**Supplementary materials**

**Framework of free volume theory**

In the earliest literatures of free volume theory, the parameter free volume *v*f is defined as the neighbor cage of an atom, in which the atom can move around without energy change [4]. The value of free volume reflects the structure feature of MGs, as a high free volume represents a high degree of loose-packing and a high atomic mobility. During loading, the viscous strain rate of a MG can be expressed as:

 (1)

The generation rate of free volume (averaged to each atom) through shear dilatation can be expressed as:

 (2)

The annihilation rate of free volume by structure relaxation can be expressed as:

 (3)

where *τ* is the shear stress, Δ*G*m is the activation energy of flow, *k* is the Boltzmann’s constant, *S*=2*G*(1+*υ*)/3(1-*υ*) is the elastic energy, *G* is the shear modulus, *υ* isthe Poisson’s ratio, *f* is the Debye frequency, Ω is the atomic volume, *v*\* is the critical size (effective hard sphere size of the atom) and usually could be taken as *v*\*=Ω, *vf* is the mean free volume content per atom, *α* and *n*D can be taken as constants with *α*≈1 and *n*D≈3 [4].

If we express the free volume content by the relative fraction, the viscous flow strain rate (Eq. 2) can be rewritten as:

 (4)

and the changing rate of free volume (Eq. 3 and Eq. 4) can be rewritten as:

 (5)

**Temperature rise during shear band sliding.**

During shear band sliding, part of the released mechanical energy will transfer into heat and cause temperature rise. According to the thin-film heat-diffusion solution in heat transfer theory, the spatial and temporal distribution of temperature rise caused by the shear band sliding can be expressed as [26]:

  (6)

where Δ*T*(*l*dis,*t*1) is the temperature rise at the position with a normal distance of *l*dis away from the shear band center at the moment of *t*1, *t*=0 is the moment when the shear band starts to slide, is the function of heat generation rate and depends on the sliding speed and shear stress *τ* with an expression of , *ρ* is the density, *c*p is the specific heat capacity, *α* is thermal diffusivity of the material. Actually, the temperature rise during stable shear band sliding is negligible. As indicated by many previous studies [26,27], this part of temperature rise will be no more than a few degrees Kelvin. Therefore, in this study, we still treat the temperature in shear band as a constant for simplification.

**Calculation of shear band nucleation**

In the article, the global shear stress during uniaxial compression of a MG is defined as:

 (7)

The material constant are set as shear modulus *G*=34.1 GPa [30], Poisson’s ratio *υ*=0.352, [30], Debye frequency *f*=6.542×1012 Hz [30], flow activation energy [5,34,35], mean atomic volume Ω=1.690×10-29 m3, according to the well-known vit-1 (Zr41.2Ti13.8Cu12.5Ni10.0Be22.5 at.%) MG [5]. The loading parameters are set as sample length *l*=4 mm,temperature *T*=300 K and press-head feeding rate *v*p=2×10-7 ms-1 (a typical press-head feeding rate in experiments [18]).

For an inhomogeneous region with an initial free volume fraction of *x*IR=0.029 (the structure relaxation rate at this free volume fraction is about 10-6 s-1) and a shear stress of (*c*=2), the evolutions of relative free volume fraction *x* and strain rate can be calculated based on the coupled equations set of Eq. 4, Eq. 5 and Eq. 7. The results are shown in Fig. 2 in the article.

For a homogeneous sample with an initial free volume fraction of *x*0=0.025 (the structure relaxation rate at this free volume fraction is about 10-15 s-1 and is negligible in the experimental time window) and no stress concentration (*c*=1), the evolutions of relative free volume fraction *x* and strain rate is calculated by the same method. The results are shown in Fig. S1. A divergence of free volume x, strain rate and free volume generation rate occurs at a critical stress of 3550 MPa.



Fig. S1 Evolutions of (a) free volume fraction *x* and strain rate , (b) free volume generation rate in a homogeneous sample during loading, with an initial free volume fraction of *x*0=0.025 and no stress concentration (*c*=1).

For a homogeneous sample with an initial free volume fraction of *x*0=0.025, 0.026, 0.027, 0.028, 0.029 and 0.030 (such a variation in initial free volume is relatively small compared to previous studies [5,7]), and the same stress condition of (Eq. 7), the evolutions of relative free volume fraction *x* and strain rate are calculated again. The results are shown in Fig. S2. The critical stress for the divergence decreases significantly with increasing initial free volume. It indicates that the critical condition for shear band nucleation is sensitive to the original structure (free volume) of the material.



Fig. S2 Evolutions of free volume fraction *x* and strain rate in a homogeneous sample during loading, with an initial free volume fraction of *x*0=0.025, 0.026, 0.027, 0.028, 0.029 and 0.030.

**Calculation of other shear band behaviors**

In this study, irreversible structure change is considered, with an expression of：

 (8)

where represents the generation rate of free volume fraction caused by shear dilatation, *x*p represents the permanent free volume, *β* is a ratio coefficient. In the calculation, we introduce a specified setting of , which represents that the free volume *x* will not drop below *x*p.

In the case of shear band sliding, the shear band sliding distance is defined as *L*(*t*), with a sliding speed of , where *d* represents the shear band thickness and is defined as *d*=20 nm. The shear stress in the band can be expressed as:

 (9)

With the coupled equation set of Eq. 4, Eq. 5 and Eq. 9, the evolution of sliding speed, sliding distance *L*, free volume fraction *x* and shear stress *τ* during the shear band sliding process can be calculated based on the initial values of *τ*g0 and *x*g0.

The results of shear band sliding (Fig.4 in the article) are based on the initial values of *τ*g0=800 MPa and *x*g0=0.1. The results of shear band arrest (Fig.5 in the article) are based on the initial values of *τ*g0=700 MPa and *x*g0=0.1.

In the case of load fluctuation, the shear stress in the band is rewritten as:

 (10)

The results (Fig.8 in the article) are calculated by the coupled equation set of Eq. 4, Eq. 5 and Eq. 10, based on the parameters of *τ*g0=750 MPa, *τ*p=100 MPa, *t*p=2s, and *x*g0=0.1.