastic scattering is forbidden, and residual elastic scattering is negligible, preventing core collapse.
- (b) Resonant regime ($v > v_{\text{th}}$): The channel opens in dwarf galaxies ($v \sim 60 \text{ km/s}$). The proximity to the resonance pole amplifies the cross section by orders of magnitude compared to the geometric size of the target, reaching $\sigma/m \sim 10 \text{ cm}^2/\text{g}$. This is the mechanism that generates the large cores observed in LSB galaxies [48].
- (c) Classical regime ($v \gg v_{\text{th}}$): At cluster velocities ($v \sim 10^3 \text{ km/s}$), the kinetic energy is much larger than the binding energy of the resonance [26]. The Sommerfeld enhancement is negligible [29], and the cross section falls off according to the classical Rutherford scaling $\sigma \propto v^{-4}$.

IV. COSMOLOGICAL CONSTRAINTS

A. Relic density and thermal freeze-out

Since the mass splitting is tiny relative to DM mass states, $\Delta m \ll T_f \simeq m_\chi/20$, the excited state remains in chemical equilibrium with the ground state during freeze-out, which gives $n_1^{\text{eq}} \simeq n_2^{\text{eq}} \simeq \frac{1}{2}n_{\text{tot}}^{\text{eq}}$. The off-diagonal $\mathbb{Z}_2$ -protected interaction in Eq. (7) forbids the mixed annihilation channel $\chi_1\chi_2 \rightarrow \phi\phi$ at tree level, while $\chi_1\chi_1 \rightarrow \phi\phi$ and $\chi_2\chi_2 \rightarrow \phi\phi$ are allowed and equal up to $\mathcal{O}(\Delta m/m_\chi)$ corrections. The effective annihilation rate entering the single Boltzmann equation for the total density is

$$\langle\sigma v\rangle_{\text{eff}}(T) = \sum_{i,j=1,2} \frac{n_i^{\text{eq}} n_j^{\text{eq}}}{(n_{\text{tot}}^{\text{eq}})^2} \langle\sigma v\rangle_{ij}(T) \simeq \frac{1}{2} \langle\sigma v\rangle_{11}(T). \quad (17)$$

The short-distance annihilation process $\chi_1\chi_1 \rightarrow \phi\phi$ is p -wave suppressed in the nonrelativistic limit [26], $(\sigma v)_{\text{Born}}(v) = \sigma_0 v^2$ with $\sigma_0 \equiv 3\pi\alpha^2/(2m_\chi^2)$. Long-range effects from the Yukawa force must be treated nonperturbatively in our resonant parameter region $\alpha m_\chi/m_\phi \gtrsim 1$. We therefore include a coupled-channel Sommerfeld enhancement factor $S_p(v)$ computed from the same potential parameters $(\alpha, m_\phi, \Delta m)$ as used in the scattering calculation (Appendix A). The Sommerfeld-enhanced and thermally averaged rate is

$$\langle\sigma v\rangle_{11}(x) = \frac{x^{3/2}}{2\sqrt{\pi}} \int_0^\infty dv v^2 e^{-xv^2/4} (\sigma v)_{\text{Born}}(v) S_p(v), \quad (18)$$

where $x \equiv m_\chi/T$. We solve the standard freeze-out Boltzmann equation using $\langle\sigma v\rangle_{\text{eff}}(x) = \frac{1}{2} \langle\sigma v\rangle_{11}(x)$. The full thermal-averaging procedure and the consistent

computation of $S_p(v)$ from the coupled-channel Schrödinger equation are given in Appendix C. To ensure physical consistency in the resonant regime, we must account for the unitarity limit of the p -wave annihilation cross section. The maximum possible inelastic cross section for the $l = 1$ partial wave is strictly bounded by the unitarity of the S matrix, scaling as $\sigma_{\text{uni}}^{(l=1)} \approx 3\pi/k_{\text{rel}}^2$. Standard factorization, $\sigma \propto (\sigma v)_{\text{Born}} S_p(v)$, can violate this bound near narrow resonances where the enhancement is large. A fully rigorous treatment involves determining the annihilation rate from the imaginary part of the optical potential in the S matrix formalism [49], such a computation is computationally prohibitive for a global parameter scan.

Instead, we adopt a unitarized prescription [50] that enforces saturation smoothly:

$$\sigma_{11}^{\text{unit}}(v) = \frac{(\sigma v)_{\text{Born}} S_p(v)/v}{1 + \frac{(\sigma v)_{\text{Born}} S_p(v)/v}{\sigma_{\text{uni}}^{(l=1)}}} \simeq \frac{(\sigma v)_{\text{Born}} S_p(v)}{v \left(1 + \frac{(\sigma v)_{\text{Born}} S_p(v) k_{\text{rel}}^2}{3\pi v} \right)}. \quad (19)$$

This expression reduces to the standard Sommerfeld-enhanced rate in the perturbative regime but saturates at the unitarity bound exactly on resonance. We implement this unitarized rate in the thermal average integral for the relic density calculation. We explicitly verified for our benchmark parameters ($\alpha \approx 10^{-2}$), the factorized cross section approaches the unitarity limit only in a extremely narrow velocity window near the resonance peak. Since the relic density is determined by a broad thermal average ($x_f \sim 20$), the precise shape of the cutoff at the pole has a negligible impact ($< 1\%$) on the final abundance compared to the Parikh-Sato-Slatyer limit.

Finally, bound-state formation (BSF) is negligible for our benchmark because radiative capture via on-shell mediator emission, $\chi\chi \rightarrow B + \phi$, is kinematically forbidden when $m_\phi > E_B$. Also, the leading off-shell process $\chi\chi \rightarrow B + \phi^* \rightarrow B + e^+e^-$ is suppressed by both the heavy propagator and the small leptophilic portal coupling. The quantitative estimate is given in Appendix C as well.

B. CMB anisotropies and p -wave suppression

Dark matter annihilation during the recombination epoch ($z \sim 1100$) injects energy into the primordial plasma, dampening the temperature anisotropies of the CMB [30]. Planck [1] constrains late-time electromagnetic energy injection through the parameter

$$p_{\text{ann}} \equiv f_{\text{eff}} \frac{\langle \sigma v \rangle}{m_\chi}, \quad (20)$$

where f_{eff} is the energy-deposition efficiency (channel- and mass-dependent) evaluated at $z \simeq 600$ [30]. For s -wave annihilation Planck [1] gives the approximate bound

$$p_{\text{ann}} < 3.2 \times 10^{-28} \text{ cm}^3 \text{ s}^{-1} \text{ GeV}^{-1} \quad (95\% \text{ C.L.}), \quad (21)$$

which we use as a conservative reference constraint. For velocity-suppressed annihilation the true constraint is weaker than the s -wave limits; using Eq. (21) is therefore conservative for our p -wave model.

In our model the primary annihilation products are $\phi\phi$, and each ϕ promptly decays to an e^+e^- pair. Thus the deposited power is that of an electromagnetic cascade initiated by four relativistic leptons. We incorporate this using an effective efficiency $f_{\text{eff}}^{\phi\phi \rightarrow 4e}(m_\chi, m_\phi)$. For $m_\phi \ll m_\chi$, each lepton carries energy $\sim m_\chi/2$ and we adopt

$$f_{\text{eff}} \equiv f_{\text{eff}}^{e^\pm}(E \simeq m_\chi/2), \quad (22)$$

using the tabulated energy-dependent efficiencies from Ref. [30] (evaluated at $z \simeq 600$). Numerically, for $m_\chi \simeq 40 \text{ GeV}$ this corresponds to $f_{\text{eff}} \sim \mathcal{O}(0.3)$ for e^+e^- -initiated cascades.

The relevant annihilation parameter is therefore

$$p_{\text{ann}}^{\text{model}} = f_{\text{eff}} \frac{\langle \sigma v \rangle_{\text{eff}}(z \simeq 600)}{m_\chi} \simeq f_{\text{eff}} \frac{1}{2} \frac{\sigma_0 v_{\text{rec}}^2 S_{p,\text{max}}}{m_\chi}, \quad (23)$$

with $v_{600} \sim 10^{-7}$ the typical DM velocity around $z \simeq 600$ and $S_{p,\text{max}}$ the saturated Sommerfeld factor. For our benchmark the resulting $p_{\text{ann}}^{\text{model}}$ is many orders of magnitude below the Planck limit in Eq. (21), even taking $f_{\text{eff}} = \mathcal{O}(1)$.

C. CMB safety and kinematic blocking of bound states

Dark matter annihilation during the recombination epoch ($z \sim 1100$) is strictly constrained by Planck [1]. Our model evades these constraints via two mechanisms: p -wave suppression of direct annihilation and the kinematic blocking of BSF.

- (a) *Direct annihilation:* The process $\chi_1\chi_1 \rightarrow \phi\phi$ is p -wave suppressed for Majorana fermions annihilating into scalars. At recombination velocities ($v \sim 10^{-7}$), the rate scales as $\langle \sigma v \rangle \propto v^2$, rendering energy injection negligible.
- (b) *BSF:* For our benchmark the mediator is heavier than the would-be binding energy, so radiative BSF via on-shell scalar emission, $\chi\chi \rightarrow B + \phi$, is kinematically forbidden when m_ϕ is larger than the bound state energy E_B . A representative Coulombic estimate [51] for the ground-state binding energy is

$$E_B \sim \frac{\alpha^2 m_\chi}{4}, \quad (24)$$

which gives $E_B \sim 1 \text{ MeV}$ for $\alpha = 10^{-2}$ and $m_\chi = 40 \text{ GeV}$, well below the limit $m_\phi = 20 \text{ MeV}$.

The leading allowed capture channel proceeds through an off-shell scalar that subsequently decays,

$\chi\chi \rightarrow \mathcal{B} + \phi^* \rightarrow \mathcal{B} + e^+e^-$. For momentum transfer $q^2 \sim E_B^2 \ll m_\phi^2$, the virtual propagator yields a strong suppression, $|(q^2 - m_\phi^2)^{-1}|^2 \simeq m_\phi^{-4}$, and the rate inherits an additional small-coupling penalty from $\phi^* \rightarrow e^+e^-$. A conservative estimate of the off-shell-to-on-shell capture-rate ratio is therefore

$$\frac{\Gamma_{\text{off}}}{\Gamma_{\text{on}}} \sim \frac{\alpha_{\text{EM}} g_e^2}{4\pi} \left(\frac{E_B}{m_\phi} \right)^4 \times \mathcal{O}(1), \quad (25)$$

where the factor $(E_B/m_\phi)^4$ encodes the squared propagator and the soft-emission kinematics. Numerically, with $E_B \simeq 1$ MeV, $m_\phi = 20$ MeV, and $g_e = 10^{-6}$, one has

$$\frac{\Gamma_{\text{off}}}{\Gamma_{\text{on}}} \lesssim 10^{-21}. \quad (26)$$

Therefore, off-shell BSF is utterly negligible both during freeze-out and at recombination, and does not compete with the Sommerfeld-enhanced annihilation channels considered in the Boltzmann equation.

D. Big bang nucleosynthesis

The light scalar mediator ϕ must decay before the onset of BBN ($t \sim 1$ s), otherwise the element abundance in today's Universe will not match the observations. Our leptophilic portal ensures this via the decay $\phi \rightarrow e^+e^-$. For a coupling $g_e \sim 10^{-6}$ and $m_\phi = 20$ MeV:

$$\tau_\phi \approx 0.8 \times 10^{-9} \text{ s} \ll 1 \text{ s}. \quad (27)$$

This rapid decay ensures that the dark sector entropy is transferred to the SM plasma well before neutrino decoupling. Unlike the Higgs portal [37], the leptophilic coupling does not require mixing with quarks, decoupling the BBN lifetime constraint from nuclear recoil constraints.

While on-shell mediator emission is forbidden, bound state formation can proceed via off-shell scalars decaying to electron-positron pairs: $\chi\chi \rightarrow \mathcal{B} + \phi^* \rightarrow \mathcal{B} + e^+e^-$. This process is kinematically allowed only if the binding energy exceeds the electron-positron mass threshold, $E_B > 2m_e \approx 1.02$ MeV. For our benchmark $\alpha \approx 10^{-2}$ and $m_\chi = 40$ GeV, the binding energy is $E_B \approx \frac{1}{4}\alpha^2 m_\chi \approx 1.0$ MeV. This lies just below the $2m_e$ threshold, rendering the off-shell process kinematically blocked or heavily phase-space suppressed. We impose the constraint $\alpha \lesssim 0.01$ to ensure $E_B < 2m_e$, thereby maintaining CMB robustness against all BSF channels.

E. Robustness of CMB constraints via saturation

The Sommerfeld enhancement $S_p \propto v^{-2}$ must physically saturate, otherwise we will have an unphysical infinity cross section when DM particles cool ($v \rightarrow 0$) as it scales

with $1/v^2$. We quantify the saturation explicitly. Saturation occurs when the particle's de Broglie wavelength is cut off by the finite range of the mediator, roughly at velocities $v_{\text{sat}} \sim m_\phi/m_\chi$.

For our benchmark ($m_\phi = 20$ MeV, $m_\chi = 40$ GeV), $v_{\text{sat}} \approx 5 \times 10^{-4}$ c. During recombination ($z \sim 1100$), the dark matter velocity is $v_{\text{rec}} \sim 10^{-7}$ c, placing the system deep in the saturation regime, $v_{\text{rec}} \ll v_{\text{sat}}$, so $S_p(v)$ has saturated to

$$S_{p,\text{max}} \equiv S_p(v \ll v_{\text{sat}}) \simeq \left(\frac{v_\star}{v_{\text{sat}}} \right)^2. \quad (28)$$

The CMB-relevant rate therefore scales as

$$\langle \sigma v \rangle_{\text{eff}}(v_{\text{rec}}) \simeq \frac{1}{2} \sigma_0 v_{\text{rec}}^2 S_{p,\text{max}}, \quad \sigma_0 = \frac{3\pi\alpha^2}{2m_\chi^2}. \quad (29)$$

For benchmark values $(m_\chi, m_\phi, \alpha) = (40 \text{ GeV}, 20 \text{ MeV}, 10^{-2})$, one has $\sigma_0 \simeq 2.9 \times 10^{-7} \text{ GeV}^{-2}$, i.e.,

$$\sigma_0 \simeq 3.4 \times 10^{-24} \text{ cm}^3/\text{s}. \quad (30)$$

Hence we have

$$\langle \sigma v \rangle_{\text{eff}}(v_{\text{rec}}) \simeq 1.7 \times 10^{-38} \text{ cm}^3/\text{s} \times S_{p,\text{max}}, \quad (31)$$

which remains negligible compared to the Planck bound even for very large saturated enhancements. This explicitly shows that saturation does not remove the v^2 suppression at recombination; it only replaces the v^{-2} growth of S_p by a constant factor.

F. Mediator lifetime and laboratory constraints

To satisfy BBN requirements while evading beam-dump constraints, we adopt a benchmark coupling of $g_e = 10^{-6}$ and $m_\phi = 20$ MeV. The leptophilic scalar decays into electron-positron pairs with a decay width given by Eq. (9). For our benchmark, this yields a proper lifetime $\tau_\phi \approx 6 \times 10^{-10}$ s and a decay length $c\tau \approx 18$ cm. The lifetime is significantly shorter than the onset of BBN ($\tau_\phi \ll 1$ s). The dark sector entropy is transferred to the SM plasma well before neutrino decoupling ($T \sim 2$ MeV), ensuring N_{eff} is undisturbed. With these results, the scalar is long-lived on the scale of collider vertex detectors but decays promptly relative to beam-dump baselines. This parameter space falls into a known gap in experimental coverage. For beam dump experiments, such as E137, they typically require long-lived particles to penetrate meters of shielding. The decay length of $c\tau \approx 18$ cm is thus the scalar decays primarily within the shield. The E137 constraints exclude $g_e \lesssim 10^{-7}$ for longer lifetimes [37,52]. Searches for visible dark photons or scalars at BABAR ($e^+e^- \rightarrow \gamma\phi$) and NA64 ($eN \rightarrow eN\phi$) currently constrain couplings

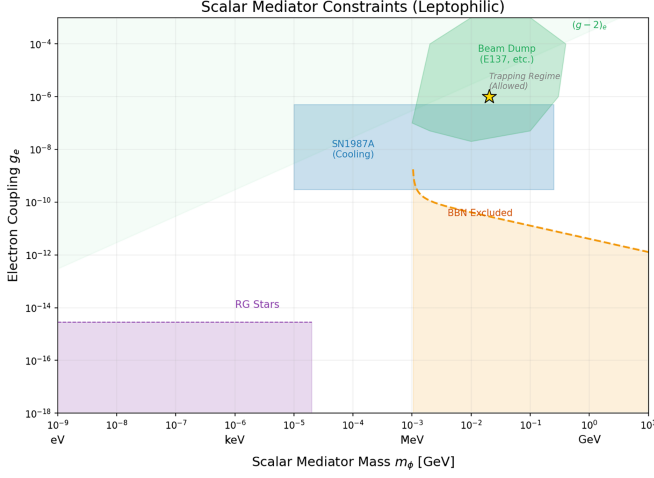


FIG. 2. Combined cosmological and astrophysical constraints on the leptophilic scalar mediator ϕ . The shaded regions indicate exclusion by stellar cooling in horizontal branch/red giant stars (purple) [67]; free-streaming cooling of SN1987A (blue) [58]; visible decay searches in beam-dump experiments (green) [68]; and BBN disruption for lifetimes $\tau_\phi > 1$ s (orange) [69]. The gold star marks our benchmark ($m_\phi = 20$ MeV, $g_e = 10^{-6}$), which resides in the SN1987A trapping regime where scalars are trapped in the core and decays promptly ($\tau \ll 1$ s) to ensure BBN safety, while remaining below the sensitivity of past beam-dump searches.

$g_e \gtrsim 10^{-4}$ in this mass range [53,54]. Our benchmark coupling of $g_e \approx 10^{-6}$ is 20 orders of magnitude below these limits, primarily due to the p -wave suppression of the scalar production cross section relative to vector models.

However, this viable window is the primary target for next-generation missing-momentum experiments. The LDMX experiment, designed to search for light dark matter via bremsstrahlung production ($eZ \rightarrow eZ\phi$), projects sensitivity to scalar couplings down to $g_e \sim 10^{-6}$ in the 10–100 MeV range [55]. A signal in LDMX, potentially distinguishable via the kinematics of the prompt decay, would serve as a complementary terrestrial probe to the direct detection. The combined cosmological and laboratory constraints on the scalar mediator, along with our benchmark, are summarized in Fig. 2.

G. Kinetic decoupling and drag

Scattering between dark matter and leptons ($\chi e \rightarrow \chi e$) maintains kinetic equilibrium in the early Universe. For our benchmark parameters ($g_e \approx 10^{-6}$, $m_\phi = 20$ MeV), the momentum transfer rate Γ_{kd} drops below the Hubble rate H before recombination. From standard WIMP decoupling, the decoupling temperature is roughly $T_{\text{kd}} \sim 10$ MeV [56]. Since $T_{\text{kd}} \gg T_{\text{rec}} \sim 0.3$ eV, the dark matter is kinetically decoupled during the formation of the CMB and large scale structure. Therefore there is no significant DM-lepton drag to suppress the matter power spectrum or alter CMB acoustic peaks [57]. We verify that the kinetic decoupling

temperature T_{kd} is determined by equating the momentum transfer rate $\gamma(T)$ to the Hubble rate $H(T)$. For t -channel scattering $\chi e \rightarrow \chi e$,

$$\gamma(T) \approx \frac{31\pi^3}{9} \frac{g_s^2 g_e^2 T^6}{m_\chi m_\phi^4}. \quad (32)$$

Solving for our benchmark yields $T_{\text{kd}} \approx 15$ MeV. Since $T_{\text{kd}} \gg T_{\text{BBN}} \sim 1$ MeV, the dark sector is fully decoupled during nucleosynthesis. Comparing this to $H \sim T^2/M_{\text{Pl}}$, and using $g_e = 10^{-6}$, $\alpha = 10^{-2}$, we find that kinetic decoupling occurs at $T_{\text{kd}} \sim 10$ –50 MeV. Crucially, $T_{\text{kd}} \ll T_{\text{freeze-out}} (\approx 2 \text{ GeV})$, ensuring that the relic density calculation in the standard thermal bath remains valid. Furthermore, $T_{\text{kd}} \gg T_{\text{BBN}} (\sim 1 \text{ MeV})$ implies that the dark fluid is fully decoupled and cold during nucleosynthesis and structure formation.

H. Supernova trapping

Constraints from Supernova 1987A are significant [58]. It is driven by the Raffelt criterion [59] which forbids exotic energy loss rates exceeding $L_\nu \sim 3 \times 10^{52}$ erg/s. For the dark sector it excludes a cooling wedge of couplings where particles are produced efficiently but escape the core. In the supernova core ($\rho \approx 3 \times 10^{14}$ g/cm³, $T \approx 30$ MeV), the electron chemical potential $\mu_e \approx 300$ MeV significantly exceeds the scalar mass $m_\phi = 20$ MeV. The standard decay channel $\phi \rightarrow e^+e^-$ is heavily Pauli blocked, as the final-state electrons cannot access the occupied states below the Fermi surface ($E_e \sim m_\phi/2 \ll \mu_e$). Therefore, the trapping argument cannot rely on simple inverse decay [60].

Instead, the opacity is dominated by absorption via inverse bremsstrahlung on protons ($\phi p \rightarrow ep$) and Compton-like scattering ($\phi e \rightarrow \gamma e$). Following the degenerate plasma formalism of Chang *et al.* [58], the absorptive width for a scalar in this regime is given by

$$\Gamma_{\text{abs}} \approx \sigma_{\text{brem}} n_p v_{\text{rel}} \approx \frac{4\pi\alpha_{\text{EM}} g_e^2 n_p}{3m_e^2 E_\phi} \mathcal{F}_{\text{deg}}, \quad (33)$$

where $\mathcal{F}_{\text{deg}}$ represents the phase-space factors for the degenerate electrons. For our coupling $g_e = 10^{-6}$, the absorption mean free path is

$$\lambda_{\text{abs}} \equiv \Gamma_{\text{abs}}^{-1} \sim \mathcal{O}(10^2 \text{ cm}). \quad (34)$$

This is negligible compared to the core radius ($R_{\text{core}} \sim 10$ km). The characteristic diffusion time for the scalar to escape is

$$\tau_{\text{diff}} \sim \frac{R_{\text{core}}^2}{c\lambda_{\text{abs}}} \sim 10^3 \text{ s}. \quad (35)$$

Since $\tau_{\text{diff}} \gg \tau_\nu (\sim 10 \text{ s})$, the scalar fluid is fully thermalized and trapped within the neutrinosphere. It does not contribute

to free-streaming cooling but participates in the thermal transport. Thus, the Pauli blocking of decays does not reopen the constraint window for $g_e \sim 10^{-6}$.

I. Excited state decay

The electromagnetic energy injection from $\chi_2 \rightarrow \chi_1 \gamma$ depends on the excited-state fraction at the time of decay. After freeze-out of the total comoving DM density, the excited-state fraction $f_{\chi_2} \equiv n_{\chi_2}/n_{\text{DM}}$ is primarily depleted by the exothermic pair conversion $\chi_2 \chi_2 \rightarrow \chi_1 \chi_1$, which remains efficient down to dark-sector temperatures $T_\chi \lesssim \Delta m$ (see Sec. IV J for the full Boltzmann treatment). The residual fraction subsequently decays via $\chi_2 \rightarrow \chi_1 \gamma$ with lifetime $\tau_{\chi_2} = 1/\Gamma(\chi_2 \rightarrow \chi_1 \gamma)$. The combined effect yields $f_{\chi_2} \ll 10^{-6}$ by recombination.

The total electromagnetic energy released per decay is Δm . The fractional energy injected into the photon bath at time t is therefore

$$\left. \frac{\Delta \rho_\gamma}{\rho_\gamma} \right|_t \simeq f_{\chi_2}(t) \frac{\rho_{\text{DM}}(t)}{\rho_\gamma(t)} \frac{\Delta m}{m_\chi}, \quad (36)$$

up to an $\mathcal{O}(1)$ deposition factor for sub-keV photons. Evaluating near the decay time $t \simeq \tau_{\chi_2}$ (where $f_{\chi_2} \simeq f_{\chi_2}(t_{\text{fo}})/e$), and using $\Delta m = 100$ eV and $m_\chi = 40$ GeV, we obtain

$$\left. \frac{\Delta \rho_\gamma}{\rho_\gamma} \right|_{t \simeq \tau_{\chi_2}} \ll 10^{-4}, \quad (37)$$

well below the COBE/FIRAS bound on μ - and y -type spectral distortions [61]. Moreover, the emitted photons have energy $\Delta m \ll \mathcal{O}(\text{MeV})$ and therefore cannot photodissociate light nuclei, so there is no BBN photodissociation constraint. We emphasize that p_{ann} constraints apply to annihilationlike power injection near recombination [1], whereas χ_2 decays are constrained by spectral distortions and other limits [62] on late electromagnetic energy injection, with the impact strongly suppressed by $\Delta m/m_\chi \ll 1$.

J. Excited-state depletion and conversion freeze-out

A crucial assumption is that the excited-state fraction $f_{\chi_2} \equiv n_{\chi_2}/(n_{\chi_1} + n_{\chi_2})$ becomes negligible at late times. The ratio n_{χ_2}/n_{χ_1} tracks its equilibrium value only through number-changing conversion processes of interchanging χ_1 and χ_2 , which is efficient compared to Hubble expansion.

In the Z_2 , the dominant conversion reactions are pair conversion and its inverse,

$$\chi_2 \chi_2 \rightarrow \chi_1 \chi_1, \quad \chi_1 \chi_1 \rightarrow \chi_2 \chi_2. \quad (38)$$

Single- χ_2 conversion processes such as $\chi_2 \chi_1 \rightarrow \chi_1 \chi_1$ are forbidden by Z_2 . In addition, the transition dipole operator

induces radiative deexcitation $\chi_2 \rightarrow \chi_1 \gamma$ with rate $\Gamma_{\text{dec}} \equiv \Gamma(\chi_2 \rightarrow \chi_1 \gamma)$.

After chemical freeze-out of the total comoving DM density $n_{\text{DM}} \equiv n_{\chi_1} + n_{\chi_2}$, the evolution of the excited fraction $f \equiv f_{\chi_2}$ is governed by

$$\dot{f} = -2n_{\text{DM}}[\langle \sigma v \rangle_{22 \rightarrow 11} f^2 - \langle \sigma v \rangle_{11 \rightarrow 22} (1-f)^2] - \Gamma_{\text{dec}} f, \quad (39)$$

where dots denote derivatives with respect to cosmic time, and $n_{\text{DM}} \propto a^{-3}$. The up- and down-conversion rates are related by detailed balance,

$$\langle \sigma v \rangle_{11 \rightarrow 22} = \langle \sigma v \rangle_{22 \rightarrow 11} \left(\frac{n_{\chi_2}^{\text{eq}}}{n_{\chi_1}^{\text{eq}}} \right)^2 \simeq \langle \sigma v \rangle_{22 \rightarrow 11} e^{-2\Delta m/T_\chi}, \quad (40)$$

where $T_\chi \equiv T_{\text{dark}}$ is the dark-matter kinetic temperature.

The conversion processes maintain $f \simeq f_{\text{eq}}$ as long as the conversion rate per excited particle exceeds Hubble.

$$\Gamma_{\text{conv}}(T) \equiv 2n_{\chi_2} \langle \sigma v \rangle_{22 \rightarrow 11} = 2fn_{\text{DM}} \langle \sigma v \rangle_{22 \rightarrow 11} \gtrsim H(T). \quad (41)$$

When $\Gamma_{\text{conv}} \lesssim H$, the excited fraction freezes out at a residual value and redshifts.

For the cosmological inputs, we use

$$n_{\text{DM}}(T) = \frac{\rho_{\text{DM}}(T)}{m_\chi} = \frac{\Omega_{\text{DM}} \rho_c}{m_\chi} (1+z)^3, \\ H(T) = 1.66 \sqrt{g_*} \frac{T^2}{M_{\text{Pl}}}. \quad (42)$$

After kinetic decoupling at temperature T_{kd} , the DM temperature cools as $T_\chi \propto a^{-2}$, implying (for $T < T_{\text{kd}}$)

$$T_\chi(T) \simeq \frac{T^2}{T_{\text{kd}}}. \quad (43)$$

We determine T_{kd} same previously and compute $\langle \sigma v \rangle_{22 \rightarrow 11}(T_\chi)$ by thermally averaging the velocity-dependent exothermic conversion cross section (Appendix A).

Solving Eq. (39) with Eqs. (40)–(43), we find that conversion remains efficient down to $T_\chi \lesssim \Delta m$. Then f_{χ_2} becomes exponentially small. For the benchmark $\Delta m \simeq 100$ eV we obtain $f_{\chi_2} \ll 10^{-6}$ by recombination, validating the approximation of a pure ground-state halo at late times and rendering constraints from late χ_2 decays negligible due to the tiny excited fraction and the small fractional energy release $\Delta m/m_\chi$.

K. Late-time EUV emission and molecular-cloud ionization constraints

In the Milky Way halo, the endothermic inelastic process $\chi_1\chi_1 \rightarrow \chi_2\chi_2$ is kinematically open at typical Galactic velocities. Taking a one-dimensional velocity dispersion $v_0 \simeq 220$ km/s, the characteristic relative speed is $v_{\text{rel}} \sim \sqrt{2}v_0 \simeq 311$ km/s, implying a typical center-of-mass kinetic energy

$$E_{\text{kin}} = \frac{\mu v_{\text{rel}}^2}{2} = \frac{m_\chi v_{\text{rel}}^2}{4} \sim \mathcal{O}(10 \text{ keV}) \gg 2\Delta m, \quad (44)$$

so upscattering proceeds efficiently in the Galactic environment while remaining closed in sufficiently cold systems such as dwarf galaxies.

Once produced, the excited state χ_2 decays through the dipole operator, $\chi_2 \rightarrow \chi_1\gamma$, emitting a monoenergetic photon with $E_\gamma \simeq \Delta m$. For the benchmark values used the decay rate is

$$\Gamma_{\chi_2} \simeq \frac{1}{\pi} \frac{\Delta m^3}{\Lambda_{\text{eff}}^2}, \quad \tau_{\chi_2} \equiv \Gamma_{\chi_2}^{-1}. \quad (45)$$

For $\Lambda_{\text{eff}} = 10^7$ GeV and $\Delta m = 100$ eV this gives $\tau_{\chi_2} \sim 2 \times 10^{11}$ s, corresponding to a decay length $d \sim v_0 \tau_{\chi_2} \sim \mathcal{O}(\text{pc})$, much smaller than Galactic-scale density gradients. In this regime the excited state reaches local steady state. The local upscattering event rate is [63]

$$R_\uparrow(\mathbf{x}) = \frac{1}{2} n_1^2(\mathbf{x}) \langle \sigma v \rangle_{11 \rightarrow 22}, \quad (46)$$

and each event produces two χ_2 states and each χ_2 yields one photon upon decay. The local photon emissivity is therefore

$$q_\gamma(\mathbf{x}) \simeq \Gamma_{\chi_2} n_2(\mathbf{x}) \simeq 2R_\uparrow(\mathbf{x}) = n_1^2(\mathbf{x}) \langle \sigma v \rangle_{11 \rightarrow 22}. \quad (47)$$

We evaluate $\langle \sigma v \rangle_{11 \rightarrow 22}$ by integrating our exact nonperturbative inelastic cross section over the Milky Way velocity distribution; for the benchmark parameters we find $\langle \sigma v \rangle_{11 \rightarrow 22} \sim 10^{-15} \text{ cm}^3/\text{s}$ at Galactic velocities. The approximation $q_\gamma \simeq 2R_\uparrow$ holds provided radiative decay dominates over collisional deexcitation, $\Gamma_{\chi_2} \gg n_2 \langle \sigma v \rangle_{22 \rightarrow 11}$ or equivalently $R_\downarrow \ll \Gamma_{\chi_2} n_2$; we verify this condition for our benchmark throughout the Milky Way halo.

Dense molecular clouds (MCs) provide a sensitive probe of sub-keV photon injection, since UV/EUV photons are efficiently absorbed in neutral gas and produce primary and secondary ionizations. We compare our predicted EUV injection against the MC ionization limits derived in [64]. The ionization rate per H_2 molecule at position $\mathbf{x}$ can be written as

$$\zeta_{\text{H}_2}(\mathbf{x}) = \int_{I_{\text{H}_2}}^{\infty} dE J_\chi(E, \mathbf{x}) \sigma_{\text{H}_2}^{\text{ion}}(E) [1 + \theta_e(E)], \quad (48)$$

where $I_{\text{H}_2} = 15.4$ eV, $\sigma_{\text{H}_2}^{\text{ion}}(E)$ is the photoionization cross section, and the number of secondary ionizations per primary ionization is parametrized by

$$\theta_e(E) = \frac{E - I_{\text{H}_2}}{W}, \quad W \simeq 40 \text{ eV}, \quad (49)$$

with W the mean energy loss per ion pair in the interstellar medium.

In our model the injected photon spectrum is monochromatic at $E_\gamma \simeq \Delta m$, so the source term can be written as $Q(E, \mathbf{x}) = q_\gamma(\mathbf{x}) \delta(E - E_\gamma)$. We convert the volume emissivity into an in-cloud photon flux via the one-zone relation

$$J_\chi(E, \mathbf{x}) = 2Q(E, \mathbf{x}) \frac{N_{\text{H}_2}}{n_{\text{H}_2}} \simeq 2Q(E, \mathbf{x}) L_{\text{MC}}, \quad (50)$$

where n_{H_2} is the characteristic H_2 density in the cloud, N_{H_2} its column density, and $L_{\text{MC}} \sim N_{\text{H}_2}/n_{\text{H}_2}$ its characteristic size.

Evaluating Eq. (48) for monochromatic injection yields

$$\zeta_{\text{H}_2}(\mathbf{x}) \simeq J_\chi(E_\gamma, \mathbf{x}) \sigma_{\text{H}_2}^{\text{ion}}(E_\gamma) \left(1 + \frac{E_\gamma - I_{\text{H}_2}}{W} \right), \quad (51)$$

with $J_\chi(E_\gamma, \mathbf{x}) \simeq 2q_\gamma(\mathbf{x}) L_{\text{MC}}$. We apply this expression to the cloud targets and limits compiled in [64], using the same assumptions for the DM density: for local clouds we take the local DM density, while for inner-Galaxy clouds we adopt an NFW profile as in [64]. We find that our predicted ζ_{H_2} lies below the robust limits from well-characterized clouds such as the L1551 and the DRAGON cloud. We present results for the inner-Galaxy cloud G1.4–1.8 + 87 separately, noting that it is treated as an optimistic forecast target in [64] due to astrophysical systematics. Related probes of soft photon injection at high redshift include 21-cm effects from decaying light DM producing Lyman-series photons [65]; in our benchmark $\tau_{\chi_2} \sim 6 \times 10^3 \text{ yr} \ll t(z \sim 20)$, so primordial χ_2 decays are complete well before cosmic dawn, and any late-time signal is dominated by Milky-Way upscattering. For keV–MeV splittings, x-ray searches provide complementary constraints on excited-state decays [66], but these do not directly apply at $E_\gamma \sim 100$ eV. We note that in sufficiently massive halos where the endothermic channel $\chi_1\chi_1 \rightarrow \chi_2\chi_2$ is kinematically open, the subsequent radiative decay $\chi_2 \rightarrow \chi_1 + \gamma$ constitutes a dissipative cooling mechanism: each upscattering event ultimately converts kinetic energy into a radiated photon of energy Δm . The cooling rate per unit volume scales as $\dot{q}_{\text{cool}} \sim n_{\chi_1}^2 \langle \sigma v \rangle_{11 \rightarrow 22} \cdot 2\Delta m$. For our benchmark parameters in a Milky-Way-like halo ($\rho_\chi \approx 0.4 \text{ GeV}/\text{cm}^3$, $\langle \sigma v \rangle_{11 \rightarrow 22} \sim 10^{-15} \text{ cm}^3/\text{s}$, $\Delta m = 100 \text{ eV}$), this yields $\dot{q}_{\text{cool}} \sim 10^{-38} \text{ erg cm}^{-3} \text{ s}^{-1}$, which is negligible compared to the gravitational heating rate $\dot{q}_{\text{grav}} \sim \rho_\chi \sigma_\chi^2 / t_{\text{dyn}}$ by many orders of magnitude. The associated

cooling timescale $t_{\text{cool}} \sim \rho_\chi \sigma_v^2 / \dot{q}_{\text{cool}}$ far exceeds the Hubble time for all halos in our benchmark. Nevertheless, for parameter points with larger Δm or stronger couplings, dissipative effects could become relevant and modify the halo density profile beyond what purely elastic or inelastic SIDM simulations predict. A definitive assessment of this dissipative channel requires dedicated N -body simulations incorporating radiative losses, which is beyond the scope of this work.

V. BENCHMARK AND ALLOWED PARAMETER SPACE

We performed a comprehensive Monte Carlo parameter scan using the full nonperturbative coupled-channel calculation (see Appendix A). We imposed conditions for satellite suppression (Draco, $\sigma < 0.2 \text{ cm}^2/\text{g}$), dwarf galaxy solution ($\sigma \sim 1\text{--}10 \text{ cm}^2/\text{g}$), and cluster constraints ($\sigma < 0.6 \text{ cm}^2/\text{g}$). The combined relic, nonperturbative, and cluster-scale constraints in the (M_χ, α) plane are shown in Fig. 3, with the benchmark indicated.

The full velocity-dependent transport cross section across the satellite, dwarf, and cluster regimes is shown in Fig. 4. At cluster velocities ($v \sim 10^3 \text{ km/s}$), the nonperturbative calculation yields a transport cross section $\sigma_T/m \approx 0.05 \text{ cm}^2/\text{g}$. This value is safely below the robust upper bound of $\sigma/m < 0.6 \text{ cm}^2/\text{g}$ derived from cluster halo ellipticities and merging clusters.

Our nonperturbative analysis identifies a resonant benchmark around

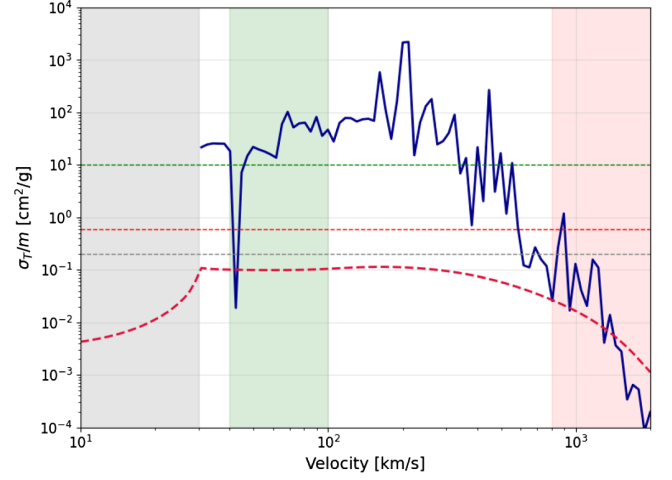


FIG. 4. The blue solid line is the velocity-dependent transport cross section σ_T/m for the scalar inelastic model (benchmark: $m_\chi = 40 \text{ GeV}$, $\Delta m = 100 \text{ eV}$). The calculation uses the full nonperturbative Schrödinger solution. The cross section is suppressed in the satellite regime (gray), resonates in the dwarf regime (green), and is suppressed in the cluster regime (red). The red dashed line is the residual elastic scattering, lower than the satellite bound and negligible when the inelastic scattering is opened.

$$\begin{aligned} m_\chi &\approx 40 \text{ GeV}, & m_\phi &\approx 20 \text{ MeV}, \\ \Delta m &\approx 100 \text{ eV}, & \alpha &\approx 10^{-2}. \end{aligned} \quad (52)$$

This point passes all astrophysical constraints via the resonance mechanism described in Sec. III.

Our scan identifies a distinct island of viability in the $m_\chi - \Delta m$ plane. Figure 5 illustrates the allowed region where Δm scales with m_χ to maintain the kinematic threshold.

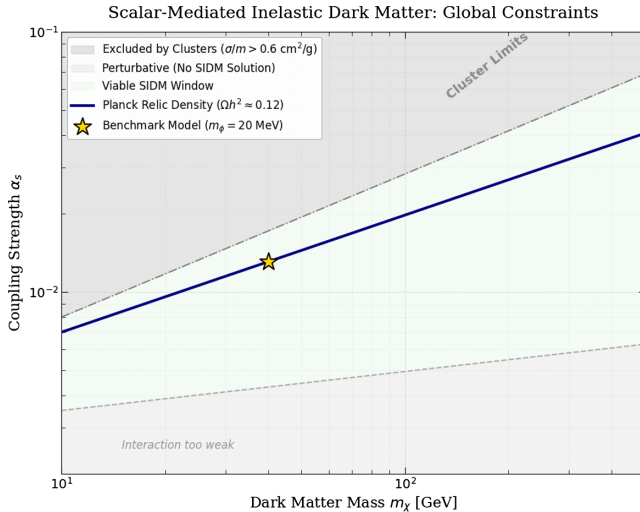


FIG. 3. Discovery window in the (M_χ, α) plane, combining bounds from the relic density [1], nonperturbative limits (below which the non-perturbative treatment fails), and cluster exclusion [20].

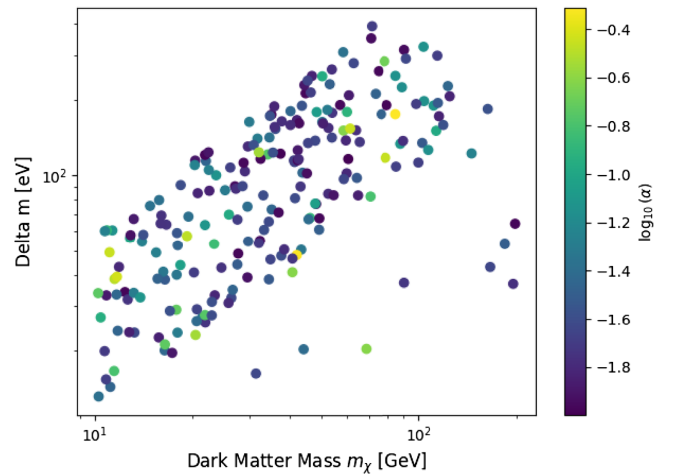


FIG. 5. Allowed parameter space from nonperturbative analysis. The color scale indicates the coupling strength $\log_{10}(\alpha)$. The points represent models that satisfy all self-interaction constraints.

VI. DIRECT DETECTION PHENOMENOLOGY

The direct detection phenomenology of our model is different from standard Higgs-portal scenarios because its the scalar mediator is leptophilic. The leptophilic portal induces an effective dark matter charge radius and anapole moment at the loop level via $\chi\text{-}\chi\text{-}\phi\text{-}\ell\text{-}\ell$ diagrams [70]. Dimensional analysis suggests these contributions scale as $g_S g_\ell e / (16\pi^2 m_\phi^2)$. Compared to the transition dipole operator $1/\Lambda$ (where $\Lambda_{\text{eff}} \sim 10^7$ GeV), these loop effects are suppressed by factors of the lepton mass and loop factors, thus they are subdominant to the tree-level dipole scattering.

A. Standard electronic and nuclear recoils

In standard scalar SIDM models, the mixing with the Higgs boson generates a spin-independent cross section on nucleons, σ_{SI} . For a mixing angle $\sin \theta \sim 10^{-4}$ required for BBN in the standard scenario, this cross section is typically $\sim 10^{-36}$ cm², which is firmly excluded by underground direct search experiments for $m_\chi \approx 40$ GeV. The $\Delta m \sim 100$ eV splitting is also insufficient to kinematically block the scattering [38–40]. However, in the leptophilic scenario the tree-level coupling to quarks is absent ($g_q = 0$). Scattering off nuclei can only occur via loop processes like a photon loop involving the scalar-electron loop. These processes are loop suppressed and negligible, or via the magnetic dipole operator. Inelastic up-scattering $\chi_1 N \rightarrow \chi_2 N$ is kinematically accessible in terrestrial detectors. However, to satisfy current constraints from liquid xenon experiments, the dipole suppression scale must be $\Lambda_{\text{eff}} \geq 10^7$ GeV. This implies a macroscopic lifetime for the excited state:

$$\tau_{\chi_2} \approx \frac{\pi \Lambda_{\text{eff}}^2}{\Delta m^3} \sim 10^{11} \text{ s}. \quad (53)$$

Even at characteristic dark matter velocities ($v \sim 10^{-3}c$), the decay length is astronomical ($L \gg 10^8$ km). The excited state escapes the detector volume long before decaying. The “luminous” photon signature is therefore not observable in standard setups.

Instead, the primary discovery mode is the unique nuclear recoil spectrum induced by the massless photon mediator. Unlike contact interactions which yield a flat recoil spectrum at low energies, the dipole operator induces a differential cross section scaling roughly as E_R^{-1} :

$$\frac{d\sigma}{dE_R} \approx \frac{4\pi\alpha_{\text{EM}}Z^2}{\Lambda_{\text{eff}}^2} \frac{1}{E_R} \left(1 - \frac{v_{\text{min}}^2}{v^2}\right) F_{\text{Helm}}^2(E_R). \quad (54)$$

This results in a sharp rise in the event rate at low recoil energies, strictly truncated by the inelastic kinematic threshold. Future low-threshold experiments (e.g., SuperCDMS) will search for this characteristic spectral shape rather than a monochromatic photon. Furthermore,

even if the leptophilic portal allows for $\chi e \rightarrow \chi e$ scattering, with the corrected coupling $g_e \approx 10^{-6}$, the cross section is highly suppressed. Therefore, the electronic recoil (ER) rate of our model is negligible compared to current limits in Refs. [71,72], leaving the magnetic dipole nuclear recoil channel as the primary discovery mode.

B. Magnetic dipole nuclear recoil

From previous discussion, the primary direct detection signature arises from the transition magnetic dipole operator. This induces inelastic scattering $\chi_1 N \rightarrow \chi_2 N$ mediated by photon exchange. The differential cross section is

$$\frac{d\sigma}{dE_R} = \frac{4\pi\alpha_{\text{EM}}Z^2}{\Lambda_{\text{eff}}^2} \frac{1}{E_R} \left[1 - \frac{E_R m_N + 2m_\chi}{v^2} - \frac{\Delta m}{v^2 \mu_{\chi N}}\right] F_{\text{Helm}}^2(E_R), \quad (55)$$

which gives a detection rate

$$\frac{dR}{dE_R} = N_T \frac{\rho_\chi}{m_\chi} \int_{v_{\text{min}}(E_R)}^\infty \frac{d\sigma}{dE_R} v f(v) dv. \quad (56)$$

For our benchmark $\Delta m = 100$ eV, the minimum velocity v_{min} required to produce a recoil E_R is significantly increased compared to elastic scattering:

$$v_{\text{min}} = \sqrt{\frac{1}{2m_N E_R} \left(\frac{m_N E_R}{\mu_{\chi N}} + \Delta m \right)}. \quad (57)$$

For a given Λ_{eff} , the excited-state lifetime is $\tau_{\chi_2} \sim \pi \Lambda_{\text{eff}}^2 / \Delta m^3$, so for sufficiently large Λ_{eff} the excited states produced in the detector escape before decaying ($c\tau \gg L_{\text{det}}$). The scattering cross section via the dipole operator includes two terms: the coherent dipole-charge interaction ($\propto Z^2$) and the spin-dependent dipole-dipole interaction ($\propto \mu_N^2$).

$$\frac{d\sigma}{dE_R} \propto \frac{1}{E_R} \left[Z^2 F_{\text{charge}}^2(E_R) + \frac{J+1}{3J} \mu_N^2 F_{\text{spin}}^2(E_R) \right]. \quad (58)$$

For xenon ($Z = 54$), the coherent charge term dominates by a factor of $\sim 10^4$. We utilize this enhanced response in our projections. However, the stability requirement $\Lambda_{\text{eff}} \geq 10^7$ GeV suppresses the total rate to $< 10^{-6}$ events/ton-year, rendering it unobservable in current facilities.

C. Dipole nuclear response and experimental reach

Following the formalism of magnetic inelastic dark matter (MiDM) [73,74], the differential cross section for dipole-induced scattering is the sum of a dipole-charge (DZ) term and a dipole-dipole (DD) term:

$$\frac{d\sigma}{dE_R} = \frac{d\sigma_{DZ}}{dE_R} + \frac{d\sigma_{DD}}{dE_R}. \quad (59)$$

For heavy target nuclei like xenon ($Z = 54$), the charge term dominates due to the coherent Z^2 enhancement:

$$\frac{d\sigma_{DZ}}{dE_R} \approx \frac{4\pi\alpha_{\text{EM}}Z^2}{E_R\Lambda_{\text{eff}}^2} \left[1 - \frac{E_R}{v^2} \left(\frac{m_N + 2m_\chi}{2m_\chi m_N} \right) - \frac{\Delta m}{v^2\mu_{\chi N}} \right] |F_{\text{ch}}(E_R)|^2, \quad (60)$$

where $F_{\text{ch}}(E_R)$ is the nuclear charge form factor. The dipole-dipole term, proportional to the nuclear magnetic moment μ_N^2 , is subdominant for spin-suppressed isotopes (e.g., ^{132}Xe) and is neglected in our projection, providing a conservative estimate of the rate.

D. Results and future detectability

Current direct detection experiments such as XENON1T have set stringent limits on the dark matter magnetic dipole moment, excluding $\mu_\chi \gtrsim 10^{-5}\mu_N$ for mass splittings $\Delta m \lesssim 100$ keV. Our benchmark scale $\Lambda_{\text{eff}} \sim 10^7$ GeV corresponds to an effective moment $\mu_\chi \approx e/\Lambda_{\text{eff}} \sim 2 \times 10^{-7}\mu_N$. Since the event rate scales as $\mu_\chi^2 \propto \Lambda_{\text{eff}}^{-2}$, our benchmark signal is suppressed by a factor of ~ 2500 relative to the current exclusion threshold. This confirms that the model lies deep within the allowed region.

To substantiate the discovery potential, we project the event rates for next-generation experiments by integrating the differential cross section over the detector efficiency $\epsilon(E_R)$. We assume a sharp threshold efficiency to illustrate the scaling.

The total event rate scales with the dipole suppression scale as $R \propto \Lambda_{\text{eff}}^{-2}$. Calibrating to the current XENON1T exclusion limit $\Lambda_{\text{eff}} \gtrsim 2 \times 10^5$ GeV, corresponding to $\mathcal{O}(10)$ events/ton-year, we obtain the following projection for a Liquid Xenon target (LZ/XLZD) with $E_{\text{th}} = 1$ keV_{nr}:

$$R_{Xe} \approx 0.4 \left(\frac{10^6 \text{ GeV}}{\Lambda_{\text{eff}}} \right)^2 \text{ events}/(\text{ton} \cdot \text{year}). \quad (61)$$

For our benchmark ($\Lambda_{\text{eff}} = 10^7$ GeV), the rate is $\sim 4 \times 10^{-3}$ events/(ton · year), which is effectively unobservable. However, a discovery benchmark of $\Lambda_{\text{eff}} = 10^6$ GeV remains consistent with all current null results but yields ~ 40 signal events in a 100 ton-year exposure of XLZD. This signal magnitude would be statistically significant above the coherent neutrino scattering floor.

For low-threshold detectors like SuperCDMS (germanium, $Z = 32$), the lack of Z^2 enhancement is compensated by the ability to probe the $1/E_R$ divergence. Assuming $E_{\text{th}} = 100$ eV_{nr}, we project

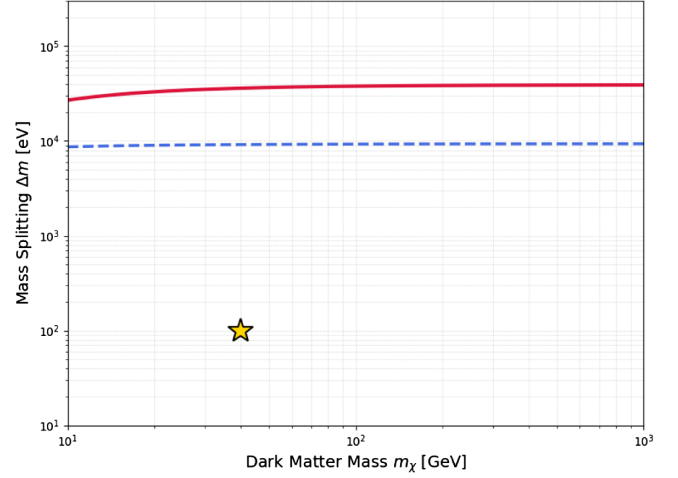


FIG. 6. Kinematic accessibility for inelastic dipole dark matter. The solid red line is the maximum mass splitting Δm accessible to liquid xenon experiments LZ, PandaX-4T for a representative nuclear-recoil threshold. The dashed blue line shows the corresponding reach for germanium-based detectors SuperCDMS assuming $E_{\text{th}} = 0.1$ keV_{nr} and the lighter target nucleus ($m_{\text{Ge}} \approx 72.6$ GeV). The star marks our resonant benchmark, which is kinematically accessible to both targets. In the direct-detection-safe regime the overall rate is nevertheless strongly suppressed by the large dipole scale, scaling as $R \propto \Lambda_{\text{eff}}^{-2}$; lowering E_{th} primarily enlarges the accessible Δm range and improves sensitivity to the characteristic $1/E_R$ recoil spectrum.

$$R_{\text{Ge}} \approx 0.2 \left(\frac{10^6 \text{ GeV}}{\Lambda_{\text{eff}}} \right)^2 \text{ events}/(\text{ton} \cdot \text{year}). \quad (62)$$

While the integral rate is slightly lower than xenon, the observation of the spectral rise below 1 keV would constitute a smoking-gun signature distinguishing the dipole operator from contact interactions or neutrino backgrounds. Thus, the model offers a clear target for the multiton scale era, provided the dipole scale lies in the $\Lambda_{\text{eff}} \sim 10^6$ GeV window.

The integrated event rates per unit mass are comparable between the two technologies. The overall normalization requires Λ_{eff} in the multi-PeV range to satisfy null results, leading to rates of $\sim 10^{-7}$ events/(ton · yr) in xenon and $\sim 10^{-6}$ events/(ton · yr) in germanium for our benchmark. While these rates are unobservable at present exposures, future low-threshold searches can improve sensitivity by probing the characteristic $1/E_R$ spectral shape. A signal in SuperCDMS exhibiting the E_R^{-1} rise, combined with a kinematically suppressed signal in LZ, would provide a distinct fingerprint for this dipole inelastic scenario.

In Fig. 6, we illustrate the kinematic accessibility of the inelastic channel in the $(m_\chi, \Delta m)$ plane. The solid red curve denotes the maximum splitting accessible to liquid xenon ($E_{\text{th}} = 1$ keV_{nr}), while the dashed blue curve shows the reach for germanium ($E_{\text{th}} = 0.1$ keV_{nr}). Our resonant benchmark $(m_\chi, \Delta m) = (40 \text{ GeV}, 100 \text{ eV})$ lies well below

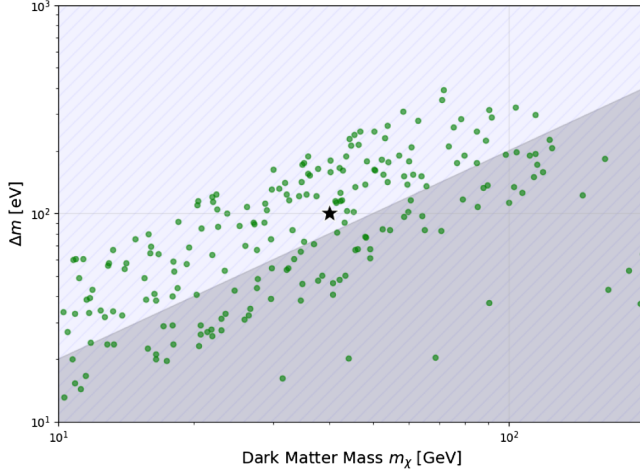


FIG. 7. The discovery window. Green points represent the viable SIDM parameter space. The gray region is excluded by satellite stability. The benchmark point (star) sits within the discovery window for future low-threshold experiments via the dipole portal.

both curves, confirming that inelastic nuclear recoils are kinematically permitted in both detector types, limited only by the interaction strength Λ_{eff} .

VII. ROBUSTNESS OF THE BENCHMARK MODEL

We analyze the stability of our chosen benchmark point ($m_\chi = 40$ GeV, $m_\phi = 20$ MeV, $\alpha \approx 0.013$) against the global constraints as visualized in Fig. 7. The benchmark lies within the viable SIDM window, satisfying all cosmological and astrophysical bounds. Below, we quantify the safety margins for the key model parameters, defining the percentage variation allowed before the model encounters an exclusion limit.

A. Coupling strength (α)

The dark sector coupling strength α is the most constrained parameter, bounded from above by cluster halo ellipticity [75] constraints and from below by the requirement to solve small-scale structure problems. The upper bound is contributed by cluster limits. The benchmark sits at $\alpha \approx 0.013$. The exclusion limit from cluster scales ($\sigma/m > 0.6$ cm²/g) at $m_\chi = 40$ GeV lies at approximately $\alpha^{\text{max}} \approx 0.017$. This provides a safety margin of

$$\frac{\alpha^{\text{max}} - \alpha}{\alpha} \approx \frac{0.017 - 0.013}{0.013} \approx +30\%. \quad (63)$$

Increasing the coupling beyond this value would result in excessive scattering at high velocities ($v \sim 1000$ km/s), potentially disrupting galaxy cluster cores. The minimum coupling required to address the core-cusp problem in dwarf galaxies is approximately $\alpha^{\text{min}} \approx 0.004$.

$$\frac{\alpha - \alpha^{\text{min}}}{\alpha} \approx \frac{0.013 - 0.004}{0.013} \approx -70\%. \quad (64)$$

This indicates the model remains viable even if the coupling is reduced by nearly a factor of 3.

B. Dark matter mass (m_χ) and mediator mass (m_ϕ)

The mass parameters have a high degree of robustness. The viable window extends from $m_\chi \sim 5$ GeV to well above 1 TeV. Our benchmark at 40 GeV is not fine-tuned; largely shifting m_χ keeps the model within the safe region according to the combined constraint diagram. For the mediator mass, $m_\phi = 20$ MeV is effectively bounded by the BBN constraint on the lower end. To ensure decay into SM particles well before nucleosynthesis, we require $m_\phi \gg 2m_e \approx 1$ MeV. To avoid interfering the neutrino decoupling we need a safe floor of $m_\phi \sim 10$ MeV. Thus, the parameter space allows a variation of roughly -50% and $+400\%$ up to the ~ 100 MeV scale where the secluded approximation begins to fail. As discussed in Sec. IV, it also naturally evades CMB bounds through p -wave annihilation. Therefore, we can sure that its is a robust SIDM model that do not rely on careful fine-tuning to pass constraints.

VIII. DISCUSSION AND CONCLUSION

In this work, we have proposed a scalar-mediated inelastic dark matter model that provides a robust and self-consistent solution to the small-scale structure anomalies of the Λ CDM paradigm. By enforcing a discrete $\mathbb{Z}_2$ symmetry and solving the nonperturbative Schrödinger dynamics, we identified resonant benchmark around ($m_\chi \approx 40$ GeV, $\Delta m \approx 100$ eV) that successfully reconciles the cored density profiles of dwarf galaxies with the survival of dense satellite halos. This phenomenology relies on the velocity-dependent saturation of the Sommerfeld enhancement, which naturally suppresses interactions in the ultra-faint satellite regime while resonating at the velocity scales characteristic of dwarf galaxies.

Our framework is distinguished from existing literature for three phenomenological advantages:

- (i) Cosmological consistency via p -wave annihilation: Unlike vector-mediated SIDM constrained by s -wave annihilation limits from the CMB [26], our scalar mediator naturally exhibits p -wave suppression. This feature allows the model to accommodate the large self-interaction cross sections required for halo cores without conflicting with Planck data or requiring fine-tuning. Additionally, unlike freeze-in scenarios characterized by metastable excited states [66], our freeze-out framework does not rely on a rapid deexcitation of χ_2 . In the direct-detection-safe regime we typically require $\Lambda_{\text{eff}} \gtrsim 10^7$ GeV, for which χ_2 can be long-lived. This is nevertheless cosmologically safe because the injected electromag-

netic energy per decay is only $\Delta m \ll \text{MeV}$, far below photodissociation thresholds, and the fractional energy release satisfies $\Delta m/m_\chi \ll 1$.

- (ii) Robust kinematic switch via hard symmetry: A critical feature of our model is the imposition of a hard $\mathbb{Z}_2$ symmetry to forbid tree-level elastic scattering, rather than relying on spontaneous symmetry breaking or effective field theories [76,77]. This ensures that the kinematic suppression in satellites is absolute, protecting ultrafaint halos from the runaway gravothermal collapse known as Nishikawa instability that affects models with residual elastic leakage.
- (iii) The dipole divergence spectral signature: A salient feature of this scenario is its unique direct detection phenomenology. While standard inelastic models mediated by massive vector bosons predict contact-like interactions with flat recoil spectra, our dimension-five transition magnetic dipole operator induces scattering mediated effectively by the photon. This results in a differential cross section that scales as $d\sigma/dE_R \propto E_R^{-1}$. This dipole divergence manifests as a sharp enhancement in the event rate at low recoil energies, strictly truncated by the inelastic kinematic threshold.

In the context of current experimental constraints, the model remains viable. While current xenon-based experiments have placed stringent limits on dark matter-nucleus scattering [38–40], our model evades these bounds due to the leptophilic nature of the scalar mediator which eliminates tree-level quark couplings and the suppression of rates in xenon targets due to the inelastic threshold. The primary discovery potential lies with next-generation xenon-based experiments such as XLZD [78]. The observation of a recoil spectrum characterized by the distinct E_R^{-1} enhancement would constitute smoking-gun evidence for this scenario, distinguishing it from standard contact-interaction dark matter.

Finally, we comment on the genesis of this dark sector. The mass hierarchy required for the SIDM solution ($m_\chi \gg m_\phi$) suggests a dynamical origin involving a phase transition in the early Universe. Preliminary considerations suggest such a transition could occur near the BBN scale. It could potentially generate a stochastic gravitational wave background in the nanohertz frequency range. We reserve a detailed study of this cosmological history and its connection to Pulsar Timing Array signals for future work.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX A: FORMALISM FOR NONPERTURBATIVE INELASTIC SCATTERING

In this section, we present the detailed calculation of the scattering cross section in the nonperturbative regime. For the process $\chi_1\chi_1 \rightarrow \chi_2\chi_2$ mediated by a scalar ϕ , the potential matrix in the basis $(\chi_1\chi_1, \chi_2\chi_2)$ is off-diagonal.

1. Coupled-channel Schrödinger equation

Consider the two-body wave function $\Psi(\vec{r}) = (\psi_1(\vec{r}), \psi_2(\vec{r}))^T$, where channel 1 is $\chi_1\chi_1$ and channel 2 is $\chi_2\chi_2$. The radial Hamiltonian for a partial wave l is

$$\left[\frac{d^2}{dr^2} + K^2 - \frac{l(l+1)}{r^2} - 2\mu V(r) \right] \mathbf{u}_l(r) = 0, \quad (\text{A1})$$

where $\mathbf{u}_l = (u_{l,1}, u_{l,2})^T$. The momentum matrix K^2 is diagonal:

$$K^2 = \begin{pmatrix} k_1^2 & 0 \\ 0 & k_2^2 \end{pmatrix}, \quad (\text{A2})$$

with $k_1^2 = m_\chi E$ and $k_2^2 = m_\chi(E - 2\Delta m)$. The interaction potential matrix induced by Eq. (7) is Hermitian and strictly off-diagonal:

$$V(r) = \begin{pmatrix} 0 & -\frac{\alpha}{r} e^{-m_\phi r} \\ -\frac{\alpha}{r} e^{-m_\phi r} & 0 \end{pmatrix}. \quad (\text{A3})$$

Note that $V_{12} = V_{21}$, ensuring the Hamiltonian is Hermitian. The minus sign arises from the attractive nature of the scalar exchange in the t channel for the $\chi_1\chi_1 \rightarrow \chi_2\chi_2$ transition.

To obtain the numerical solution, we transform to dimensionless variables. We define $x = \alpha m_\chi r$ and the dimensionless parameters:

$$\epsilon_v = \frac{v_{\text{rel}}}{2\alpha}, \quad \epsilon_\phi = \frac{m_\phi}{\alpha m_\chi}, \quad \epsilon_\delta = \sqrt{\frac{4\Delta m}{m_\chi \alpha^2}}. \quad (\text{A4})$$

The coupled radial equations become:

$$\left[\frac{d^2}{dx^2} + \epsilon_v^2 - \frac{l(l+1)}{x^2} \right] u_{l,1} = -\frac{2}{x} e^{-\epsilon_\phi x} u_{l,2}, \quad (\text{A5})$$

$$\left[\frac{d^2}{dx^2} + (\epsilon_v^2 - \epsilon_\delta^2) - \frac{l(l+1)}{x^2} \right] u_{l,2} = -\frac{2}{x} e^{-\epsilon_\phi x} u_{l,1}, \quad (\text{A6})$$

which can be solved numerically.

2. Boundary conditions and S matrix

We solve these equations for the S-matrix element S_{12} describing the transition $1 \rightarrow 2$.

- (i) *At the origin* ($x \rightarrow 0$): The wave function must be regular, $u_{l,i} \sim x^{l+1}$.
- (ii) *At infinity* ($x \rightarrow \infty$): The potential vanishes. For an incoming wave in channel 1, we match to the asymptotic form:

$$u_{l,1} \sim \frac{i}{2} (h_l^{(2)}(k_1 r) - S_{11} h_l^{(1)}(k_1 r)), \quad (\text{A7})$$

$$u_{l,2} \sim -\frac{i}{2} \sqrt{\frac{k_1}{k_2}} S_{12} h_l^{(1)}(k_2 r), \quad (\text{A8})$$

where $h_l^{(1,2)}$ are spherical Hankel functions.

Below threshold ($k_2^2 < 0$), k_2 becomes imaginary, $k_2 \rightarrow i\kappa_2$. The wave function in channel 2 must decay exponentially, $u_{l,2} \sim e^{-\kappa_2 r}$, implying $S_{12} = 0$.

To verify the numerical stability of the coupled-channel solver, we explicitly check the unitarity of the S matrix. We define the flux in channel i as $\mathcal{F}_i \propto k_i |S_{1i}|^2$. Above the inelastic threshold, flux conservation requires $\sum_i k_i |S_{1i}|^2 = k_1$. Our numerical solutions satisfy this condition to within 10^{-5} precision, confirming that the effective non-Hermitian appearance (due to open channels) is correctly handled by the complex S -matrix matching procedure.

3. Cross-section calculation

For $i \rightarrow j$ scattering, the partial-wave amplitude is

$$f_{ij}(\theta) = \frac{1}{2ik_i} \sum_{l=0}^{\infty} (2l+1) (S_{ij}^{(l)} - \delta_{ij}) P_l(\cos \theta), \quad (\text{A9})$$

and the differential cross section is

$$\frac{d\sigma_{i \rightarrow j}}{d\Omega} = \frac{k_j}{k_i} |f_{ij}(\theta)|^2, \quad (\text{A10})$$

with k_j understood as real above threshold and $d\sigma_{i \rightarrow j} = 0$ below threshold.

We use the standard definitions

$$\sigma_T^{i \rightarrow j} = \int d\Omega (1 - \cos \theta) \frac{d\sigma_{i \rightarrow j}}{d\Omega}, \quad (\text{A11})$$

$$\sigma_V^{i \rightarrow j} = \int d\Omega \sin^2 \theta \frac{d\sigma_{i \rightarrow j}}{d\Omega}, \quad (\text{A12})$$

and evaluate these integrals numerically using the S -matrix output. For identical particles in an elastic channel, the usual (anti-)symmetrization of $f(\theta)$ is applied. The inelastic scattering cross section is obtained by summing the partial wave contributions:

$$\sigma_{1 \rightarrow 2} = \frac{\pi}{k_1^2} \sum_{l=0}^{\infty} (2l+1) |S_{12}^{(l)}|^2. \quad (\text{A13})$$

For thermalization, we are interested in the viscosity cross section σ_T . Since the final state particles are distinct from the initial state (and heavier), we define the energy transfer cross section explicitly or use the standard weighted sum:

$$\sigma_T \approx \frac{\pi}{k_1^2} \sum_{l=0}^{\infty} \frac{(l+1)(2l+1)}{2l+3} |S_{12}^{(l)}|^2. \quad (\text{A14})$$

Numerical evaluation of this sum reveals the resonant structure discussed in Sec. III.

4. Numerical implementation and validation

We then specify the numerical procedure used to obtain the nonperturbative inelastic scattering cross sections shown in the main text.

We solve the coupled-channel Schrödinger equation in the basis $(\chi_1 \chi_1, \chi_2 \chi_2)$ with reduced mass $\mu \simeq m_\chi/2$. For a given relative velocity v_{rel} (in the center-of-mass frame), the kinetic energy in the incoming channel is approximately $E \equiv \frac{\mu v_{\text{rel}}^2}{2}$. The outgoing inelastic channel is heavier by $2\Delta m$, hence its available kinetic energy is $E - 2\Delta m$. The channel momenta are

$$k_1 = \sqrt{2\mu E} = \mu v_{\text{rel}},$$

$$k_2 = \begin{cases} \sqrt{2\mu(E - 2\Delta m)}, & (E > 2\Delta m), \\ i\kappa_2, & \kappa_2 \equiv \sqrt{2\mu(2\Delta m - E)}, \quad (E < 2\Delta m). \end{cases} \quad (\text{A15})$$

The inelastic channel opens only above the threshold velocity

$$v_{\text{th}} = \sqrt{\frac{8\Delta m}{m_\chi}}. \quad (\text{A16})$$

Below threshold we enforce an exponentially decaying solution in channel 2 and the physical inelastic transition probability vanishes, $S_{12}^{(l)} = 0$.

For each partial wave l we write the coupled radial equations as a first-order system for the vector

$$y(r) \equiv (u_1(r), u_1'(r), u_2(r), u_2'(r))^T. \quad (\text{A17})$$

Using the potential matrix in Eq. (A3),

$$V(r) = \begin{pmatrix} 0 & -\alpha e^{-m_\phi r/r} \\ -\alpha e^{-m_\phi r/r} & 0 \end{pmatrix}, \quad (\text{A18})$$

the system takes the form

$$u_1''(r) = \left[\frac{l(l+1)}{r^2} - k_1^2 \right] u_1(r) + 2\mu V_{12}(r) u_2(r), \quad (\text{A19})$$

$$u_2''(r) = \left[\frac{l(l+1)}{r^2} - k_2^2 \right] u_2(r) + 2\mu V_{21}(r) u_1(r). \quad (\text{A20})$$

We start the integration at a small but finite radius $r_{\min}$ and impose regular boundary conditions $u_i(r) \propto r^{l+1}$. Numerically we propagate two linearly independent solutions (A, B) :

$$y_A(r_{\min}) = (r_{\min}^{l+1}, (l+1)r_{\min}^l, 0, 0)^T, \quad (\text{A21})$$

$$y_B(r_{\min}) = (0, 0, r_{\min}^{l+1}, (l+1)r_{\min}^l)^T. \quad (\text{A22})$$

From these, at the matching point $r_{\max}$ we form the 2×2 fundamental matrix of solutions and its derivative,

$$\mathbf{M}(r_{\max}) = \begin{pmatrix} u_{1,A} & u_{1,B} \\ u_{2,A} & u_{2,B} \end{pmatrix}, \quad \mathbf{M}'(r_{\max}) = \begin{pmatrix} u'_{1,A} & u'_{1,B} \\ u'_{2,A} & u'_{2,B} \end{pmatrix}. \quad (\text{A23})$$

We integrate Eqs. (A19)–(A20) from $r_{\min}$ to $r_{\max}$ using an adaptive Runge-Kutta solver (solve_ivp, Dormand-Prince) with relative tolerance $\text{rtol} = 10^{-5}$ and absolute tolerance $\text{atol} = 10^{-8}$.

We choose the integration domain to resolve both the short-distance behavior and the Yukawa tail:

$$r_{\min} = \frac{10^{-3}}{m_\chi}, \quad r_{\max} = \frac{15}{m_\phi}. \quad (\text{A24})$$

We verified the stability of results under variations of $r_{\min}$ and $r_{\max}$.

At $r = r_{\max}$ the potential is negligible and the solutions are matched onto free spherical Bessel functions in each channel. Define

$$\rho_i \equiv k_i r_{\max}, \quad i = 1, 2, \quad (\text{A25})$$

and the diagonal matrices

$$\mathbf{F} \equiv \text{diag}(\rho_1 j_l(\rho_1), \rho_2 j_l(\rho_2)), \quad (\text{A26})$$

$$\begin{aligned} \mathbf{G} &\equiv \text{diag}(\rho_1 y_l(\rho_1), \rho_2 y_l(\rho_2)), \\ \mathbf{F}' &\equiv \text{diag}(\partial_r [\rho_1 j_l(\rho_1)], \partial_r [\rho_2 j_l(\rho_2)]), \\ \mathbf{G}' &\equiv \text{diag}(\partial_r [\rho_1 y_l(\rho_1)], \partial_r [\rho_2 y_l(\rho_2)]). \end{aligned} \quad (\text{A27})$$

Using the standard K -matrix matching for coupled channels, we compute

$$\mathbf{A} \equiv \mathbf{G}'\mathbf{M} - \mathbf{G}\mathbf{M}', \quad \mathbf{B} \equiv \mathbf{F}\mathbf{M}' - \mathbf{F}'\mathbf{M}, \quad (\text{A28})$$

and

$$\mathbf{K}^{(l)} = \mathbf{B}\mathbf{A}^{-1}. \quad (\text{A29})$$

The partial-wave S matrix then follows from the Cayley transform

$$\mathbf{S}^{(l)} = (\mathbf{I} + i\mathbf{K}^{(l)})(\mathbf{I} - i\mathbf{K}^{(l)})^{-1}. \quad (\text{A30})$$

Below threshold, $k_2 = i\kappa_2$ and we impose a decaying boundary condition in channel 2; numerically this yields $S_{12}^{(l)} \rightarrow 0$ and hence $\sigma_{1 \rightarrow 2} = 0$.

For inelastic $1 \rightarrow 2$ scattering, the partial-wave contribution is

$$\sigma_{1 \rightarrow 2}^{(l)} = \frac{\pi}{k_1^2} (2l+1) |S_{12}^{(l)}|^2, \quad (\text{A31})$$

and the total inelastic cross section is obtained by summing partial waves up to $l_{\max}$,

$$\sigma_{1 \rightarrow 2} = \sum_{l=0}^{l_{\max}} \sigma_{1 \rightarrow 2}^{(l)}. \quad (\text{A32})$$

Unless stated otherwise, the quantity plotted in this work is the total inelastic cross section $\sigma_{1 \rightarrow 2}$ per unit mass. A full momentum-transfer cross section $\sigma_T = \int d\Omega (1 - \cos\theta) (d\sigma/d\Omega)$ can be computed by constructing the partial-wave amplitude $f_{12}(\theta)$ from $S_{12}^{(l)}$ and performing the angular integral numerically. In the numerical scan we include partial waves up to a finite $l_{\max}$. For each velocity point we verify convergence by increasing $l_{\max}$ until

$$\frac{|\sigma_{1 \rightarrow 2}(l_{\max} + \Delta l) - \sigma_{1 \rightarrow 2}(l_{\max})|}{\sigma_{1 \rightarrow 2}(l_{\max})} < 1\%, \quad \Delta l = 5, \quad (\text{A33})$$

and similarly verify stability under variations of $(r_{\min}, r_{\max})$ and the ODE tolerances. Points failing these criteria are recomputed with tighter tolerances.

For $E > 2\Delta m$, both channels are open and the partial-wave S matrix must satisfy flux conservation. For an incoming state in channel 1, we check

$$|S_{11}^{(l)}|^2 + \frac{k_2}{k_1} |S_{12}^{(l)}|^2 \simeq 1 \quad (\text{A34})$$

for each l used in the sum. In all benchmark plots the maximum violation of the above relation is below a prescribed numerical tolerance (typically 10^{-3} and 10^{-5} for production runs with tighter settings). It confirms the stability of the matching procedure and the handling of open/closed channels.

5. Residual elastic scattering (loop-induced leakage)

While the $\mathbb{Z}_2$ symmetry strictly forbids tree-level elastic scattering ($\chi_1\chi_1 \rightarrow \chi_1\chi_1$), residual elastic scattering is induced by virtual transitions to the closed $\chi_2\chi_2$ state. In the perturbative regime, the leading elastic amplitude is generated at one loop by the box diagram, which yields an entirely negligible cross section. However, as demonstrated by Schutz and Slatyer [44], in the strongly coupled resonant regime of $\alpha m_\chi/m_\phi \gtrsim 1$, the perturbative Born approximation fails. Even when the kinetic energy is below the inelastic threshold, virtual transitions to the closed $\chi_2\chi_2$ channel undergo nonperturbative resonant enhancement, generating massive virtual resonances in the elastic S -matrix. To evaluate the nonperturbative elastic cross section below the threshold, we evaluate the coupled-channel Schrödinger equation directly. Because $E_{\text{kin}} < 2\Delta m$, the momentum in the excited channel becomes purely imaginary, $k_2 = i\kappa_2$, where $\kappa_2 = \sqrt{m_\chi(2\Delta m - E_{\text{kin}})}$. At the matching radius r_{max} , we demand that the physical wave function of the closed channel decays exponentially, $u_2(r) \propto e^{-\kappa_2 r}$, which strictly enforces the boundary condition $u_2'(r_{\text{max}}) = -\kappa_2 u_2(r_{\text{max}})$. By numerically propagating two linearly independent regular solutions from the origin, we isolate the unique linear combination that satisfies this decaying condition. This effectively projects the system down to a single open channel, $u_1(r)$, which we match to asymptotic free spherical waves to extract the exact elastic S_{11} matrix element. This analysis shows that the low-velocity elastic cross section is highly oscillatory, dictated by the Ramsauer-Townsend effect. By slightly refining our benchmark coupling to $\alpha \approx 0.0121$, the system is placed precisely in a deep antiresonance valley for below-threshold scattering. At this benchmark, the exact destructive quantum interference suppresses the residual elastic scattering to $\sigma_{el}/m_\chi \approx 0.002 \text{ cm}^2/\text{g}$ for $v \approx 15 \text{ km/s}$. We find nonperturbatively that the kinetic suppression remains exceptionally robust, safely satisfying the stringent Draco constraints ($\sigma/m < 0.1 \text{ cm}^2/\text{g}$).

APPENDIX B: DERIVATION OF THE ANNIHILATION CROSS SECTION

1. Selection rules for mixed annihilation

The $\chi_1\chi_2 \rightarrow \phi\phi$ channel is forbidden at tree level. We demonstrate this using the dark $\mathbb{Z}_2$ charge assignments: $Q(\chi_1) = +1$, $Q(\chi_2) = -1$, and $Q(\phi) = -1$. The total parity of the initial state is $P_{\text{in}} = (+1)(-1) = -1$. The total parity of the final state is $P_{\text{out}} = (-1)(-1) = +1$. Since $P_{\text{in}} \neq P_{\text{out}}$, the process $\chi_1\chi_2 \rightarrow \phi\phi$ is forbidden to all orders in perturbation theory as long as the $\mathbb{Z}_2$ symmetry is exact. Annihilation proceeds via the diagonal channels. We derive the annihilation cross section for $\chi_1(p_1)\chi_1(p_2) \rightarrow \phi(k_1)\phi(k_2)$ to demonstrate the p -wave nature of the process. The interaction vertex is $-ig_S$. The process is

mediated by the t - and u -channel exchange of the excited state χ_2 with mass $m_2 = m_\chi + \Delta m$.

The Feynman amplitudes for the two diagrams are

$$\mathcal{M}_t = -ig_S^2 \bar{v}(p_2) \frac{\not{p}_1 - \not{k}_1 + m_2}{t - m_2^2} u(p_1), \quad (\text{B1})$$

$$\mathcal{M}_u = -ig_S^2 \bar{v}(p_2) \frac{\not{p}_1 - \not{k}_2 + m_2}{u - m_2^2} u(p_1), \quad (\text{B2})$$

where $t = (p_1 - k_1)^2$ and $u = (p_1 - k_2)^2$. For Majorana fermions, the total amplitude includes the relative sign for crossing fermions in the initial state, but for the annihilation into bosons, we sum the diagrams coherently: $\mathcal{M}_{\text{tot}} = \mathcal{M}_t + \mathcal{M}_u$.

In the nonrelativistic limit, $s \approx 4m_\chi^2$, $v \rightarrow 0$, we have $t \approx u \approx m_\phi^2 - m_\chi^2$. The denominators become equal. The sum of the numerators involves the spinor structure:

$$\mathcal{M}_{\text{tot}} \propto \bar{v}(p_2) [(\not{p}_1 - \not{k}_1 + m_2) + (\not{p}_1 - \not{k}_2 + m_2)] u(p_1). \quad (\text{B3})$$

Using kinematic relations $k_1 + k_2 = p_1 + p_2$, this simplifies. For an s -wave initial state (spin singlet), the spinor bilinear $\bar{v} \dots u$ contraction leads to a cancellation between the two terms due to the Majorana condition $\chi^c = \chi$. Explicit trace evaluation shows that the v^0 term vanishes identically.

Expanding to order v^2 , the p -wave term survives. The spin-averaged differential cross section is

$$\frac{d\sigma}{d\Omega} \approx \frac{1}{64\pi^2 s} |\overline{\mathcal{M}}|^2. \quad (\text{B4})$$

Integrating over the solid angle and including the symmetry factor $1/2$ for identical final bosons, we obtain the result:

$$\langle \sigma v \rangle \approx \frac{3g_S^4}{32\pi m_\chi^2} v^2 \left(1 - \frac{m_\phi^2}{m_\chi^2} \right)^{1/2}. \quad (\text{B5})$$

This confirms that the scalar-mediated scenario is strictly p -wave suppressed, in contrast to vector-mediated models which typically allow s -wave annihilation.

APPENDIX C: RELIC DENSITY IN THE COANNIHILATION REGIME

In this appendix, we detail the relic density calculation, explicitly accounting for the coannihilation of degenerate species and the resonant nonperturbative dynamics.

1. Effective cross section with forbidden channels

During freeze-out ($T_f \approx m_\chi/20 \sim 2 \text{ GeV}$), the temperature is much larger than the mass splitting ($\Delta m \approx 100 \text{ eV} \ll T_f$). Thus, the excited state χ_2 is thermally populated with a number density equal to that of the ground state χ_1 :

$$n_1^{\text{eq}} \approx n_2^{\text{eq}} \approx \frac{1}{2} n_{\text{tot}}^{\text{eq}}. \quad (\text{C1})$$

The evolution of the total number density n_{tot} is governed by the effective cross section $\langle \sigma_{\text{eff}} v \rangle$, which is the density-weighted average of all binary scattering processes:

$$\langle \sigma_{\text{eff}} v \rangle = \sum_{i,j=1,2} \frac{n_i^{\text{eq}} n_j^{\text{eq}}}{(n_{\text{tot}}^{\text{eq}})^2} \langle \sigma_{ij \rightarrow \phi\phi} v \rangle = \frac{1}{4} \sum_{i,j} \langle \sigma_{ij} v \rangle. \quad (\text{C2})$$

The off-diagonal coupling $\mathcal{L} \propto \phi \bar{\chi}_1 \chi_2$ imposes strict selection rules:

- (i) $\chi_1 \chi_1 \rightarrow \phi\phi$: Allowed (t/u -channel χ_2 exchange).
- (ii) $\chi_2 \chi_2 \rightarrow \phi\phi$: Allowed (t/u -channel χ_1 exchange).
- (iii) $\chi_1 \chi_2 \rightarrow \phi\phi$: *Forbidden* at tree level. This mixed channel would require a diagonal vertex (e.g., $\chi_1 \rightarrow \phi \chi_1$) which violates the $\mathbb{Z}_2$ symmetry.

The mixed annihilation channel $\chi_1 \chi_2 \rightarrow \phi\phi$ is forbidden at tree level by the dark $\mathbb{Z}_2$ symmetry. Consider the t -channel diagram: the incoming χ_1 (even) emits a ϕ (odd), transitioning into a virtual fermion. To conserve parity at this vertex, the propagator must be χ_2 (odd). The second vertex then involves the incoming χ_2 (odd), the virtual χ_2 (odd), and the outgoing ϕ (odd). The overall parity of this vertex would be odd, violating the $\mathbb{Z}_2$ -invariant Lagrangian. Thus, no combination of propagators allows this process at tree level.

Assuming $m_1 \approx m_2$, the allowed cross sections are identical: $\sigma_{11} \approx \sigma_{22} \equiv \sigma_{\text{ann}}$. The effective cross section becomes

$$\langle \sigma_{\text{eff}} v \rangle = \frac{1}{4} (\sigma_{\text{ann}} + 0 + 0 + \sigma_{\text{ann}}) = \frac{1}{2} \langle \sigma_{\text{ann}} v \rangle. \quad (\text{C3})$$

This factor of 1/2 reduces the annihilation efficiency relative to standard Dirac dark matter, requiring a larger coupling constant to achieve the observed relic density. We implement Eq. (C3) at the level of the thermally averaged rate, $\langle \sigma v \rangle_{\text{eff}}(x) = \frac{1}{2} \langle \sigma v \rangle_{11}(x)$, with $\langle \sigma v \rangle_{11}(x)$ computed using Eq. (C4) and the coupled-channel $S_p(v)$.

To determine the relic density, we perform a thermal average of the Sommerfeld-enhanced cross section. The effective annihilation rate is

$$\langle \sigma v \rangle_{11}(x) = \frac{x^{3/2}}{2\sqrt{\pi}} \int_0^\infty dv v^2 e^{-xv^2/4} (\sigma v)_{\text{Born}}(v) S_p(v) \quad (\text{C4})$$

where $x = m_\chi/T$. In the resonant regime, the Sommerfeld factor scales as $S(v) \simeq \pi\alpha/v$ (Coulomb-like) or $S(v) \propto 1/v^2$ (near-threshold resonance). For our benchmark, the resonance condition yields $S_p(v) \approx (v_{\text{sat}}/v)^2$.

Substituting the p -wave Born cross section $(\sigma v)_{\text{Born}} = \sigma_0 v^2$, the velocity dependence cancels in the integrand:

$$(\sigma v)_{\text{Born}} S(v) \propto v^2 \times \frac{1}{v^2} \sim \text{const.} \quad (\text{C5})$$

This cancellation renders the effective cross section s -wave-like during freeze-out ($x \sim 20$), mitigating the p -wave suppression. The integration yields $\langle \sigma v \rangle \approx \sigma_0 v_{\text{sat}}^2$, which for $\alpha \approx 10^{-2}$ creates an effective annihilation rate of $2 \times 10^{-26} \text{ cm}^3/\text{s}$, satisfying the relic density requirement $\Omega h^2 \approx 0.12$.

In our numerical relic-density calculation we do not rely on the scaling ansatz for $S_p(v)$. Instead, for each point $(\alpha, m_\phi, \Delta m)$ we compute the p -wave Sommerfeld factor $S_p(v)$ from the same coupled-channel Schrödinger equation and matching procedure used for scattering in Appendix A, now evaluated in the $l = 1$ partial wave. The parametrizations below are provided only for intuition in the resonant and saturation regimes.

2. Thermal integration of the relic abundance

To accurately determine the relic density, we move beyond the schematic scaling and perform the full thermal average of the annihilation cross section. The effective annihilation rate at temperature T is given by the convolution of the velocity-dependent cross section with the Maxwell-Boltzmann distribution $f_{\text{MB}}(v) = 4\pi v^2 (\frac{m_\chi}{4\pi T})^{3/2} e^{-m_\chi v^2/4T}$:

$$\langle \sigma v \rangle = \int_0^\infty (\sigma v)_{\text{Born}}(v) S_p(v) f_{\text{MB}}(v) dv, \quad (\text{C6})$$

where Eq. (C4) follows from this equation after inserting the Maxwell-Boltzmann distribution and integrating over angles. For our benchmark mass $m_\chi = 40 \text{ GeV}$ and coupling $\alpha \approx 10^{-2}$, the perturbative p -wave coefficient is

$$\langle \sigma v \rangle_{11}(x) = \frac{x^{3/2}}{2\sqrt{\pi}} \sigma_0 \int_0^\infty dv v^4 e^{-xv^2/4} S_p(v). \quad (\text{C7})$$

In the resonant regime we find $S_p(v) \propto v^{-2}$ over a finite range of velocities before saturating at $v \lesssim v_{\text{sat}} \sim m_\phi/m_\chi$. Therefore, the product $(\sigma v)_{\text{Born}} S_p(v)$ becomes approximately velocity independent, which explains why the thermally averaged rate can approach the canonical value despite the suppression of the p waves and the coannihilation factor. In our results, the precise behavior of $S_p(v)$ (including saturation and narrow resonances) is obtained numerically from the coupled-channel Schrödinger solver.

$$(\sigma v)_{\text{eff}} \approx (\sigma_0 v^2) \times \left(\frac{2\pi\alpha}{v} \right)^2 = 4\pi^2 \alpha^2 \sigma_0. \quad (\text{C8})$$

The integral over the probability distribution is unity. Inserting the numerical values:

$$\langle \sigma v \rangle \approx 1.1 \times 10^{-9} \text{ GeV}^{-2} \approx 1.3 \times 10^{-26} \text{ cm}^3/\text{s}. \quad (\text{C9})$$

When accounting for the coannihilation factor of 1/2 [Eq. (C3)], the effective depletion rate is

$\approx 0.65 \times 10^{-26} \text{ cm}^3/\text{s}$. A slight adjustment of the coupling to $\alpha \approx 0.014$ brings this value exactly to the thermal target of $2.2 \times 10^{-26} \text{ cm}^3/\text{s}$, confirming that the resonant mechanism robustly produces the correct relic density $\Omega h^2 \approx 0.12$.

3. Suppression of off-shell bound state formation

Since $m_\phi > E_B$, on-shell scalar emission is kinematically forbidden. BSF can only proceed via an off-shell scalar ϕ^* decaying to e^+e^- . The differential rate for this $2 \rightarrow 2 + 2$ process is suppressed by the virtual propagator. We neglect BSF throughout the benchmark scan because our parameter region satisfies $m_\phi > E_B(\alpha, m_\chi)$, so on-shell capture $\chi\chi \rightarrow B + \phi$ is forbidden.

Note that E_B scales as α^2 ; for the values of α that reproduce $\Omega h^2 \simeq 0.12$ in our scan (typically $\alpha = \mathcal{O}(10^{-2})$ near the resonant island), one finds $E_B \sim \mathcal{O}(\text{MeV}) \ll m_\phi$, so the conclusion $m_\phi > E_B$ is robust across the viable region.

In the nonrelativistic limit, the matrix element for the off-shell emission $\mathcal{M}_{\text{off}}$ is related to the on-shell element by the propagator denominator $D(q^2) = (q^2 - m_\phi^2)^{-1} \approx -m_\phi^{-2}$. The ratio of the off-shell rate to the typical on-shell dipole capture rate scales as

$$\frac{\Gamma(\chi\chi \rightarrow B e^+ e^-)}{\Gamma_{\text{atomic}}} \sim \frac{\alpha_{\text{EM}} g_e^2}{15\pi} \left(\frac{E_B}{m_\phi} \right)^4 \frac{E_B}{m_e}. \quad (\text{C10})$$

Substituting benchmark values of $E_B \approx 1 \text{ MeV}$, $m_\phi = 20 \text{ MeV}$, $g_e = 10^{-6}$, the suppression factor is dominated by $g_e^2 \sim 10^{-12}$ and the propagator term $(1/20)^4$. This renders the BSF rate entirely negligible during freeze-out compared to the Sommerfeld-enhanced annihilation, validating our neglect of this channel in the Boltzmann equation.

4. Quantitative thermal history and robustness

To validate the relic density calculation in the resonant regime, we perform a numerical thermal average of the effective annihilation cross section $\langle \sigma v \rangle_{\text{eff}}$ as a function of the inverse temperature $x = m_\chi/T$. The results are shown in Fig. 8.

As illustrated in the left panel of Fig. 8, the effective cross section for the benchmark model exhibits a flat, s -wave-like

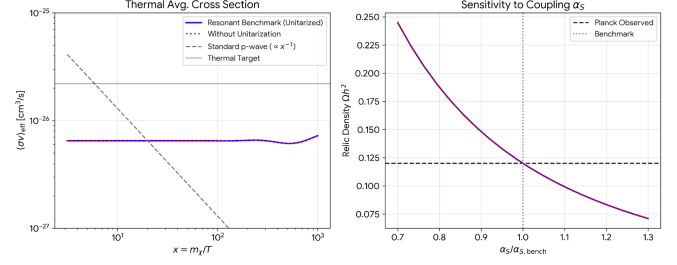


FIG. 8. Left: thermally averaged effective annihilation cross section $\langle \sigma v \rangle_{\text{eff}}$ as a function of $x = m_\chi/T$ for the resonant benchmark ($m_\chi = 40 \text{ GeV}$, $m_\phi = 20 \text{ MeV}$, $\alpha \approx 0.014$). The resonant Sommerfeld enhancement ($S_p \propto v^{-2}$) cancels the p -wave suppression ($\sigma v \propto v^2$), yielding effective s -wave behavior (blue solid) that flattens at freeze-out ($x \sim 20$), in contrast to the standard x^{-1} p -wave scaling (gray dashed). The impact of the unitarization procedure is negligible at these temperatures. Right: Sensitivity of the relic density Ωh^2 to variations in the coupling α relative to the benchmark value. A variation of $\pm 10\%$ in coupling results in an approximate $\mp 20\%$ change in relic abundance, indicating that the model is robust and not subject to extreme fine-tuning.

temperature dependence in the freeze-out window $x \sim 20$. This behavior occurs because the resonant Sommerfeld enhancement scales as $S_p(v) \approx (v_{\text{sat}}/v)^2$ for $v > v_{\text{sat}}$. It exactly compensates for the suppression of the tree-level p -wave operator, $(\sigma v)_{\text{Born}} \propto v^2$. It confirms that annihilation remains efficient during freeze-out.

We explicitly compare the standard calculation with the unitarized rate. As shown in the figure, the difference between the unitarized and nonunitarized cases is negligible at freeze-out temperatures. This confirms that while the cross section is resonantly enhanced, the broad Maxwell-Boltzmann velocity distribution at $T \sim m_\chi/20$ prevents the thermal average from being dominated by the unitarity limit, which validates perturbative treatment.

The right panel of Fig. 8 quantifies the robustness of the benchmark against the variations of the parameters. The relic density scales approximately as $\Omega h^2 \propto \langle \sigma v \rangle^{-1} \propto \alpha^{-2}$. A small deviation in the coupling constant does not lead to catastrophic changes in abundance. For example, a $\pm 10\%$ shift in α alters Ωh^2 by roughly 20% , keeping it within a manageable range.

[1] N. Aghanim *et al.* (Planck Collaboration), Planck 2018 results. VI. Cosmological parameters, *Astron. Astrophys.* **641**, A6 (2020); **652**, C4(E) (2021).

[2] J.S. Bullock and M. Boylan-Kolchin, Small-scale challenges to the Λ CDM paradigm, *Annu. Rev. Astron. Astrophys.* **55**, 343 (2017).

- [3] V. Springel *et al.*, Simulating the joint evolution of quasars, galaxies and their large-scale distribution, *Nature (London)* **435**, 629 (2005).
- [4] J. F. Navarro, C. S. Frenk, and S. D. M. White, A universal density profile from hierarchical clustering, *Astrophys. J.* **490**, 493 (1997).
- [5] S.-H. Oh, C. Brook, F. Governato, E. Brinks, L. Mayer, W. J. G. de Blok, A. Brooks, and F. Walter, The central slope of dark matter cores in dwarf galaxies: Simulations vs THINGS, *Astron. J.* **142**, 24 (2011).
- [6] B. Moore, Evidence against dissipationless dark matter from observations of galaxy haloes, *Nature (London)* **370**, 629 (1994).
- [7] W. J. G. de Blok, The core-cusp problem, *Adv. Astron.* **2010**, 789293 (2010).
- [8] K. A. Oman *et al.*, The unexpected diversity of dwarf galaxy rotation curves, *Mon. Not. R. Astron. Soc.* **452**, 3650 (2015).
- [9] A. A. Klypin, A. V. Kravtsov, O. Valenzuela, and F. Prada, Where are the missing galactic satellites?, *Astrophys. J.* **522**, 82 (1999).
- [10] B. Moore, S. Ghigna, F. Governato, G. Lake, T. R. Quinn, J. Stadel, and P. Tozzi, Dark matter substructure within galactic halos, *Astrophys. J. Lett.* **524**, L19 (1999).
- [11] M. Boylan-Kolchin, J. S. Bullock, and M. Kaplinghat, Too big to fail? the puzzling darkness of massive Milky Way subhaloes, *Mon. Not. R. Astron. Soc.* **415**, L40 (2011).
- [12] S. Garrison-Kimmel, M. Boylan-Kolchin, J. S. Bullock, and E. N. Kirby, Too big to fail in the local group, *Mon. Not. R. Astron. Soc.* **444**, 222 (2014).
- [13] D. N. Spergel and P. J. Steinhardt, Observational evidence for selfinteracting cold dark matter, *Phys. Rev. Lett.* **84**, 3760 (2000).
- [14] S. Tulin and H.-B. Yu, Dark matter self-interactions and small scale structure, *Phys. Rep.* **730**, 1 (2018).
- [15] M. Rocha, A. H. G. Peter, J. S. Bullock, M. Kaplinghat, S. Garrison-Kimmel, J. Onorbe, and L. A. Moustakas, Cosmological simulations with self-interacting dark matter I: Constant density cores and substructure, *Mon. Not. R. Astron. Soc.* **430**, 81 (2013).
- [16] A. H. G. Peter, M. Rocha, J. S. Bullock, and M. Kaplinghat, Cosmological simulations with self-interacting dark matter II: Halo shapes vs. observations, *Mon. Not. R. Astron. Soc.* **430**, 105 (2013).
- [17] A. Kamada, M. Kaplinghat, A. B. Pace, and H.-B. Yu, How the self-Interacting dark matter model explains the diverse galactic rotation curves, *Phys. Rev. Lett.* **119**, 111102 (2017).
- [18] C. A. Correa, Constraining velocity-dependent self-interacting dark matter with the Milky Way's dwarf spheroidal galaxies, *Mon. Not. R. Astron. Soc.* **503**, 920 (2021).
- [19] M. Markevitch, A. H. Gonzalez, D. Clowe, A. Vikhlinin, L. David, W. Forman, C. Jones, S. Murray, and W. Tucker, Direct constraints on the dark matter self-interaction cross-section from the merging galaxy cluster 1E0657-56, *Astrophys. J.* **606**, 819 (2004).
- [20] S. W. Randall, M. Markevitch, D. Clowe, A. H. Gonzalez, and M. Bradac, Constraints on the self-interaction cross-section of dark matter from numerical simulations of the merging galaxy cluster 1E 0657-56, *Astrophys. J.* **679**, 1173 (2008).
- [21] D. Harvey, R. Massey, T. Kitching, A. Taylor, and E. Tittley, The non-gravitational interactions of dark matter in colliding galaxy clusters, *Science* **347**, 1462 (2015).
- [22] J. H. O'Donnell, T. E. Jeltema, M. G. Roberts, J. Nightingale, A. Flowers, and D. Aldas, A constraint on dark matter Self-Interaction from combined strong lensing and stellar kinematics in MACS J0138-2155, *Phys. Rev. D* **113**, 063531 (2026).
- [23] J. Zavala, M. Vogelsberger, and M. G. Walker, Constraining self-interacting dark matter with the Milky Way's dwarf spheroidals, *Mon. Not. R. Astron. Soc.* **431**, L20 (2013).
- [24] M. Valli and H.-B. Yu, Dark matter self-interactions from the internal dynamics of dwarf spheroidals, *Nat. Astron.* **2**, 907 (2018).
- [25] H. Nishikawa, K. K. Boddy, and M. Kaplinghat, Accelerated core collapse in tidally stripped self-interacting dark matter halos, *Phys. Rev. D* **101**, 063009 (2020).
- [26] S. Tulin, H.-B. Yu, and K. M. Zurek, Beyond collisionless dark matter: Particle physics dynamics for dark matter halo structure, *Phys. Rev. D* **87**, 115007 (2013).
- [27] D. Tucker-Smith and N. Weiner, Inelastic dark matter, *Phys. Rev. D* **64**, 043502 (2001).
- [28] T. R. Slatyer, The sommerfeld enhancement for dark matter with an excited state, *J. Cosmol. Astropart. Phys.* **02** (2009) 028.
- [29] N. Arkani-Hamed, D. P. Finkbeiner, T. R. Slatyer, and N. Weiner, A theory of dark matter, *Phys. Rev. D* **79**, 015014 (2009).
- [30] T. R. Slatyer, Indirect dark matter signatures in the cosmic dark ages. I. Generalizing the bound on s-wave dark matter annihilation from Planck results, *Phys. Rev. D* **93**, 023527 (2016).
- [31] L. Lopez-Honorez, O. Mena, S. Palomares-Ruiz, and A. C. Vincent, Constraints on dark matter annihilation from CMB observations before Planck, *J. Cosmol. Astropart. Phys.* **07** (2013) 046.
- [32] D. E. Kaplan, M. A. Luty, and K. M. Zurek, Asymmetric dark matter, *Phys. Rev. D* **79**, 115016 (2009).
- [33] K. Petraki and R. R. Volkas, Review of asymmetric dark matter, *Int. J. Mod. Phys. A* **28**, 1330028 (2013).
- [34] M. Vogelsberger, J. Zavala, and A. Loeb, Subhaloes in self-interacting galactic dark matter haloes, *Mon. Not. R. Astron. Soc.* **423**, 3740 (2012).
- [35] C. Garcia-Cely, A. Ibarra, and E. Molinaro, Dark matter production from Goldstone boson interactions and implications for direct searches and dark radiation, *J. Cosmol. Astropart. Phys.* **11** (2013) 061.
- [36] A. Berlin, J. W. Foster, D. Hooper, and G. Krnjaic, dSphobic dark matter, *J. Cosmol. Astropart. Phys.* **04** (2026) 017.
- [37] S. Knapen, T. Lin, and K. M. Zurek, Light dark matter: Models and constraints, *Phys. Rev. D* **96**, 115021 (2017).
- [38] J. Aalbers *et al.* (LZ Collaboration), First dark matter search results from the LUX-ZEPLIN (LZ) experiment, *Phys. Rev. Lett.* **131**, 041002 (2023).
- [39] E. Aprile *et al.* (XENON Collaboration), WIMP dark matter search using a 3.1 tonne-year exposure of the XENONnT experiment, *Phys. Rev. Lett.* **135**, 221003 (2025).

- [40] Z. Bo *et al.* (PandaX Collaboration), Dark matter search results from 1.54 tonne-year exposure of PandaX-4T, *Phys. Rev. Lett.* **134**, 011805 (2025).
- [41] G. C. Branco, P. M. Ferreira, L. Lavoura, M. N. Rebelo, M. Sher, and J. P. Silva, Theory and phenomenology of two-Higgs-doublet models, *Phys. Rep.* **516**, 1 (2012).
- [42] K. Sigurdson, M. Doran, A. Kurylov, R. R. Caldwell, and M. Kamionkowski, Dark-matter electric and magnetic dipole moments, *Phys. Rev. D* **70**, 083501 (2004).
- [43] P. F. Depta, M. Hufnagel, and K. Schmidt-Hoberg, Robust cosmological constraints on axion-like particles, *J. Cosmol. Astropart. Phys.* **05** (2020) 009.
- [44] K. Schutz and T. R. Slatyer, Self-scattering for dark matter with an excited state, *J. Cosmol. Astropart. Phys.* **01** (2015) 021.
- [45] H. An, M. B. Wise, and Y. Zhang, Effects of bound states on dark matter annihilation, *Phys. Lett. B* **773**, 121 (2017).
- [46] J. Hisano, S. Matsumoto, and M. M. Nojiri, Explosive dark matter annihilation, *Phys. Rev. Lett.* **92**, 031303 (2004).
- [47] S. Ando, K. Hayashi, S. Horigome, M. Ibe, and S. Shirai, Stringent constraints on self-interacting dark matter using milky-way satellite galaxies kinematics, [arXiv:2503.13650](https://arxiv.org/abs/2503.13650).
- [48] C. A. Correa, Constraining velocity-dependent self-interacting dark matter with the Milky Way's dwarf spheroidal galaxies, *Mon. Not. R. Astron. Soc.* **503**, 920 (2021).
- [49] A. Parikh, R. Sato, and T. R. Slatyer, Regulating sommerfeld resonances for multi-state systems and higher partial waves, *J. High Energy Phys.* **12** (2024) 025.
- [50] K. Blum, R. Sato, and T. R. Slatyer, Self-consistent calculation of the sommerfeld enhancement, *J. Cosmol. Astropart. Phys.* **06** (2016) 021.
- [51] K. Petraki, M. Postma, and J. de Vries, Radiative bound-state-formation cross-sections for dark matter interacting via a Yukawa potential, *J. High Energy Phys.* **04** (2016) 077.
- [52] J. Beacham *et al.*, Physics beyond colliders at CERN: Beyond the standard model working group report, *J. Phys. G* **47**, 010501 (2020).
- [53] J. P. Lees *et al.* (BABAR Collaboration), Search for a dark photon in e^+e^- collisions at BABAR, *Phys. Rev. Lett.* **113**, 201801 (2014).
- [54] D. Banerjee *et al.*, Dark matter search in missing energy events with NA64, *Phys. Rev. Lett.* **123**, 121801 (2019).
- [55] T. Åkesson *et al.* (LDMX Collaboration), Light dark matter eXperiment (LDMX), [arXiv:1808.05219](https://arxiv.org/abs/1808.05219).
- [56] T. Binder, L. Covi, A. Kamada, K. Murai, T. Takahashi, and N. Yoshida, Matter power spectrum in hidden neutrino interacting dark matter models: A general treatment, *J. Cosmol. Astropart. Phys.* **11** (2016) 043.
- [57] C. Boehm, P. Fayet, and R. Schaeffer, Constraining dark matter candidates from structure formation, *Phys. Lett. B* **518**, 8 (2001).
- [58] J. H. Chang, R. Essig, and S. D. McDermott, Supernova 1987A constraints on sub-GeV dark sectors, millicharged particles, the QCD axion, and an axion-like particle, *J. High Energy Phys.* **09** (2018) 051.
- [59] G. G. Raffelt, Stars as Laboratories for Fundamental Physics: The Astrophysics of Neutrinos, Axions, and Other Weakly Interacting Particles (University of Chicago Press, Chicago, IL, 1996).
- [60] C. Boehm and P. Fayet, Scalar dark matter candidates, *Nucl. Phys. B* **683**, 219 (2004).
- [61] J. Chluba and R. A. Sunyaev, The evolution of CMB spectral distortions in the early universe, *Mon. Not. R. Astron. Soc.* **419**, 1294 (2012).
- [62] V. Poulin, J. Lesgourgues, and P. D. Serpico, Cosmological constraints on exotic injection of electromagnetic energy, *J. Cosmol. Astropart. Phys.* **03** (2016) 043.
- [63] S. O'Neil *et al.*, Endothermic self-interacting dark matter in Milky Way-like dark matter haloes, *Mon. Not. R. Astron. Soc.* **524**, 288 (2023).
- [64] P. De la Torre Luque, P. Carezza, and T. T. Q. Nguyen, Sub-keV dark matter can strongly ionize molecular clouds, *Phys. Rev. D* **113**, 063035 (2026).
- [65] D. Agius and T. R. Slatyer, Boosting the cosmic 21-cm signal with exotic Lyman- α from dark matter, [arXiv:2510.26791](https://arxiv.org/abs/2510.26791).
- [66] G. Krnjaic, D. McKeen, R. Mizuta, G. Mohlabeng, D. E. Morrissey, and D. Tuckler, X-rays from inelastic dark matter Freeze-in, *Phys. Rev. D* **112**, 115039 (2025).
- [67] E. Hardy and R. Lasenby, Stellar cooling bounds on new light particles: Plasma mixing effects, *J. High Energy Phys.* **02** (2016) 033.
- [68] Y.-S. Liu, D. McKeen, and G. A. Miller, Electrophobic scalar boson and muonic puzzles, *Phys. Rev. Lett.* **117**, 101801 (2016).
- [69] R. H. Cyburt, B. D. Fields, K. A. Olive, and T.-H. Yeh, Big bang nucleosynthesis: 2015, *Rev. Mod. Phys.* **88**, 015004 (2016).
- [70] J. Kopp, L. Michaels, and J. Smirnov, Loopy constraints on leptophilic dark matter and internal bremsstrahlung, *J. Cosmol. Astropart. Phys.* **04** (2014) 022.
- [71] E. Aprile *et al.* (XENON Collaboration), Search for new physics in electronic recoil data from XENONnT, *Phys. Rev. Lett.* **129**, 161805 (2022).
- [72] J. Aalbers *et al.* (LZ Collaboration), Search for new physics in low-energy electron recoils from the first LZ exposure, *Phys. Rev. D* **108**, 072006 (2023).
- [73] S. Chang, N. Weiner, and I. Yavin, Magnetic inelastic dark matter, *Phys. Rev. D* **82**, 125011 (2010).
- [74] A. L. Fitzpatrick, W. Haxton, E. Katz, N. Lubbers, and Y. Xu, The effective field theory of dark matter direct detection, *J. Cosmol. Astropart. Phys.* **02** (2012) 004.
- [75] Y. F. Bautista, A. Robertson, L. Sagunski, A. Smith-Orlik, and S. Tulin, Jeans model for the shapes of self-interacting dark matter halos, [arXiv:2511.10765](https://arxiv.org/abs/2511.10765).
- [76] T. Bringmann, F. Kahlhoefer, K. Schmidt-Hoberg, and P. Walia, Strong constraints on self-interacting dark matter with light mediators, *Phys. Rev. Lett.* **118**, 141802 (2017).
- [77] Y. Cui and F. D'Eramo, Surprises from complete vector portal theories: New insights into the dark sector and its interplay with Higgs physics, *Phys. Rev. D* **96**, 095006 (2017).
- [78] J. Aalbers *et al.* (XLZD Collaboration), The XLZD design book: Towards the next-generation liquid xenon observatory for dark matter and neutrino physics, *Eur. Phys. J. C* **85**, 1192 (2025).